# Non-uniform Analysis for Non-convex Optimization in Machine Learning

by

Jincheng Mei

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistical Machine Learning

Department of Computing Science

University of Alberta

# Abstract

The optimization of non-convex objective functions is a topic of central interest in machine learning. Remarkably, it has recently been shown that simple gradient-based optimization can achieve globally optimal solutions in important non-convex problems that arise in machine learning, including policy gradient optimization in reinforcement learning (RL), generalized linear model training in supervised learning (SL), and over-parameterized neural network training in deep learning. However, previous work generally relies on *uniform* properties of the optimization landscape, ignoring relevant problem structure, which limits both the applicability and strength of the theoretical results that can be obtained.

In this thesis, motivated by fundamental problems in RL and SL, I investigate a **non-uniform analysis** for non-convex optimization.

Chapter 2 studies policy gradient optimization (PG) in RL and resolves three open problems in the literature by introducing a new analysis tool called non-uniform Łojasiewicz inequality (NŁ). In particular, this chapter shows that *(i)* PG optimization with a softmax parameterization converges to a globally optimal policy at a $O(1/t)$ rate; *(ii)* adding entropy regularization improves the convergence rate of PG to $O(e^{-c \cdot t})$ (where $c > 0$) to a regularized optimal policy; and *(iii)* an $\Omega(1/t)$ lower bound can be established on the worst case convergence of softmax PG. The separation of rates is further explained using the concept of the NŁ degree. These results provide a theoretical explanation of the optimization advantage of entropy regularization.

Next, Chapter 3 reconsiders a common policy parameterization used in machine learning: the softmax transform. Two negative results are established for using the softmax transform in gradient based optimization. In particular, this chapter shows that *(i)* optimizing any expectation with respect to the softmax must exhibit sensitivity to parameter initialization ("softmax gravity well"); and *(ii)* optimizing log-probabilities under the softmax must exhibit slow convergence ("softmax damping"). I propose an alternative *escort* mapping that demonstrates better optimization properties for PG and cross entropy minimization in SL. This analysis is based on the NŁ inequality and a new non-uniform smoothness (NS) property. These difficulties with the softmax and the advantage of the escort transform are further explained by the concept of the NŁ coefficient.

Chapter 4 then introduces a non-uniform analysis that combines the non-uniform smoothness (NS) property and the NŁ inequality, using the combination to more accurately characterize non-convex objective landscapes and inspire new geometry-aware gradient descent methods. One interesting result for general optimization is that geometry-aware first-order methods can converge to global optimality faster than the classical $\Omega(1/t^2)$ lower bounds if one additionally considers these non-uniform properties. This chapter then applies new geometry-aware first-order methods to PG and generalized linear model training (GLM). For PG, it is shown that normalizing gradient ascent can accelerate convergence to $O(e^{-c \cdot t})$ for some $c > 0$, while incurring less overhead than existing algorithms. For GLM, it is shown that geometry-aware normalized gradient descent can also achieve a linear convergence rate, which significantly improves the best known results. Additionally, I show that these geometry-aware gradient descent methods can escape landscape plateaus faster than standard gradient descent.

Finally, Chapter 5 extends the analysis to stochastic policy optimization,

and shows that the preferability of optimization methods depends critically on whether stochastic versus exact gradients are used. By introducing the concept of *committal rate*, this chapter contributes two key findings: *(i)* identifying a criterion for determining almost sure global convergence; and *(ii)* revealing an inherent trade-off between exploiting geometry to accelerate convergence versus achieving almost sure global optimality. This committal rate theory is then used to explain why practical policy optimization methods are sensitive to random initialization, leading to the development of an ensemble method that can be guaranteed to achieve near-optimal solutions with high probability.

*One's destiny, of course, depends on self-struggle, but also takes into account the historical schedule.*

– Jiang Zemin

# Acknowledgements

I am most grateful to my supervisor, Dale Schuurmans, for accepting me as his student, for his long-term collaboration and supervision, from which I learned a lot about research and beyond. In retrospect, I especially owe him many thanks for his important advice. In 2017, Dale advised me to explore multiple research directions, which later turned out to be critical to my research. In 2018 and 2019, Dale helped me with two internship opportunities, which have been precious experiences that have benefited my career. I have been always impressed by his brilliant talent of naming things, including many results in this dissertation. Without his excellent writing, it would be hard to get my research work published. I am extremely fortunate to have such a supportive supervisor and to work with him.

I would like to thank several senior researchers who collaborated with me as my mentors. I am grateful to Ruitong Huang for his mentorship at Borealis AI. I sincerely thank Lihong Li for his help and mentorship at Google Brain. I greatly appreciate Bo Dai for taking responsibility of mentoring me after Lihong, and for being supportive to our research and my career. I would like thank Csaba Szepesvári for his collaboration and valuable advice. Moreover, thanks to Csaba for teaching me climbing skills, which I enjoyed a lot.

I am thankful to a number of students and friends who helped me. I greatly appeciate Csaba Szepesvári and Martha White for serving in my supervisory committee. I would like to thank Lihong Li, Michael Bowling, and Lin Xiao for being my candidacy / thesis examiners.

Finally, I would like to thank my family and friends, especially my parents, who have always supported me during my study.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| DL | Deep Learning |
| RL | Reinforcement Learning |
| SL | Supervised Learning |
| NS | Non-uniform Smoothness, Definition 6 |
| NŁ | Non-uniform Łojasiewicz, Definition 7 |
| GD | Gradient Descent, Definition 8 |
| GNGD | Geometry-aware Normalized Gradient Descent, Definition 9 |
| PG | Policy Gradient, Algorithm 1 |
| SPG | Softmax Policy Gradient, Algorithm 1 |
| GNPG | Geometry-aware Normalized Policy Gradient, Algorithm 2 |
| MDP | Markov Decision Process, Section 2.2 |
| SGW | Softmax Gravity Well, Section 3.2 |
| EPG | Escort Policy Gradient, Section 3.3 |
| SCE | Softmax Cross Entropy, Section 3.5 |
| ECE | Escort Cross Entropy, Section 3.5 |
| MD | Mirror Descent, Remark 8 |
| KL | Kullback–Leibler |
| PŁ | Polyak-Łojasiewicz, Section 4.3 |
| KŁ | Kurdyka-Łojasiewicz, Section 4.3 |
| NGD | Normalized Gradient Descent, Section 4.7 |
| GLM | Generalized Linear Model, Section 4.7 |
| MSE | Mean Squared Error, Eq. (4.22) |
| NPG | Natural Policy Gradient, Section 5.2 |

# List of Symbols

| | |
|---|---|
| $\mathbb{R}^d$ | $d$-dimensional Euclidean space |
| $t$ | iteration number / time |
| $\eta$ | learning rate / stepsize |
| $\delta$ | sub-optimality gap |
| $\Delta$ | reward / value gap |
| softmax | softmax transform, Eq. (2.5) |
| $\mathrm{diag}(x)$ | diagonal matrix that has vector $x$ as the diagonal |
| $\mathbb{E}$ | expectation |
| $\|x\|_2$ | $\ell_2$ norm of vector $x \in \mathbb{R}^K$, $\left[\sum_{i=1}^K x(i)^2\right]^{\frac{1}{2}}$ |
| $\|x\|_1$ | $\ell_1$ norm of vector $x \in \mathbb{R}^K$, $\sum_{i=1}^K \|x(i)\|$ |
| $\|x\|_\infty$ | $\ell_\infty$ norm of vector $x \in \mathbb{R}^K$, $\max_{i\in[K]} \|x(i)\|$ |
| $\xi$ | NŁ degree, Definition 1 |
| $C(\theta)$ | NŁ coefficient, Definition 3 |
| $\beta$ | smoothness coefficient, Definition 4 |
| $\beta(\theta)$ | NS coefficient, Definition 6 |
| $D_{\mathrm{KL}}(\cdot\|\cdot)$ | KL divergence |

# Chapter 1

# Introduction

Optimization plays a central role in machine learning: many machine learning problems can be formulated as optimizing some form of objective function. During the past few decades, the focus of machine learning has shifted from the classical, well developed topic of convex analysis (Boyd et al., 2004; Nesterov, 2018; Rockafellar, 2015) to non-convex optimization. The key reason is that current machine learning techniques, such as deep learning (DL) (LeCun et al., 2015) and reinforcement learning (RL) (Sutton and Barto, 2018), almost always involve optimizing a non-convex objective. Deep neural networks achieve outstanding performance in practice, but have resisted a clear theoretical understanding from the perspectives of optimization (Sun, 2020) and generalization (Zhang et al., 2016), while value based objectives in RL are naturally non-concave in complex decision processes.

Despite these challenges, there have been a number of exciting recent advances in non-convex optimization research. In particular, Ge et al. (2016) study the matrix completion problem and find that it has no spurious local minima; i.e., all local minima achieve the globally minimal loss value. Kawaguchi (2016) have shown that deep linear network training also avoids bad local minima. Allen-Zhu et al. (2019), Du et al. (2018), and Zou et al. (2020) proved that, with high probability over random initializations, (stochastic) gradient descent methods converge to globally optimal solutions in over-parameterized neural networks. Agarwal et al. (2019) and Bhandari and Russo (2019) showed that policy gradient based methods converge to a globally op-

timal policy in tabular settings.

## 1.1 Examples

Although this recent progress is impressive, there still exist many open problems, and the understanding and theoretical justification for many empirically successful methods remains lacking. I will use two important examples to illustrate many of the key findings in this thesis.

**Example 1** (Policy gradient (PG) optimization)**.** *Given a reward vector $r \in \mathbb{R}^K$ and a probability distribution $\pi_\theta$ over a $K$-dimensional action space $[K] \coloneqq \{1, 2, \ldots, K\}$, parameterized by $\theta \in \Theta$ where $\pi_\theta(a) \geq 0$ and $\sum_{a\in[K]} \pi_\theta(a) = 1$, we would like to maximize the expected reward of $\pi_\theta$,*

$$\max_{\theta:[K]\to\mathbb{R}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} [r(a)], \tag{1.1}$$

*by performing gradient ascent updates on $\theta$, i.e., for all $t \geq 1$,*

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \quad \forall a \in [K]. \tag{1.2}$$

The policy optimization problem in Example 1, also known as Policy gradient (PG), is a representative problem in reinforcement learning (RL). Understanding the optimization behaviour of gradient descent in this scenario is of foundational significance to understanding many RL methods, such as policy search and actor critic methods. However, the convergence properties of PG using standard parameterizations remains underdeveloped.

**Example 2** (Generalized linear model (GLM) training)**.** *Consider a finite dataset $\mathcal{D} = \{(x_i, y_i)\}_{i\in[N]}$, where $y_i \in [0, 1]$ is a target prediction and $x_i$ is a data point. A conditional prediction $\pi_i$ of $y_i$ given $x_i$ can be expressed in terms of a linear feature mapping $\phi : x_i \mapsto \phi_i \in \mathbb{R}^d$ and parameters $\theta \in \mathbb{R}^d$, via*

$$\pi_i = \sigma(\phi_i^\top \theta), \tag{1.3}$$

*where $\sigma$ is a certain non-linear transform. We would like to learn a good predictor by minimizing a loss function with gradient descent, i.e.,*

$$\min_\theta \mathcal{L}(\theta) = \min_{\theta\in\mathbb{R}^d} \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - y_i)^2. \tag{1.4}$$

This second example, the generalized linear model (GLM) training of Example 2, is widely applied in supervised learning (SL). However, the optimization behavior of gradient descent is still not thoroughly understood for different combinations of non-linear transforms and loss functions. There also remain questions about whether better performing methods can be devised for this problem.

## 1.2 Approach and Overview

The thesis begins by carefully analyzing Example 1. I propose a new analytical tool, the non-uniform Łojasiewicz (NŁ) inequality, motivated by the failure of standard techniques to fully characterize gradient descent optimization for this problem. These alternative NŁ inequalities are used to resolve three open problems in the PG optimization literature, which provides a new understanding and explanation for the optimization advantage of entropy regularization. These results are presented in Chapter 2.

I then further develop these novel NŁ inequality tools to reveal negative results to accompany the above findings. In particular, several optimization disadvantages of the standard softmax transform are revealed, both for the RL and SL settings, using the concept of NŁ coefficient to provide explanations and solutions. This part is presented in Chapter 3.

I then expand the above negative results by introducing another important new concept, the non-uniform smoothness (NS) property. By combining the NŁ and NS properties, a novel geometry-aware first-order method is developed, with a corresponding non-uniform analysis. This analysis applies to general non-convex optimization, PG optimization, and GLM training, while significantly improving existing results and mitigating negative aspects of existing algorithms. This part is presented in Chapter 4.

Finally, I extend the analysis to the stochastic setting, where an anomaly is observed that faster policy gradient algorithms can become dominated by slower counterparts. To understand this phenomenon, I introduce a new concept called the committal rate to explain the results, and reveal an inherent

trade-off between convergence speed and almost sure global convergence, which characterizes the fundamental difficulty of stochastic policy optimization. This part is presented in Chapter 5.

## 1.3 Contributions

The main contributions of this dissertation are the following.

**Non-uniform properties.** This thesis introduces two new non-uniform properties, the non-uniform smoothness property (NS) and the non-uniform Łojasiewicz (NŁ) inequality, to characterize non-convex objective landscapes. These properties generalize and unify existing concepts in the optimization literature. I apply these to fundamental problems in reinforcement learning and supervised learning, and show that the key quantities in these properties—the NS coefficient, NŁ degree, and NŁ coefficient—can be used to explain the optimization advantages of regularization, parameterzation, surrogate objectives, and label smoothing. This work was published as (Mei et al., 2021b).

**Non-uniform analysis and geometry-aware first-order methods.** The thesis also introduces a non-uniform analysis for non-convex optimization, and a novel family of geometry-aware normalized gradient descent (GNGD) methods. By exploiting non-uniform landscape information, GNGD can be shown to achieve linear convergence rates of $O(e^{-c\cdot t})$ (where $c > 0$), which overcomes the classical $\Omega(1/t^2)$ lower bounds for convex-smooth optimization. These results broaden our fundamental knowledge of the set of objectives that admit efficient global optimization. This work was published as (Mei et al., 2021b).

**Softmax policy gradient (PG) convergence analysis.** Next, the thesis analyzes softmax policy gradient (PG) methods using the non-uniform Łojasiewicz inequality (NŁ). Three open problems are resolved by providing value function objectives that satisfy the NŁ inequalities. First, it is shown that softmax PG converges to a globally optimal policy with rate $O(1/t)$. Second, entropy regularized PG is shown to converge to a regularized optimal

policy at a linear rate of $O(e^{-c \cdot t})$. Finally, unregularized PG is shown to follow a rate lower bound of $\Omega(1/t)$. These results reveal an optimization advantage of entropy regularization: it accelerates the convergence of PG methods. This acceleration is further explained using the concept of NŁ degree, a key quantity in the NŁ inequality. This work was published as (Mei et al., 2020b).

**Softmax gravity well phenomenon and the escort transform for PG.** Using empirical and theoretical results, the thesis illustrates a fundamental disadvantage of the common practice in machine learning of using the softmax transform. Optimizing an expectation over softmax probabilities is shown to exhibit initialization sensitivity and slow escaping behavior from landscape plateaus. This phenomenon is referred to as the "softmax gravity well (SGW)" for PG. An alternative escort transform is proposed, and the resulting escort PG methods provably mitigate the SGW problem. The difficulties with softmax and the effectiveness of the escort transform are both explained using the concept of NŁ coefficient, another key quantity in the NŁ inequality. This work was published as (Mei et al., 2020a) and received an oral presentation.

**Softmax damping phenomenon and escort cross entropy minimization.** A disadvantage of using the softmax transform in cross entropy minimization, called "softmax damping", is revealed. Here it is shown that the convergence rate degenerates from $O(e^{-c \cdot t})$ to $O(1/t)$ when minimizing the log-probabilities of a softmax transform in supervised learning. Using the escort transform for cross entropy minimization provably achieves fast linear convergence rates. These results are explained by observing that vanishing NŁ coefficients lead to decreasing NŁ degrees, an interplay between the two key quantities in the NŁ inequality. This work was published as part of the same paper (Mei et al., 2020a) above.

**Geometry-aware normalized PG.** A geometry-aware normalized gradient descent (GNGD) method is developed and applied to PG optimization in RL, where it is shown that the resulting geometry-aware normalized PG

(GNPG) method achieves a global linear convergence rate of $O(e^{-c \cdot t})$ without using regularization or introducing an arg max operation in each iteration. This result breaks the $\Omega(1/t)$ lower bound of standard softmax PG with bounded learning rate of $O(1)$. The key reason is that the policy value function satisfies the non-uniform smoothness (NS) property, while the NS coefficient is exactly the PG norm. This allows GNPG to better leverage the NŁ inequality compared to standard PG, accelerating both its convergence rate and escaping behavior from landscape plateaus. This work was published as (Mei et al., 2021b).

**Geometry-aware normalization in generalized linear model training.** The geometry-aware normalized gradient descent (GNGD) method is also applied to generalized linear model (GLM) training in supervised learning. It is shown that GNGD achieves a global linear convergence rate of $O(e^{-c \cdot t})$ for minimizing the mean squared error (MSE), which significantly improves the best existing result of $O(1/\sqrt{t})$. These results are achieved by observing that MSE in a GLM satisfies a new NŁ inequality and NS property, which enables both gradient descent (GD) and GNGD to achieve fast convergence. Using the NS and NŁ properties, I further show that GNGD escapes from landscape plateaus strictly faster than GD, providing new understanding of using geometry-aware normalization in GLM training. This work was published as part of the same paper (Mei et al., 2021b) above.

**Stochastic policy optimization and committal rate theory.** Finally, the previous results are extended to the stochastic setting, revealing an apparent anomaly that the preferability of policy optimization algorithms changes dramatically depending on whether true versus on-policy stochastic gradients are considered. To explain the anomaly, the concept of committal rate is introduced, which serves as a criterion for determining almost sure global convergence. Using the committal rate theory, the thesis uncovers a fundamental trade-off between leveraging geometry to accelerate convergence and achieving almost sure global convergence; in particular, no uninformed algorithm can

improve the $O(1/t)$ convergence rate without incurring a positive probability of failure (i.e. diverging or converging to a sub-optimal stationary point). This finding explains the sensitivity to random initialization in practical policy optimization algorithms, which motivates the development of an ensemble method that can achieve fast convergence to global optima with high probability. This work has been submitted for review as (Mei et al., 2021a).

### 1.3.1 Publications

The papers related to the topics covered in this dissertation are as follows.

- Jincheng Mei, Chenjun Xiao, Ruitong Huang, Dale Schuurmans, and Martin Müller. On Principled Entropy Exploration in Policy Optimization. IJCAI 2019. See Mei et al. (2019)
- Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. ICML 2020. See Mei et al. (2020b)
- Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the Gravitational Pull of Softmax. NeurIPS 2020, oral. See Mei et al. (2020a)
- Jincheng Mei*, Yue Gao*, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. Leveraging Non-uniformity in First-order Non-convex Optimization. ICML 2021. See Mei et al. (2021b)
- Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvári†, and Dale Schuurmans†. Understanding the Effect of Stochasticity in Policy Optimization. Preprint 2021 (under review). See Mei et al. (2021a).
- * indicates equal contribution, and † indicates equal advising.

# Chapter 2

# Global Convergence Rates of Softmax Policy Gradient

We start by considering a special fundamental non-convex optimization problem, i.e., policy gradient optimization in reinforcement learning. It is necessary to introduce a new analysis tool which we call non-uniform Łojasiewicz inequality, to resolve several open problems in the reinforcement learning literature.

The results in this chapter appeared in Mei et al. (2020b).

## 2.1 Introduction

The *policy gradient* is one of the most foundational concepts in Reinforcement Learning (RL) (Sutton and Barto, 2018), lying at the core of policy-search and actor-critic methods. The policy gradient theorem (Sutton et al., 2000), in particular, establishes a general foundation for policy search methods, by showing that an unbiased estimate of the gradient of a policy's expected return with respect to its parameters can still be recovered from an approximate value function (provided the approximation is a best fit). As an approach to RL, policy gradient ascent is particularly appealing due to its simplicity and directness: it targets the quantity of interest, it is inherently sound given appropriate step size control, and it can be readily combined with network function approximation to achieve effective empirical performance (e.g., Schulman et al. (2015) and Schulman et al. (2017)).

Despite the prevalence and importance of policy optimization in RL, the

theoretical understanding of the policy gradient method has, until recently, been severely limited. A key barrier to understanding is the inherent non-convexity of the value landscape with respect to standard policy parametrizations. As a result, little has been known about the global convergence behavior of policy gradient methods. Recently, important new progress in understanding the convergence behavior of policy gradient has been achieved. In this dissertation we will restrict ourselves to the tabular setting, we analyze the part of the literature that also deals with this setting. While the tabular setting is clearly limiting, this is the setting where so far the cleanest results have been achieved and understanding this setting is a necessary first step towards the bigger problem of understanding RL algorithms. Returning to the discussion of recent work, Fazel et al. (2018) showed that gradient based methods achieve global convergence in special linear quadratic regulator settings. For general Markov decision processes, Bhandari and Russo (2019) showed that, with direct parametrization, projected gradient ascent on the simplex does not suffer from spurious local optima. In concurrent work, Agarwal et al. (2019) showed that *(i)* with direct parametrization, projected gradient ascent converges at rate $O(1/\sqrt{t})$ to a global optimum; and *(ii)* with softmax parametrization, policy gradient converges asymptotically. Agarwal et al. (2019) also analyze other variants of policy gradient, and show that policy gradient with relative entropy regularization converges at rate $O(1/\sqrt{t})$, natural policy gradient (mirror descent) converges at rate $O(1/t)$, and given a "compatible" function approximation (thus, going beyond the tabular case) natural policy gradient converges at rate $O(1/\sqrt{t})$. Shani et al. (2020) obtains the slower rate $O(1/\sqrt{t})$ for mirror descent. They also proposed a variant that adds entropy regularization and prove a rate of $O(1/t)$ for this modified problem.

Despite these advances, many open questions remain in understanding the behavior of policy gradient methods, even in the tabular setting and even when the true gradient is available in the updates. In this chapter, we provide answers to the following three questions left open by previous work in this area:

**(i)** What is the convergence rate of policy gradient methods with softmax

parametrization? The best previous result, due to Agarwal et al. (2019), established asymptotic convergence but gave no rates.

**(ii)** What is the convergence rate of entropy regularized softmax policy gradient? Figuring out the answer to this question was explicitly stated as an open problem by Agarwal et al. (2019).

**(iii)** Empirical results suggest that entropy helps optimization (Ahmed et al., 2019). Can this empirical observation be turned into a rigorous theoretical result?[1]

In this chapter, we answer the above three open questions and our contributions are summarized as follows.

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

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

[1] While Shani et al. (2020) suggest that entropy regularization speeds up mirror descent to achieve the rate of $O(1/t)$, in light of the corresponding result of Khodadadian et al. (2021) who established the $O(e^{-c \cdot t})$ rate for the unregularized version of mirror descent, the conclusion is questionable: entropy does not speed up mirror descent in in terms of convergence rate. It remains open whether it speeds up mirror descent in terms of better constant dependence.

updates behave similarly to the contraction operator in value learning, with a contraction factor that depends on the current policy.

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

The remainder of the chapter is organized as follows. After introducing notation and defining the setting in Section 2.2, we present the three main contributions in Sections 2.3 to 2.5 as aforementioned. Section 2.7 gives our brief summary.

## 2.2 Notations and Settings

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

An agent interacts with the environment, i.e., the MDP $\mathcal{M}$, using a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. Given a state $s_t$, the agent takes an action $a_t \sim \pi(\cdot|s_t)$,

receives a one-step scalar reward $r(s_t, a_t)$ and a next-state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. The long-term expected reward, also known as the value function of $\pi$ under $s$, is defined as

$$V^\pi(s) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0=s, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \tag{2.1}$$

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

$$Q^\pi(s, a) \coloneqq r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a)V^\pi(s'). \tag{2.2}$$

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

$$d^\pi_{s_0}(s) \coloneqq (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s|s_0, \pi, \mathcal{P}), \tag{2.3}$$

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

$$V^{\pi^*}(\rho) = \max_{\pi:\mathcal{S}\to\Delta(\mathcal{A})} V^\pi(\rho). \tag{2.4}$$

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

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

Consider a tabular representation, i.e., $\theta(s, a) \in \mathbb{R}$ for all $(s, a)$, so that the policy $\pi_\theta$ can be parameterized by $\theta$ using the softmax transform, which exponentiates the components of the vector and normalizes it so that the result lies in the simplex. This can be used to transform vectors assigned to state-action pairs into policies:

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

$$\pi_\theta(a|s) = \frac{\exp\{\theta(s, a)\}}{\sum_{a'} \exp\{\theta(s, a')\}}. \tag{2.5}$$

We also extend this notation to the case when there are no states. Given a finite action set $[K] \coloneqq \{1, 2, \ldots, K\}$, for $\theta : [K] \to \mathbb{R}$, we define $\pi_\theta \coloneqq \mathrm{softmax}(\theta)$ using

$$\pi_\theta(a) = \frac{\exp\{\theta(a)\}}{\sum_{a'} \exp\{\theta(a')\}}, \quad \forall a \in [K]. \tag{2.6}$$

The problem of RL is then to find a policy $\pi_\theta$ that maximizes the value function, i.e.,

$$\sup_{\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}} V^{\pi_\theta}(\rho). \tag{2.7}$$

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

$$\left(\frac{d\pi_\theta}{d\theta}\right)^\top = H(\pi_\theta). \tag{2.8}$$

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

**Smoothness.** A function $f : \Theta \to \mathbb{R}$ with $\Theta \subset \mathbb{R}^d$ is $\beta$-smooth (w.r.t. $\ell_2$ norm, $\beta > 0$) if for all $\theta$, $\theta' \in \Theta$,

$$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2. \tag{2.9}$$

## 2.3 Policy Gradient

Policy gradient is a special policy search method. In policy search, one considers a family of policies parametrized by finite-dimensional parameter vectors, reducing the search for a good policy to searching in the space of parameters. This search is usually accomplished by making incremental changes (additive updates) to the parameters. Representative policy-based RL methods include

REINFORCE (Williams, 1992), natural policy gradient (Kakade, 2002), deterministic policy gradient (Silver et al., 2014), and trust region policy optimization (Schulman et al., 2015). In policy gradient methods, the parameters are updated by following the gradient of the map that maps policy parameters to values. Under mild conditions, the gradient can be reexpressed in a convenient form in terms of the policy's action-value function and the gradients of the policy parametrization:

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \cdot \mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_\theta}} \left[ \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot Q^{\pi_\theta}(s,a) \right]. \tag{2.10}$$

### 2.3.1 Vanilla Softmax Policy Gradient

We focus on the policy gradient method that uses the softmax parametrization Eq. (2.5). Since we consider the tabular case, the policy is then parametrized using the parameter $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ function and $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s,\cdot))$. The vanilla form of policy gradient for this case is shown in Algorithm 1.

**Algorithm 1** Policy Gradient Method

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

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

**Lemma 1.** *Softmax policy gradient w.r.t.* $\theta$ *is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a). \tag{2.11}$$

Recently, Agarwal et al. (2019) showed that softmax policy gradient asymptotically converges to $\pi^*$, i.e., $V^{\pi_{\theta_t}}(\rho) \to V^*(\rho)$ as $t \to \infty$ provided that

$\mu(s) > 0$ holds for all states $s \in \mathcal{S}$. We strengthen this result to show that the rate of convergence (in terms of value sub-optimality) is $O(1/t)$. The next section is devoted to this result. For better accessibility, we start with the result for the bandit case which presents an opportunity to explaining the main ideas underlying our result in a clean fashion.

### 2.3.2 Convergence Rate: One-state MDPs

As promised, in this section we consider the "bandit case". In particular, assume that the MDP has a single state and the discount factor $\gamma$ is zero: $\gamma = 0$. In this case, Eq. (2.1) reduces to maximizing the expected reward,

$$\max_{\theta:\mathcal{A}\to\mathbb{R}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} [r(a)]. \tag{2.12}$$

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

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

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

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r)\,. \tag{2.13}$$

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

**Update 1** (Softmax policy gradient, expected reward)**.**

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r), \quad \forall a \in [K]. \tag{2.14}$$

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

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

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

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r \,. \tag{2.15}$$

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

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(t \cdot c_t^2), \qquad \text{and} \tag{2.16}$$

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \leq \min \left\{ \sqrt{5T}/c_T,\ (5 \log T)/c_T^2 + 1 \right\}. \tag{2.17}$$

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

**Remark 1.** *The value of $\pi_{\theta_t}(a^*)$, while it is nonzero (and so is $c_t$) can be small (e.g., because of the choice of $\theta_1$). Consequently, its minimum $c_t$ can be*

*quite small and the upper bound in Lemma 4 can be large, or even vacuous. The dependence of the previous result on $\pi_{\theta_t}(a^*)$ comes from Lemma 3. As it turns out, it is not possible to eliminate or improve the dependence on $\pi_\theta(a^*)$ in Lemma 3. To see this consider $r = (5, 4, 4)^\top$, $\pi_\theta = (2\epsilon, 1/2 - 2\epsilon, 1/2)$ where $\epsilon > 0$ is small number. By algebra, we have,*

$$(\pi^* - \pi_\theta)^\top r = 1 - 2\epsilon > 1/2, \tag{2.18}$$

$$\frac{d\pi_\theta^\top r}{d\theta} = (2\epsilon - 4\epsilon^2, -\epsilon + 4\epsilon^2, -\epsilon)^\top, \tag{2.19}$$

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \epsilon \cdot \sqrt{6 - 24\epsilon + 32\epsilon^2} \leq 3\epsilon. \tag{2.20}$$

*Hence, we have, for any constant $C > 0$,*

$$C \cdot (\pi^* - \pi_\theta)^\top r > C/2 > 3\epsilon \geq \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2, \tag{2.21}$$

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

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

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

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

**Proposition 2.** *For any initialization there exist $t_0 \geq 1$ such that for any $t \geq t_0$, $t \mapsto \pi_{\theta_t}(a^*)$ is increasing. In particular, when $\pi_{\theta_1}$ is the uniform distribution, $t_0 = 1$.*

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

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(c^2 \cdot t), \tag{2.22}$$

*where $c = \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ is a constant that depends on $r$ and $\theta_1$, but it does not depend on the time $t$.*

Proposition 2 suggests that one should set $\theta_1$ so that $\pi_{\theta_1}$ is uniform. Using this initialization, we can show that $\inf_{t \geq 1} \pi_{\theta_t}(a^*) \geq 1/K$, strengthening Theorem 2:

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5K^2/t, \qquad \text{and} \tag{2.23}$$

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \leq \min\left\{K\sqrt{5T},\ 5K^2 \log T + 1\right\}. \tag{2.24}$$

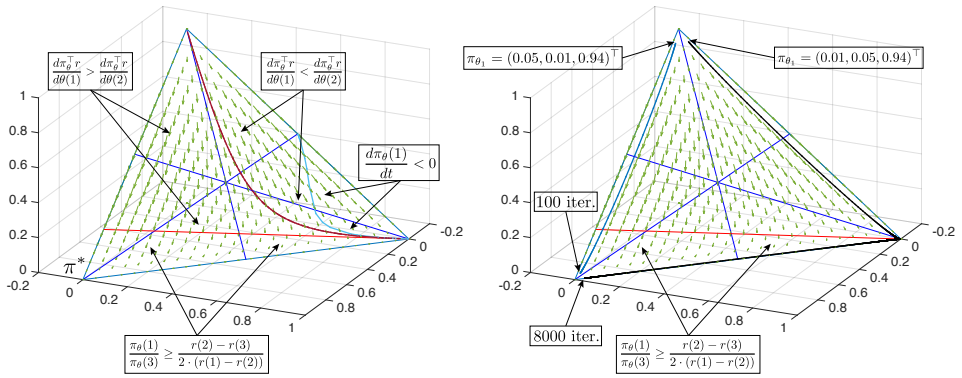
**Remark 2.** *In Section 2.5, we prove a lower bound $\Omega(1/t)$ for the same update rule, showing that the upper bound $O(1/t)$ of Theorem 2, apart from constant factors, is unimprovable.*

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

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

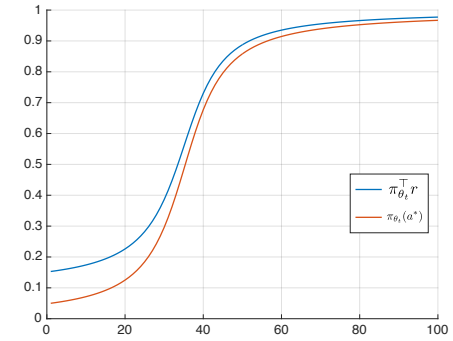
$$t_0 = \min\left\{t \geq 1 \ :\ \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \geq \frac{r(2) - r(3)}{2 \cdot (r(1) - r(2))}\right\}. \tag{2.25}$$

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

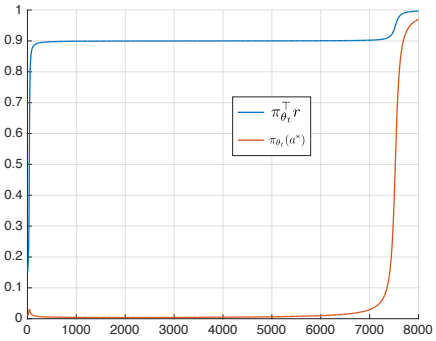



(a) Softmax gradient flow.

(b) Good & bad initializations.




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




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

Figure 2.1: Visualization of proof idea for Lemma 5.

**Visualization.** In Fig. 2.1(a), the region below the red line corresponds to

$$\mathcal{R} = \left\{ \pi_\theta : \frac{\pi_\theta(1)}{\pi_\theta(3)} \geq \frac{r(2) - r(3)}{2 \cdot (r(1) - r(2))} \right\}, \tag{2.26}$$

where $r = (1.0, 0.9, 0.1)^\top$. Any globally convergent iteration will enter $\mathcal{R}$ within finite time (the closure of $\mathcal{R}$ contains $\pi^*$) and never leaves $\mathcal{R}$ (this is the main idea in Lemma 5). Subfigure (b) shows the behavior of the gradient updates with "good" ($\pi_{\theta_1} = (0.05, 0.01, 0.94)^\top$) and "bad" ($\pi_{\theta_1} = (0.01, 0.05, 0.94)^\top$) initial policies. While these are close to each other, the iterates behave quite differently (in both cases $\eta = 2/5$). From the good initialization, the iterates converge quickly: after 100 iterations the distance to the optimal policy is already quite small. At the same time, starting from a "bad" initial value, the iterates are first attracted toward a sub-optimal action. It takes more than 7000 iterations for the algorithm to escape this sub-optimal

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

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

### 2.3.3 Convergence Rate: General MDPs

For general MDPs, the optimization problem takes the form

$$\max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} V^{\pi_\theta}(\rho) = \max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathbb{E}_{s\sim\rho} \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a). \tag{2.27}$$

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

**Assumption 2** (Sufficient exploration)**.** *The initial state distribution satisfies*

$$\min_s \mu(s) > 0. \tag{2.28}$$

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

**Proposition 3.** *There exists an MDP and $\mu$ with $\min_s \mu(s) = 0$ such that there exists $\theta^* : \mathcal{S} \times \mathcal{A} \to [0,\infty]$ such that $\theta^*$ is the stationary point of $\theta \mapsto V^{\pi_\theta}(\mu)$*

*while $\pi_{\theta^*}$ is not an optimal policy. Furthermore, this stationary point is an attractor, hence, starting gradient ascent in a small enough vicinity of $\theta^*$ will make it converge to $\theta^*$.*

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

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

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

As mentioned in Section 2.3.2, smoothness and (uniform) Łojasiewicz inequality are sufficient to prove a convergence rate. As noted by Agarwal et al. (2019), the main difficulty is to establish a (uniform) Łojasiewicz inequality for the softmax parametrization. As it turns out, the results from the bandit case carry over to multi-state MDPs.

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

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

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot [V^*(\rho) - V^{\pi_\theta}(\rho)] \,. \tag{2.29}$$

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

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

**Lemma 9.** *Let Assumption 2 hold. Using Algorithm 1, we have,*

$$c \coloneqq \inf_{s\in\mathcal{S},t\geq 1} \pi_{\theta_t}(a^*(s)|s) > 0. \tag{2.30}$$

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

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

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{16\cdot S}{c^2\cdot(1-\gamma)^6\cdot t}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2\cdot\left\|\frac{1}{\mu}\right\|_\infty. \tag{2.31}$$

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

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

## 2.4 Entropy Regularized Policy Gradient

Agarwal et al. (2019) considered relative-entropy regularization in policy gradient to get an $O(1/\sqrt{t})$ convergence rate. As they note, relative-entropy is more "agressive" in penalizing small probabilities than the more "common"

entropy regularizer (cf. Remark 5.2 in their paper) and it remains unclear whether this latter regularizer leads to an algorithm with the same rate. In this section, we answer this positively and in fact prove a much better rate. In particular, we show that entropy regularized policy gradient with the softmax parametrization enjoys a linear rate of $O(e^{-c\cdot t})$. In retrospect, perhaps this is unsurprising as entropy regularization bears a strong similarity to introducing a strongly convex regularizer in convex optimization, where this change is known to significantly improve the rate of convergence of first-order methods (e.g., Nesterov, 2018, Chapter 2).

### 2.4.1 Maximum Entropy RL

In entropy regularized RL, or sometimes called maximum entropy RL, near-deterministic policies are penalized (Haarnoja et al., 2018; Mei et al., 2019; Mnih et al., 2016; Nachum et al., 2017; Neu et al., 2017; Williams and Peng, 1991; Xiao et al., 2019; Xiao et al., 2018; Ziebart, 2010; Ziebart et al., 2008), which is achieved by modifying the value of a policy $\pi$ to

$$\tilde{V}^{\pi}(\rho) \coloneqq V^{\pi}(\rho) + \tau \cdot \mathbb{H}(\rho, \pi)\,, \tag{2.32}$$

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

$$\mathbb{H}(\rho,\pi) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}}\left[\sum_{t=0}^{\infty} -\gamma^t \log\pi(a_t|s_t)\right], \tag{2.33}$$

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

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

$$\frac{\partial\tilde{V}^{\pi_\theta}(\mu)}{\partial\theta(s,a)} = \frac{1}{1-\gamma}\cdot d_{\mu}^{\pi_\theta}(s)\cdot\pi_\theta(a|s)\cdot\tilde{A}^{\pi_\theta}(s,a), \tag{2.34}$$

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

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

$$\tilde{A}^{\pi_\theta}(s,a) \coloneqq \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s), \tag{2.35}$$

$$\tilde{Q}^{\pi_\theta}(s,a) \coloneqq r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s'). \tag{2.36}$$

### 2.4.2 Convergence Rate: One-state MDPs

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

In the one-state case with $\gamma = 0$, Eq. (2.32) reduces to maximizing the entropy-regularized reward,

$$\max_{\theta:\mathcal{A}\to\mathbb{R}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} [r(a) - \tau \log \pi_\theta(a)]. \tag{2.37}$$

Again, Eq. (2.37) is a non-concave function of $\theta$. In this case, regularized policy gradient reduces to

$$\frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} = H(\pi_\theta)(r - \tau \log \pi_\theta), \tag{2.38}$$

where $H(\pi_\theta)$ is the same as in Eq. (2.8). Using the above gradient in Algorithm 1 we have the following update rule:

**Update 2** (Softmax policy gradient, maximum entropy reward)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot H(\pi_{\theta_t})(r - \tau \log \pi_{\theta_t}). \tag{2.39}$$

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

**Softmax optimal policy.** $\pi_\tau^* \coloneqq \text{softmax}(r/\tau)$ is the optimal solution of Eq. (2.37).

**Remark 4.** *At this stage, we could use arguments similar to those of Section 2.3 to show the* $O(1/t)$ *convergence of* $\pi_{\theta_t}$ *to* $\pi_\tau^*$*. However, we can use an alternative idea to show that entropy-regularized policy gradient converges significantly faster. The issue of bias will be discussed later.*

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

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

$$\|\zeta_{t+1}\|_2 \leq \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a)\right) \cdot \|\zeta_t\|_2, \tag{2.40}$$

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

This lemma immediately implies the following bound:

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

$$\|\zeta_t\|_2 \leq \frac{2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\exp\left\{\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}}. \tag{2.41}$$

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

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

$$\sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)] \geq c \cdot (t-1). \tag{2.42}$$

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

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

$$\tilde{\delta}_t \leq \frac{2(\tau\|\theta_1\|_\infty + 1)^2 K/\tau}{\exp\{2\tau\eta \cdot c \cdot (t-1)\}}, \tag{2.43}$$

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

### 2.4.3 Convergence Rate: General MDPs

For general MDPs, the problem is to maximize $\tilde{V}^{\pi_\theta}(\rho)$ in Eq. (2.32). The soft-max optimal policy $\pi_\tau^*$ is known to satisfy the following consistency conditions (Nachum et al., 2017):

$$\pi_\tau^*(a|s) = \exp\left\{(\tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s))/\tau\right\}, \tag{2.44}$$

$$\tilde{V}^{\pi_\tau^*}(s) = \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\tau^*}(s,a)/\tau\right\}. \tag{2.45}$$

Using a somewhat lengthy calculation, we show that the discounted entropy in Eq. (2.33) is smooth:

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

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

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

$$\left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq C(\theta) \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}}, \tag{2.46}$$

*where*

$$C(\theta) := \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\|\frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}}. \tag{2.47}$$

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

**Lemma 16.** *Using Algorithm 1 with the entropy regularized objective, we have*

$$c := \inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0. \tag{2.48}$$

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

**Theorem 6.** *Let Assumption 2 hold and let $c$ be the positive constant from Lemma 16. Using Algorithm 1 with the entropy regularized objective and softmax parametrization and*

$$\eta = \frac{(1-\gamma)^3}{8 + \tau(4 + 8\log A)}, \tag{2.49}$$

*we have, for all $t \geq 1$,*

$$\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \left\| \frac{1}{\mu} \right\|_\infty \cdot \frac{1 + \tau \log A}{(1-\gamma)^2} \cdot e^{-C(t-1)}, \tag{2.50}$$

*where*

$$C = \frac{(1-\gamma)^4}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot c^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{\mu} \right\|_\infty^{-1} > 0, \tag{2.51}$$

*is independent with $t$.*

### 2.4.4 Controlling the Bias

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

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(c^2 \cdot t_2), \tag{2.52}$$

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

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

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

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

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{K}{t^{1/\alpha}} + \frac{C \cdot \log t}{\exp\left\{\sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}}, \tag{2.53}$$

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

The final rates then depend on how fast $\min_a \pi_{\theta_t}(a)$ diminishes as a function of $t$. We conjecture that the rate in some cases degenerates to $O\big(\frac{\log t}{t^{1/\alpha}}\big)$, which is strictly faster than $O(1/t)$ in the non-regularized case when $\alpha \in (0, 1)$ and is observed in simulations in Fig. 2.5. We leave it as an open problem to study decaying entropy in general MDPs.

## 2.5 A Theoretical Understanding of Entropy Regularization in Policy Gradient

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

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

### 2.5.1 Lower Bounds

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

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \leq \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r. \tag{2.54}$$

Using this result gives the desired lower bound:

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

$$(\pi^* - \pi_{\theta_t})^\top r \geq \Delta^2/(6 \cdot t). \tag{2.55}$$

Note that Theorem 9 is a special case of general MDPs. Next, we show that similar to Lemma 17, the progress in each iteration of softmax policy gradient in any MDP can be bounded in terms of sub-optimality gap.

**Lemma 18** (Reversed Łojasiewicz)**.** *Denote*

$$\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s,a) > 0, \tag{2.56}$$

*as the optimal value gap of state $s$, where $a^*(s)$ is the action that the optimal policy selects under state $s$, and $\Delta^* = \min_{s \in \mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. Then we have,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \leq \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \left[ V^*(\mu) - V^{\pi_\theta}(\mu) \right]. \tag{2.57}$$

With Lemma 18, we then strengthen Theorem 9 and show that the $\Omega(1/t)$ lower bound also holds for *any* MDP:

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

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \geq \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12 \cdot t}, \tag{2.58}$$

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

**Remark 5.** *Our convergence rates in Section 2.3 match the lower bounds up to constant. However, the constant gap is large, e.g., $K^2$ in Theorem 3, and $\Delta^2$ in Theorem 9. The gap is because the reversed Łojasiewicz inequality of Lemma 17 uses $\Delta$, which is unavoidable when $\pi_\theta$ is close to $\pi^*$. We leave it as an open problem to close this gap.*

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

### 2.5.2 Non-uniform Łojasiewicz (NŁ) Degree

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

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

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi}, \tag{2.59}$$

$\forall \theta \in \Theta$*, where* $C(\theta) > 0$ *holds for all* $\theta \in \Theta$*.*

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

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

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

Note that according to Remark 1, it is necessary that $C(\theta)$ depends on $\pi_\theta(a^*)$. The difference between Proposition 4 and the reversed Łojasiewicz inequality of Lemma 17 is subtle. Lemma 17 is a condition that implies

impossibility to get rates faster than $O(1/t)$, while Proposition 4 says it is not sufficient to get rates faster than $O(1/t)$ *using the same technique as in Lemma 4*. However, this does not preclude that other techniques could give faster rates.

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

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

## 2.6 Experimental Verification

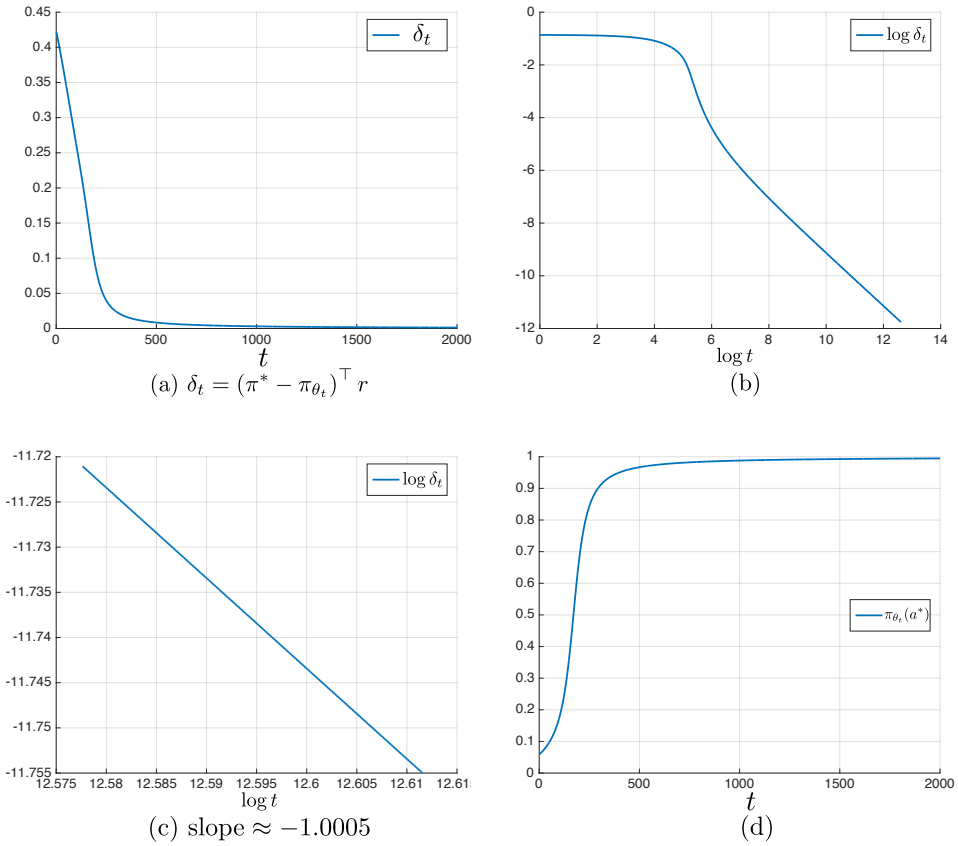
To verify the convergence rates in the chapter, we conduct experiments on one-state MDPs with $K$ actions.

### 2.6.1 Softmax Policy Gradient

$K = 20$, $r \in [0,1]^K$ is randomly generated, and $\pi_{\theta_1}$ is randomly initialized. Softmax policy gradient, i.e., Update 1 is used with learning rate $\eta = 2/5$ and maximum iteration number $T = 3 \times 10^5$. As shown in Fig. 2.2(a), the sub-optimality $\delta_t = (\pi^* - \pi_{\theta_t})^\top r$ approaches 0. Subfigures (b) and (c) show $\log \delta_t$ as a function of $\log t$. As $\log t$ increases, the slope is approaching $-1$, indicating that $\log \delta_t = -\log t + C$, which is equivalent to $\delta_t = C'/t$. Subfigure (d) shows $\pi_{\theta_t}(a^*)$ as a function of $t$.

### 2.6.2 Entropy Regularized Softmax Policy Gradient

$K = 20$, $r \in [0,1]^K$ and $\pi_{\theta_1}$ are the same as above. Entropy regularized softmax policy gradient, i.e., Update 2 is used with temperature $\tau = 0.2$, learning rate $\eta = 2/5$ and iteration number $T = 5 \times 10^4$. As shown in Fig. 2.3(a), the soft sub-optimality $\tilde{\delta}_t = {\pi^*_\tau}^\top (r - \tau \log \pi^*_\tau) - {\pi_{\theta_t}}^\top (r - \tau \log \pi_{\theta_t})$ approaches 0. Subfigure (b) shows $\log \tilde{\delta}_t$ as a function of $t$. As $t$ increases, the curve approaches a straight line, indicating that $\log \tilde{\delta}_t = -C_1 \cdot t + C_2$, which is equivalent to $\tilde{\delta}_t = C_2' / \exp\{C_1' \cdot t\}$. Subfigure (c) shows $\zeta_t$ as defined in Lemma 11 as a function of $t$, which verifies Lemma 12. Subfigure (d) shows $\min_a \pi_{\theta_t}(a)$ as




(a) $\delta_t = (\pi^* - \pi_{\theta_t})^\top r$ (b) (c) slope $\approx -1.0005$ (d)

Figure 2.2: Softmax policy gradient, Update 1.

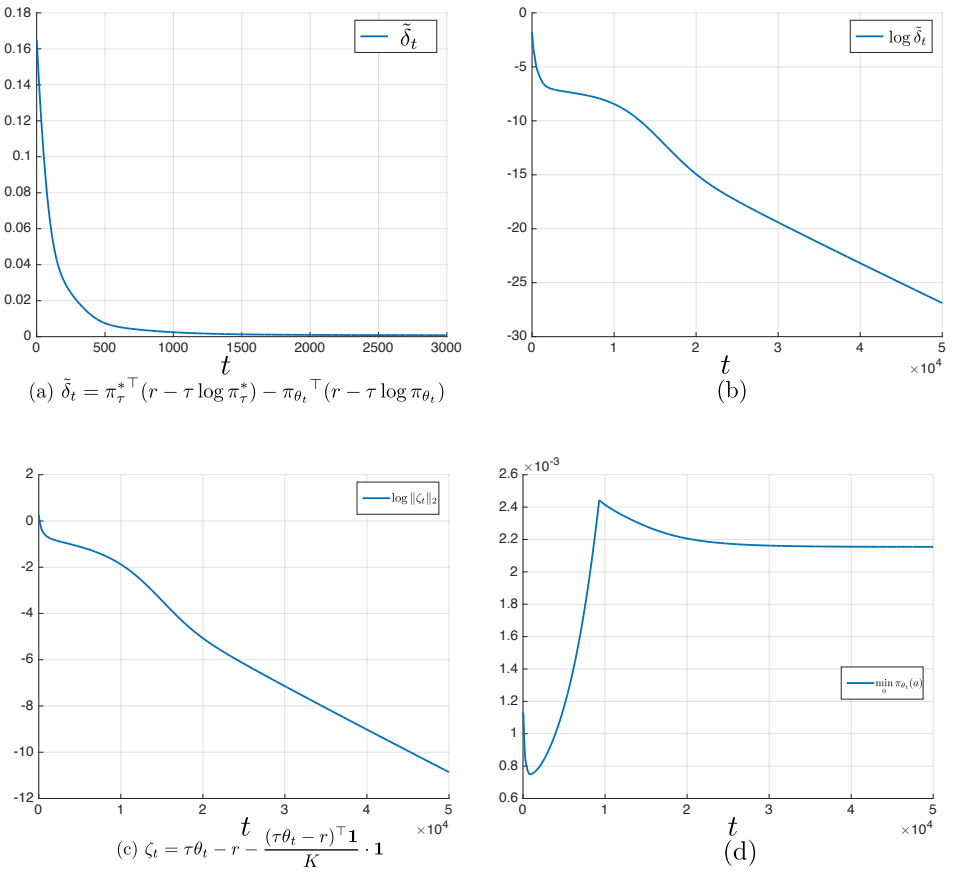
a function of $t$. As $t$ increases, $\min_a \pi_{\theta_t}(a)$ approaches constant values, which verifies Lemma 13.

### 2.6.3 "Bad" Initializations for Softmax Policy Gradient (PG)

As illustrated in Fig. 2.1, "bad" initializations lead to attraction toward sub-optimal corners and slowly escaping for softmax policy gradient. Fig. 2.4 shows one example with $K = 5$. Softmax policy gradient takes about $8 \times 10^6$ iterations around a sub-optimal corner. While with entropy regularization ($\tau = 0.2$), the convergence is significantly faster.

### 2.6.4 Decaying Entropy Regularization

We run entropy regularized policy gradient with decaying temperature $\tau_t = \frac{\alpha \cdot \Delta}{\log t}$ for $t \geq 2$, i.e., Update 3. Fig. 2.5 shows one example with $K = 10$




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

(b)

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

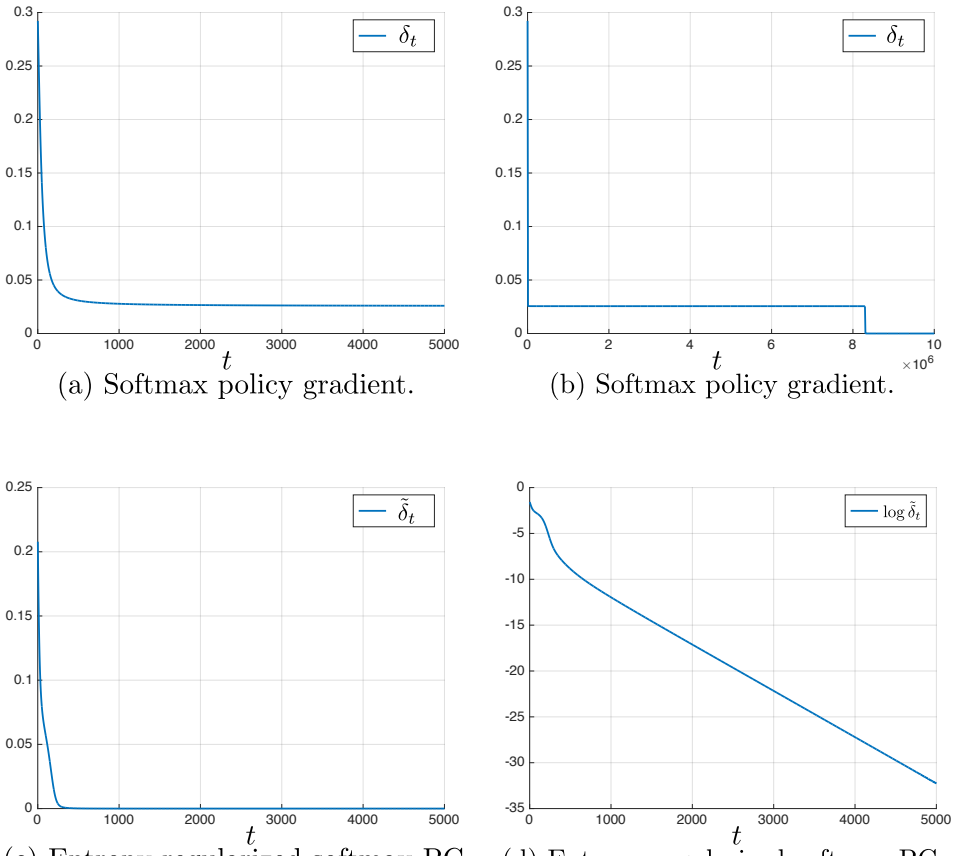
(d)

Figure 2.3: Entropy regularized softmax policy gradient, Update 2.

and different $\alpha$ values. The actual rate is $O\big(\frac{1}{t^{-\text{slope}}}\big)$, and the partial rate in Theorem 8 is $O\big(\frac{1}{t^{1/\alpha}}\big)$.

## 2.7 Summary

We set out to study the convergence speed of softmax policy gradient methods with and without entropy regularization in the tabular setting. Here, the error is measured in terms of the sub-optimality of the policy obtained after some number of updates. Our main finding is that without entropy regularization, the rate is $\Theta(1/t)$, which is faster than rates previously obtained. Our analysis also uncovered an unpleasant dependence on the initial parameter values. With entropy regularization, the rate becomes linear $O(e^{-c \cdot t})$, where now the constant in the exponent is influenced by the initial choice of parameters. Thus, our analysis shows that entropy regularization substantially changes the

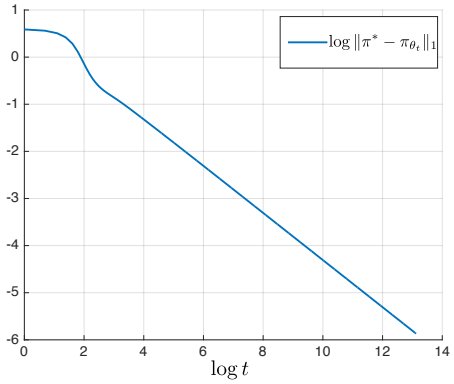



(a) Softmax policy gradient. (b) Softmax policy gradient.

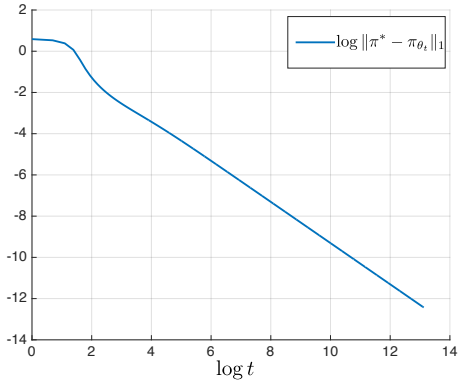
(c) Entropy regularized softmax PG. (d) Entropy regularized softmax PG.

Figure 2.4: Bad initialization for softmax policy gradient.

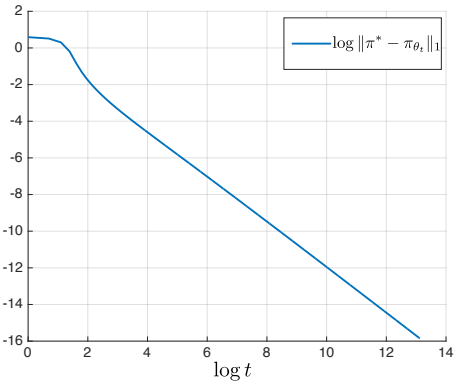
rate at which gradient methods converge. Our main technical innovation is the introduction of a new non-uniform Łojasiewicz (NŁ) inequality. The deeper reason of entropy accelerating convergence is explained using the notion of NŁ degree, which is a key quantity to characterize different NŁ inequalities.



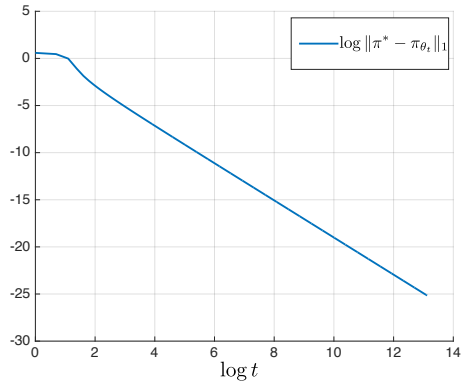

(a) $\alpha = 2.0,\ \text{slope} \approx -0.5.$




(b) $\alpha = 1.0,\ \text{slope} \approx -0.9999.$




(c) $\alpha = 0.8,\ \text{slope} \approx -1.2516.$




(d) $\alpha = 0.5,\ \text{slope} \approx -1.9222.$

Figure 2.5: Decaying entropy regularization, Update 3.

# Chapter 3

# Escaping the Gravitational Pull of Softmax

The convergence rate results we established in Chapter 2 are general and substantive. However, in this chapter we reveal a negative side of those results, and provide an explanation for the gap between theory and practice. This leads to a deeper understanding of the coefficient in non-uniform Łojasiewicz inequality.

The results in this chapter appeared in Mei et al. (2020a).

## 3.1 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 (Sutton and Barto, 2018). 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 (Friedman et al., 2001). Attention models (Vaswani et al., 2017) 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. According to Eq. (2.6), for $\theta \in \mathbb{R}^K$, the transformation $\pi_\theta = \text{softmax}(\theta)$ is defined by

$$\pi_\theta(a) = \frac{\exp\{\theta(a)\}}{\sum_{a'} \exp\{\theta(a')\}}, \quad \forall a \in [K] \coloneqq \{1, ..., K\}, \tag{3.1}$$

which ensures $\pi_\theta(a) > 0$ and $\sum_a \pi_\theta(a) = 1$ (Bridle, 1989). The softmax transform can also be extended to continuous output spaces through the concept of a Gibbs function (LeCun et al., 2006), but for concreteness we restrict our 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 (de Brébisson and Vincent, 2015; Laha et al., 2018), 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 computational overhead. This chapter seeks to fill the gap in understanding key optimization properties of probability transformations in general and how they compare to the softmax. In particular, we make the following three main contributions.

**First**, we start by investigating the softmax policy gradient (SPG) in Chapter 2. 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.

**Second**, 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. Interestingly, in the case of SCE, the vanishing NŁ coefficient leads to decreasing NŁ degree, which is introduced in Definition 1, indicating the two concepts in the NŁ inequality are not independent and can have nice interplay with each other. 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.

## 3.2 Illustrating the Softmax Gravity Wells with Softmax Policy Gradient

We begin by considering the domain of reinforcement learning (RL), using the *softmax* policy gradient (SPG) method to maximize long-term expected reward. As shown in Chapter 2, SPG enjoys a $\Theta(1/t)$ bound on the *rate of convergence*, with constants that depend on the problem and initialization.

Although the $\Theta(1/t)$ convergence rate results are general and impressive, they seem at odds with the behavior of policy gradient methods, which are

notoriously difficult to tune in practice (Schulman et al., 2015). 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.

### 3.2.1 Initialization Sensitivity

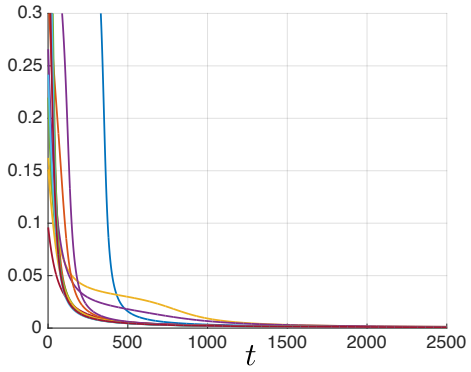
To illustrate the point concretely, consider a simple experiment on one-state Markov Decision Processes (MDPs) (i.e., a multi-armed bandit) with $K = 6$ actions. According to Update 1, in this case, the SPG of a policy $\pi_\theta$ for a given reward vector $r \in [0,1]^K$ reduces to the update

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left[r(a) - \pi_{\theta_t}^\top r\right], \tag{3.2}$$

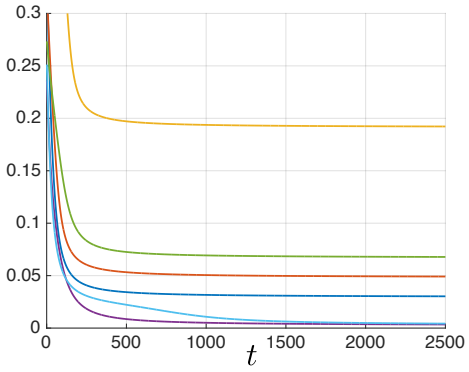
$\forall a \in [K] \coloneqq \{1, ..., K\}$, and $\pi_{\theta_{t+1}} = \text{softmax}(\theta_{t+1})$.

Fig. 3.1 shows the result of multiple runs using SPG with full gradients. Depending on whether the last iteration satisfies $\pi_{\theta_T}(a^*) \geq 0.99$, we group the 20 runs as "good" and "bad" initializations. As shown in Fig. 3.1(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 Theorem 2 that from any initializations SPG must *eventually* converge to the optimal policy $\pi^*$, and that is indeed the case here: Subfigure (d) shows the long-run time to convergence (boxes are 25 to 75th 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.

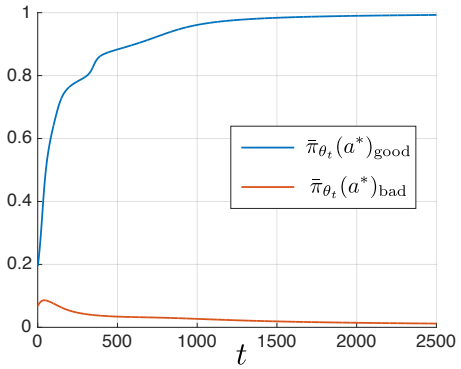
Although these findings seem not to comport with theory, they can in fact be explained by delving deeper into the detailed nature of the $\Theta(1/t)$ rates proved in Chapter 2.



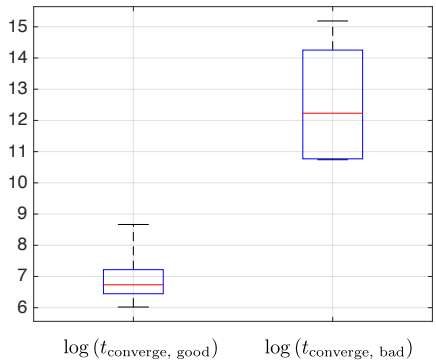

(a) "good" initializations: 14/20 runs.




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




(c) Average probability of optimal action.

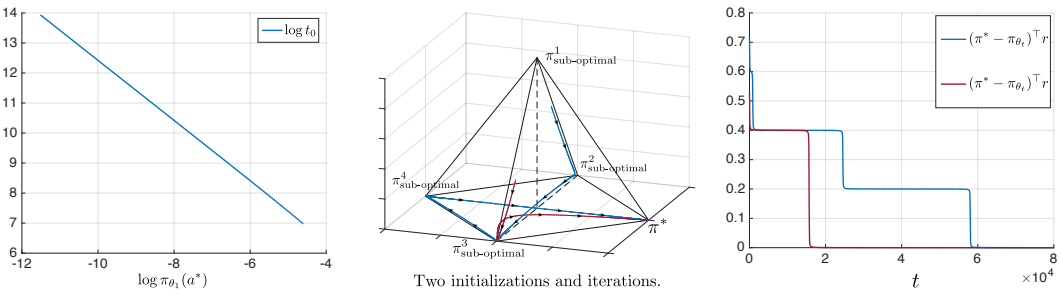



(d) Average convergence time.

Figure 3.1: SPG behavior (sub-optimality $(\pi^* - \pi_{\theta_t})^\top r$) 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 $\eta = 0.4$ from Theorem 2 for $T = 3 \times 10^4$ steps. An initialization is "good" if $\pi_{\theta_T}(a^*) \geq 0.99$ at the last iterate.

### 3.2.2 Escape Time

To control the effect of initialization, consider a specialization of the previous problem where we let $r = (b + \Delta, b, \ldots, b)^\top \in [0, 1]^K$ 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. 3.2(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 $t_0 = C'/\pi_{\theta_1}(a^*)$.




(a) Escape time $t_0$.

Two initializations and iterations.

(b) Softmax gravity wells.

(c) Multiple plateaus.

Figure 3.2: Dependence on initialization and softmax gravity wells.

### 3.2.3 Multiple Plateaus

Two trajectories for SPG on a single-state MDP with $K = 5$ is shown in Fig. 3.2(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.

### 3.2.4 Theoretical Justification

**Explaining attraction toward suboptimal deterministic policy.** 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

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = \pi_\theta(a) \cdot \left[ r(a) - \pi_\theta^\top r \right]. \tag{3.3}$$

Note that it is possible for an optimal action, $a^*$, to be less attractive than a suboptimal action $a$, even when $r(a^*) > r(a)$, since it is possible to have both

$$r(a^*) - \pi_{\theta_t}^\top r > r(a) - \pi_{\theta_t}^\top r > 0, \tag{3.4}$$

and $\pi_{\theta_t}(a) > \pi_{\theta_t}(a^*)$, and yet still have

$$\pi_{\theta_t}(a) \cdot \left[ r(a) - \pi_{\theta_t}^\top r \right] > \pi_{\theta_t}(a^*) \cdot \left[ r(a^*) - \pi_{\theta_t}^\top r \right]. \tag{3.5}$$

This configuration causes the probability on the suboptimal action to stay above the optimal action probability,

$$\theta_{t+1}(a) = \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left[r(a) - \pi_{\theta_t}^\top r\right] \tag{3.6}$$

$$\geq \theta_t(a^*) + \eta \cdot \pi_{\theta_t}(a^*) \cdot \left[r(a^*) - \pi_{\theta_t}^\top r\right] \tag{3.7}$$

$$= \theta_{t+1}(a^*), \tag{3.8}$$

and thus $\pi_{\theta_{t+1}}(a) > \pi_{\theta_{t+1}}(a^*)$, which in turn leads to Eq. (3.5) holds for $\pi_{\theta_{t+1}}$. Repeatedly, this vicious circle will finally attract the iteration toward and approach the suboptimal deterministic policy with $\pi_{\theta_t}(a) \approx 1$ for the suboptimal action $a$.

Even though the examples and analysis above might seem specific, they provide the foundation for a useful and informative lower bound.

**Theorem 11** (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 $\pi_{\theta_1}$ and a positive constant $C$, such that SPG with full gradients cannot escape a suboptimal corner before time*

$$t_0 \coloneqq \frac{C}{\Delta \cdot \pi_{\theta_1}(a^*)}, \tag{3.9}$$

*i.e., it will hold that*

$$(\pi^* - \pi_{\theta_t})^\top r \geq 0.9 \cdot \Delta, \tag{3.10}$$

*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 11 shows that for SPG with bounded learning rates, 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.

**Existing observations of plateaus.** SPG plateaus have been observed in the literature. Previous work including Chen et al. (2019) and Chapter 2 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 (Schaul et al., 2019). 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 (Hennes et al., 2019). However, the analysis in this chapter shows that SPG still suffers from plateaus even in the simplest setting (exact gradients, no approximation, stationary environments). In Section 3.4 we provide additional mathematical insight to explain why the softmax transformation itself is the root cause, which also justifies the name SGW.

## 3.3 Escort Transform for Policy Gradient

### 3.3.1 Escort Transform

As explained, a difficulty encountered by SPG comes from the $\pi_\theta(a)$ factor that appears in the gradient, Eq. (3.3). 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 11). Therefore, we study the following alternative transform, which we refer to as the "escort transform" (Beck and Schögl, 1995; Tsallis et al., 1998).

**Definition 2** (Escort transform)**.** *Given $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define $\pi_\theta = f_p(\theta)$ for $p \geq 1$ by*

$$\pi_\theta(a|s) = \frac{|\theta(s,a)|^p}{\sum_{a'} |\theta(s,a')|^p}. \tag{3.11}$$

*If there is only one state, the escort transform is defined as*

$$\pi_\theta(a) = \frac{|\theta(a)|^p}{\sum_{a'} |\theta(a')|^p}, \qquad \forall a \in [K]. \tag{3.12}$$

### 3.3.2 Escort Policy Gradient

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, $\forall a \in [K]$:

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = p \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r\right] \tag{3.13}$$

$$= \frac{p}{\|\theta\|_p} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a)^{1-1/p} \cdot \left[r(a) - \pi_\theta^\top r\right]. \tag{3.14}$$

Note the key difference between SPG and EPG, in which the $\pi_\theta(a)$ term in Eq. (3.3) now becomes $\pi_\theta(a)^{1-1/p}$ in Eq. (3.14). 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 6.** *$\pi_\theta(a)^{1-1/p} \to \pi_\theta(a)$ 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 $\pi_\theta$.*

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

**Proposition 6.** *$\theta \mapsto \pi_\theta^\top r$ 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 (NS) property 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.

**Lemma 19** (Non-uniform smoothness)**.** *Suppose $r \in [0, 1]^K$. Let $\pi_\theta \coloneqq f_p(\theta)$, and $\pi_{\theta'} \coloneqq f_p(\theta')$. Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. Then, we have,*

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_\zeta\|_p^2} \cdot \|\theta' - \theta\|_2^2. \tag{3.15}$$

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{K}{\|\theta_\zeta\|_1^2} \cdot \|\theta' - \theta\|_2^2. \tag{3.16}$$

**Lemma 20** (Non-uniform Łojasiewicz)**.** *Let $\pi_\theta = f_p(\theta)$. For any $p > 0$, we have,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \frac{p}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r, \tag{3.17}$$

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

With the NS property and NŁ inequality, we can prove a $O(1/t)$ global convergence rate for EPG using similar techniques in SPG analysis Chapter 2.

**Theorem 12.** *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 gap for all $t \geq 1$:*

**(gradient flow)** *for $p \geq 1$, with $\eta_t = \|\theta_t\|_p^2$,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{1}{c^{2-2/p} \cdot t + 1}, \tag{3.18}$$

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{9 \cdot K^{1/p}}{c^{2-2/p}} \cdot \frac{1}{t}, \tag{3.19}$$

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{9K}{t}, \tag{3.20}$$

*where $c \coloneqq \inf_t \pi_{\theta_t}(a^*) > 0$ is a problem- and initialization-dependent, but time-independent constant.*[1]

When $p$ is very large positive values, Theorem 12 implies a close to $O(1/(c^2 \cdot t))$ convergence rate, recovering the same rate for SPG Theorem 2, as expected (Remark 6). 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 6.

Similar results can in fact be obtained for EPG in general MDPs.

[1] 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 $\frac{d\theta_t}{dt} = \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$.

**Theorem 13.** *Following the escort policy gradient with any initialization such that* $|\theta_1(s,a)| > 0$, $\forall(s,a)$ *to get* $\{\theta_t\}_{t\geq 1}$, *for any* $t \geq 1$, *the following upper bounds hold for* $\pi_{\theta_t}$,

**(i)** *for* $p \geq 2$, *with* $\eta_t = \frac{(1-\gamma)^3}{10\cdot p^2\cdot A^{1/p}}$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{20\cdot A^{1/p}\cdot S}{c^{2-2/p}\cdot(1-\gamma)^6\cdot t}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2\cdot\left\|\frac{1}{\mu}\right\|_\infty, \tag{3.21}$$

**(ii)** *for* $p = 1$, *with* $\eta_t = \frac{(1-\gamma)^3}{10\cdot A}$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{20\cdot A\cdot S}{(1-\gamma)^6\cdot t}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2\cdot\left\|\frac{1}{\mu}\right\|_\infty, \tag{3.22}$$

*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 method.*

**Remark 7.** *Using* $p = 1$ *in Theorem 13, the iteration complexity of EPG depends on polynomial functions of* $S$ *and* $A$, *which significantly improves the corresponding results for SPG Theorem 4, where the worst case dependence can be exponential in* $S$ *and* $A$.

### 3.3.3 Entropy Regularization

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 $O(e^{-c^2\cdot t})$ with dependence on $c = \inf_{t\geq 1}\min_{(s,a)}\pi_{\theta_t}(a|s)$ Theorem 6. Our results show that EPG with entropy regularization has strictly better constant dependence than SPG.

**Theorem 14.** *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*

$$\eta_t = \frac{(1-\gamma)^3}{10\cdot p^2\cdot A^{1/p} + c_\tau}, \tag{3.23}$$

*to get* $\{\theta_t\}_{t\geq 1}$*, for all* $t \geq 1$*, the following sub-optimality upper bounds hold for* $\pi_{\theta_t}$*, for* $p \geq 2$*:*

$$\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \frac{\|1/\mu\|_\infty}{\exp\{C_\tau \cdot c'^2 \cdot t\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2}\,, \tag{3.24}$$

*where* $c' > c \coloneqq \inf_{(s,a)} \inf_{t\geq 1} \pi_{\theta_t}(a|s) > 0$*,* $\tau$ *is the temperature for entropy regularization,* $\pi^*_\tau$ *is the softmax optimal policy, and* $c_\tau$*,* $C_\tau$ *are problem-dependent constants.*

### 3.3.4 Relationship to Mirror Descent (MD)

As an additional observation, note that simply removing $\pi_\theta(a)$ in Eq. (3.3) yields an update $\theta_{t+1} = \theta_t(a) + \eta_t \cdot r(a)$ and $\pi_{\theta_{t+1}} = \text{softmax}(\theta_{t+1})$, which can be combined to yield an update

$$\pi_{\theta_{t+1}}(a) = \frac{\pi_{\theta_t}(a) \cdot \exp\{\eta_t \cdot r(a)\}}{\sum_{a' \in [K]} \pi_{\theta_t}(a') \cdot \exp\{\eta_t \cdot r(a')\}}, \tag{3.25}$$

that is equivalent to Mirror Descent (MD) with KL divergence. 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 8** (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*

$$\pi_{t+1} = \arg\max_{\pi \in \Delta} \; \pi^\top r - \frac{1}{\eta_t} \cdot D_\Phi(\pi \| \pi_t). \tag{3.26}$$

*In particular, using* $\Phi(\pi) = \pi^\top \log \pi$ *as the potential and* $D_\Phi(\pi\|\pi') = D_{\text{KL}}(\pi\|\pi')$ *as the divergence one obtains* $\pi_{\theta_{t+1}}(a) \propto \pi_{\theta_t}(a) \cdot \exp\{\eta_t \cdot r(a)\}$*. Equivalently, this update can be expressed*

$$\pi_{\theta_{t+1}} = \arg\max_{\pi \in \Delta} \pi^\top \theta_{t+1} - \Phi(\pi), \tag{3.27}$$

*where* $\theta_{t+1} = \theta_t(a) + \eta_t \cdot r(a)$*.*

*Now suppose EPG is MD, i.e., there is some $\Phi$ such that*

$$f_p(\theta_{t+1}) = \arg\max_{\pi \in \Delta} \pi^\top \theta_{t+1} - \Phi(\pi). \tag{3.28}$$

*Then we would have to have $f_p(\theta_{t+1}) = \nabla \Phi^*(\theta_{t+1})$ where $\Phi^*$ is the Fenchel conjugate of $\Phi$. Taking the derivative w.r.t. $\theta$ yields*

$$\left(\frac{d\pi_\theta}{d\theta}\right)^\top = \left(\frac{df_p(\theta)}{d\theta}\right)^\top = p \cdot diag(1/\theta)\left(diag(\pi_\theta) - \pi_\theta \pi_\theta^\top\right) \overset{(?)}{=} \frac{d^2\Phi^*(\theta)}{d\theta^2}. \tag{3.29}$$

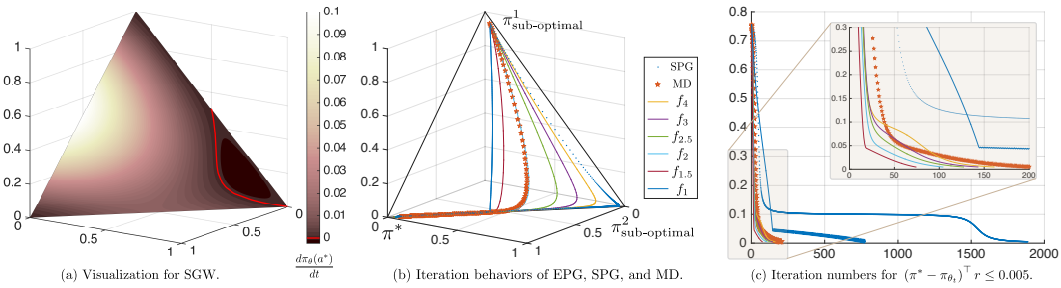
*By Schwarz's theorem, $\frac{d^2\Phi^*(\theta)}{d\theta^2}$ is symmetric, however $diag(1/\theta)\left(diag(\pi_\theta) - \pi_\theta \pi_\theta^\top\right)$ is not symmetric. Therefore, there cannot be a regularizer $\Phi$ that makes EPG equivalent to MD.*

Remark 8 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 12 to 14.

### 3.3.5 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. 3.3(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 11. 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 6, 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 divergence) has similar performance to EPG with $p = 2$ in this case. Finally, subfigure (c) plots the sub-optimaliy 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 3.3: Empirical visualization for EPG and SPG.

## 3.4 Non-uniform Łojasiewicz Coefficient: An Underlying Explanation

Remark 6 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 in Chapter 2 and EPG analyses are based on NŁ inequalities.

**Definition 3** (Non-uniform Łojasiewicz (NŁ) coefficient)**.** *A function $f : \Theta \to \mathbb{R}$ has NŁ coefficient $C(\theta) > 0$ if it satisfies NŁ inequality with coefficient $C(\theta)$, i.e., there exists $\xi \in [0, 1]$ such that for all $\theta \in \Theta$,*

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi}. \tag{3.30}$$

Recall that in Definition 1, $\xi$ is called NŁ degree, which impacts the convergence rates of SPG methods as shown in Section 2.5. According to the result of Lemma 3, if $\pi_\theta = \text{softmax}(\theta)$, then $\pi_\theta^\top r$ has NŁ coefficient $\pi_\theta(a^*)$; that is

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left\| \left( \text{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) r \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \tag{3.31}$$

Moreover, this coefficient is not improvable (Remark 1) and it appears in the SPG convergence rate $O(1/(c^2 \cdot t))$ (Theorem 2), where $c \coloneqq \inf_{t \geq 1} \pi_{\theta_t}(a^*)$ (Lemma 5). Now consider EPG. If $\pi_\theta = f_p(\theta)$, then we have

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left\| p \cdot \operatorname{diag}(1/\theta) \left( \operatorname{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) r \right\|_2 \tag{3.32}$$

$$\geq p \cdot \frac{\pi_\theta(a^*)}{|\theta(a^*)|} \cdot (\pi^* - \pi_\theta)^\top r \tag{3.33}$$

$$= \frac{1}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r, \tag{3.34}$$

where $\pi_\theta(a^*)^{1-1/p} > \pi_\theta(a^*)$ provides strictly larger (partial) NŁ coefficient, hence in Theorem 12 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.

## 3.5 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:

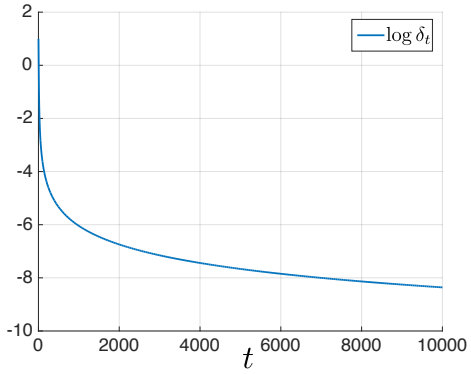
$$\min_{\theta: \mathcal{A} \to \mathbb{R}} -\log \pi_\theta(a_y) = \mathcal{H}(y) + \min_{\theta: \mathcal{A} \to \mathbb{R}} D_{\mathrm{KL}}(y \| \pi_\theta), \tag{3.35}$$

where $\pi_\theta = \operatorname{softmax}(\theta)$, $y \in \{0, 1\}^K$ 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 $\mathcal{H}(y) \coloneqq -y^\top \log y = 0$ here. The objective in Eq. (3.35) is smooth and convex in $\theta$, which implies that gradient descent will achieve an $O(1/t)$ rate (Nesterov, 2018). Furthermore, for $\theta$ that satisfies $\min_a \pi_\theta(a) \geq \pi_{\min}$ with some constant $\pi_{\min} > 0$ ($\pi_\theta$ is bounded away from the simplex boundary), the objective is strongly convex, resulting in an even better, linear rate $O(e^{-c \cdot t})$.

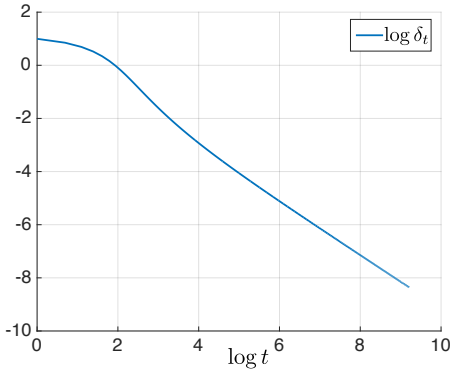
Despite these nice properties, we still find that the softmax transform proves problematic for gradient descent optimization. We refer to this new disadvantage as "softmax damping".

### 3.5.1 Softmax Damping

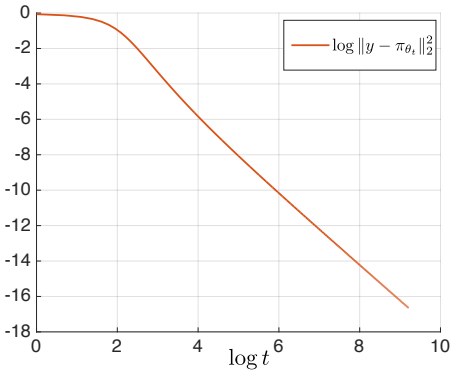
Consider running gradient descent in a simple experiment where $K = 10$ and $y$ is a one-hot vector. Let $\delta_t := -\log \pi_{\theta_t}(a_y)$. If one hopes for a linear convergence rate, i.e., $\delta_t = O(e^{-c \cdot t})$, then $\log \delta_t = -O(t)$ is expected. But Fig. 3.4(a) shows a different 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 $O(e^{-c \cdot t})$ rate. Subfigure (c) shows the $\ell_2$ measure $\|y - \pi_{\theta_t}\|_2^2$ also has a sublinear rate, indicating that this is an inherent optimization phenomenon and is independent of the measurement.



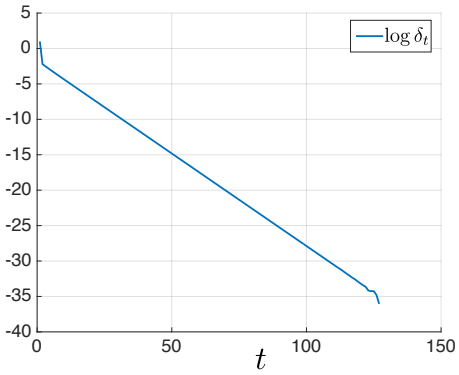

(a) Softmax damping: degenerating rates.




(b) Softmax cross entropy.




(c) $\ell_2$ measure for SCE.




(d) Escort cross entropy.

Figure 3.4: Softmax damping phenomenon and escort cross entropy.

### 3.5.2 NŁ Coefficient Explanation

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

$$\left\| \frac{d\{D_{\text{KL}}(y\|\pi_\theta)\}}{d\theta} \right\|_2^2 = \|y - \pi_\theta\|_2^2 \geq \min_a \pi_\theta(a) \cdot D_{\text{KL}}(y\|\pi_\theta). \tag{3.36}$$

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 $D_{\text{KL}}(y\|\pi) = -\log(1 - \epsilon) \geq \epsilon$ and $\|y - \pi\|_2^2 = 2 \cdot \epsilon^2$, which means for any constant $C > 0$, we have

$$C \cdot D_{\text{KL}}(y\|\pi) \geq C \cdot \epsilon > 2 \cdot \epsilon^2 = \|y - \pi\|_2^2. \tag{3.37}$$

Therefore, for any Łojasiewicz-type inequality, $C$ necessarily depends on $\min_a \pi_\theta(a)$. Now for any convergent sequence $\pi_{\theta_t}$, i.e., such that $D_{\text{KL}}(y\|\pi_{\theta_t}) \to 0$, 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.

Moreover, we can explain why the rate degenerates to $O(1/t)$ asymptotically as $t \to \infty$, using the interplay between the NŁ coefficient and NŁ degree. Consider the same example where $\pi_\theta = (\epsilon, 1 - \epsilon)^\top$ and $D_{\text{KL}}(y\|\pi_\theta) = -\log(1 - \epsilon)$. Note that we have $e^{-2 \cdot x} \leq 1 - x$ for all $x \in [0, 1/2]$. Then we have,

$$\begin{aligned}
\left\| \frac{d\{D_{\text{KL}}(y\|\pi_\theta)\}}{d\theta} \right\|_2^2 &\geq \min_a \pi_\theta(a) \cdot D_{\text{KL}}(y\|\pi_\theta) \qquad \text{(by Eq. (3.36))} && (3.38)\\
&= \frac{1}{2} \cdot 2 \cdot \epsilon \cdot D_{\text{KL}}(y\|\pi_\theta) && (3.39)\\
&\geq -\frac{1}{2} \cdot \log(1 - \epsilon) \cdot D_{\text{KL}}(y\|\pi_\theta) \qquad (\epsilon \in [0, 1/2]) && (3.40)\\
&= \frac{1}{2} \cdot D_{\text{KL}}(y\|\pi_\theta)^2, && (3.41)
\end{aligned}$$

which means the NŁ degree becomes $\xi = 0$ as $\min_a \pi_\theta(a) \to 0$, according to Definition 1. Comparing to Eq. (3.36) with $\xi = 1/2$, this is a strictly weaker NŁ inequality and can only lead to a $O(1/t)$ convergence rate.

### 3.5.3 Label smoothing, soft target, reward-augmented maximum likelihood

As shown in Eq. (3.36), the reason for softmax damping happens is $\min_a \pi_\theta(a) \to 0$ as $\pi_\theta \to y$ if $y$ is a one-hot distribution. One might consider non-deterministic target $y$ to avoid degenerating convergence since we would have $\min_a \pi_\theta(a) \to \min_a y(a) > 0$. In fact, there exist several existing work implementing this idea, including label smoothing (Szegedy et al., 2016), soft target (Hinton et al., 2015), and reward augmented maximum likelihood (Norouzi et al., 2016). Those techniques are usually considered to have generalization benefits. Here we provide an optimization advantage, which is a byproduct of our NŁ coefficient explanation. For example, instead of using a one-hot true label distribution $y$ in $D_{\mathrm{KL}}(y\|\pi_\theta)$, label smoothing has a regularized target as

$$y_{\mathrm{LS}} \coloneqq (1-\alpha)\cdot y + \alpha \cdot \frac{1}{K}, \tag{3.42}$$

where $K$ is the total number of classes, and $\alpha > 0$ is the label smoothing hyperparameter. It is then obvious that $\min_a \pi_\theta(a) \to \min_a y_{\mathrm{LS}}(a) = \alpha/K > 0$, and the softmax damping will not happen.

### 3.5.4 Escort Cross Entropy

As in Section 3.3 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

$$\left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta} \right\|_2^2 = \|p \cdot \mathrm{diag}(1/\theta)(y - \pi_\theta)\|_2^2 \tag{3.43}$$

$$\geq \frac{p^2}{\|\theta\|_p^2} \cdot \min_a \pi_\theta(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_\theta). \tag{3.44}$$

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 $\pi_\theta$. This leads to our last main result, which restores the linear convergence rate.

**Theorem 15.** *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 $t \geq 1$,*

$$-\log \pi_{\theta_t}(a_y) = D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \leq D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \cdot \exp\Big\{-\frac{(t-1)}{2\cdot(3+c_1^2)}\Big\}, \tag{3.45}$$

*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. 3.4(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 15.

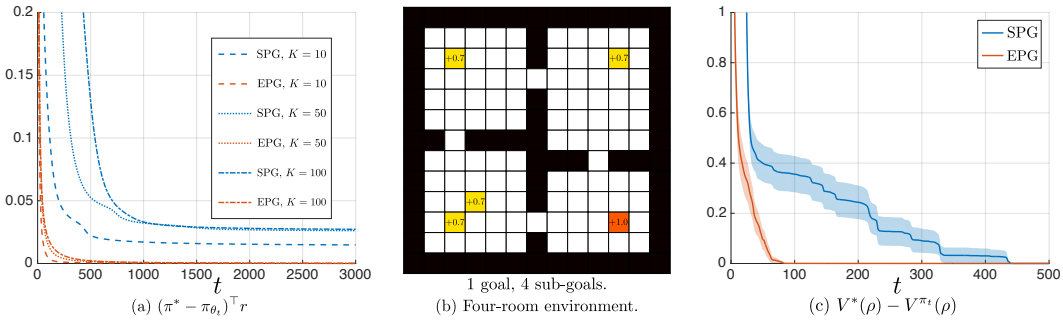
## 3.6 Experimental Results

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

### 3.6.1 One-state MDPs

First, we conduct experiments on one-state MDPs with $K = 10$, 50, and 100. 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 $\pi_{\theta_1}$ within the $(K-1)$-dimensional probability simplex. SPG and EPG start from the same initial policy $\pi_{\theta_1}$. The total number of iterations is $T = 5 \times 10^4$.

Fig. 3.5 shows the results of SPG and EPG with $p = 2$. The learning rate of SPG is set to be $\eta = 0.4$ (Theorem 2). The learning rate of EPG is $\eta_t = 0.2 \cdot \|\theta_t\|_p^2$ (Theorem 12). As shown in Fig. 3.5(a), EPG with $p = 2$ quickly converges to optimal policies consistently across all the $K$ values, significantly outperforming SPG.




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

1 goal, 4 sub-goals.
(b) Four-room environment.

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

Figure 3.5: Results on one-state MDPs and Four-room.

**Different $p$ values.** Fig. 3.6(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.

### 3.6.2 Four-room

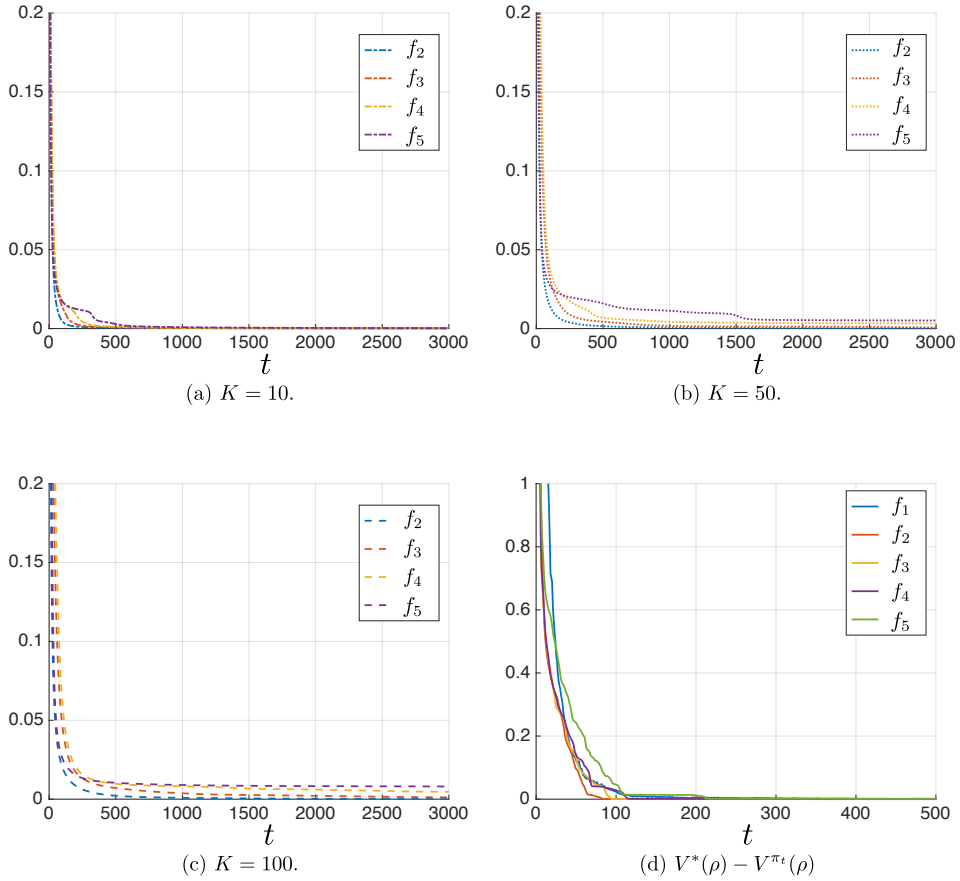
Second, we compare the algorithms on Four-room environment for 20 runs. There is one goal with reward 1.0 and 4 sub-goals ("sub-goals" mean goals with lower rewards) with reward 0.7 as shown in Fig. 3.5(b). At a (sub-)goal state, the agent can step away then step back to receive rewards. The policy is $\pi_\theta = \text{softmax}(\theta)$ for SPG, and $\pi_\theta = f_p(\theta)$ for EPG, and $\theta$ is the output of one parameterized by one hidden layer neural network with ReLU activation function and 64 hidden nodes.

In each run, the starting position is randomly generated. The optimal value function $V^*$ is approximately calculated using value iteration with threshold of two consecutive iterations $\|V_t - V_{t+1}\|_2^2 \leq 1 \times 10^{-10}$. 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 (Kingma and Ba, 2014) and the total number of iterations is $T = 500$.

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 rate 0.01 is used for both SPG and EPG as a result of searching within $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.

Fig. 3.5 shows the results of SPG and EPG with $p = 2$. As shown in Fig. 3.5(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.

**Different $p$ values.** Fig. 3.6(d) shows the results of EPG for $p \in \{1, 2, 3, 4, 5\}$ in Four-room task, where each curve is the averaged result of 20 runs.




(a) $K = 10$. (b) $K = 50$.

(c) $K = 100$. (d) $V^*(\rho) - V^{\pi_t}(\rho)$

Figure 3.6: Results of EPG with different $p$ values on one-state MDPs and Four-room.

### 3.6.3 MNIST

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 misclassification probability minimization problem is a special case of expected reward maximization. The policy is $\pi_\theta = \text{softmax}(\theta)$ for SPG and SCE, and $\pi_\theta = f_p(\theta)$ for EPG and ECE, where $\theta$ is the output of one hidden layer neural network with ReLU activation function and 512 hidden nodes. The dataset is split into training set with 55000, validation set with 5000, and testing set with 10000 data points.

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. 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. 3.7(a) and (b) show the results of SPG and EPG with $p = 4$. For both training objective and test error, SPG has plateaus due to SGWs, which is consistent with the observation in (Chen et al., 2019). At the same time, EPG with $p = 4$ does not have this disadvantage: it converges quickly and achieves better results than SPG.

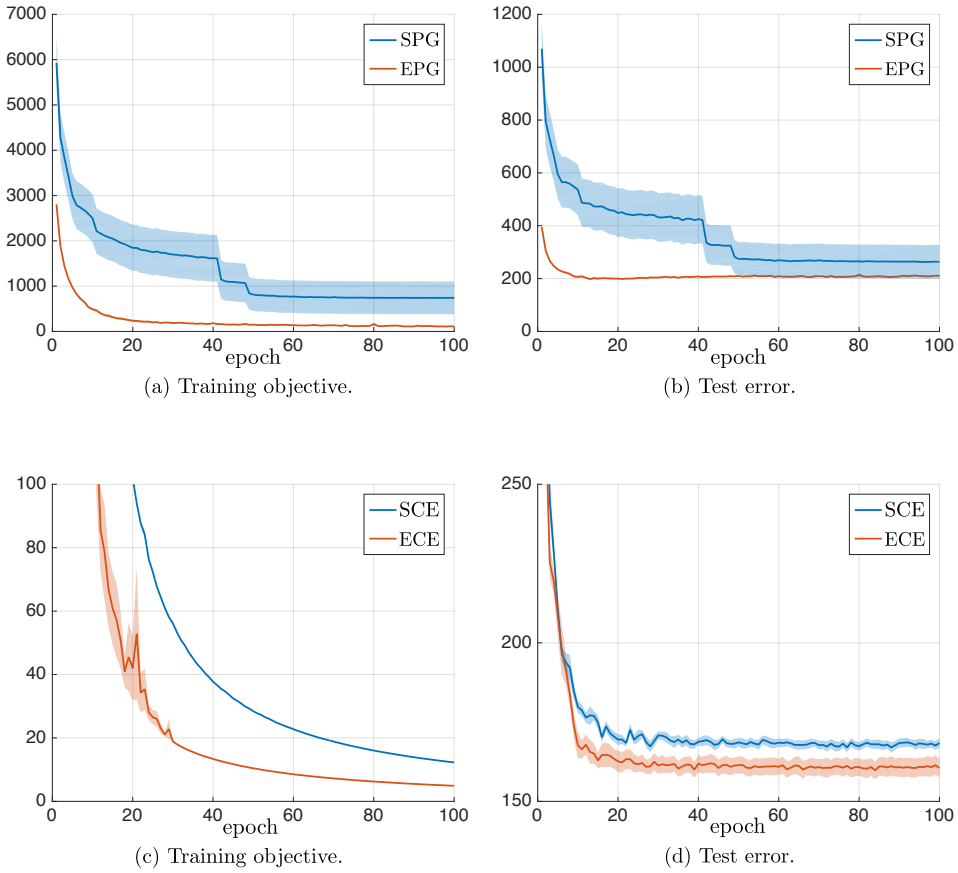
The results show that with stochastic gradients and neural network function approximations, *(i)* SPG still plateaus even when starting from nearly uniform initializations; *(ii)* EPG outperforms SPG in terms of not suffering from plateaus even with estimated gradients.

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 are the same as above. The learning rate and batchsizes 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. As shown in Fig. 3.7(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.

**Different $p$ values.** Fig. 3.8 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 $p = 5$.

### 3.6.4 Comparing SPG, EPG, and MD

As noted in Remark 8, EPG cannot be reduced to MD with any regularizer. Also as shown in Fig. 3.3(b), EPG and MD with KL divergence behave similarly in the 3-action case. We conduct experiments on bandit problems with $K \in \{50, 100, 500\}$ actions to compare EPG with MD. In each iteration, all




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

(c) Training objective. (d) Test error.

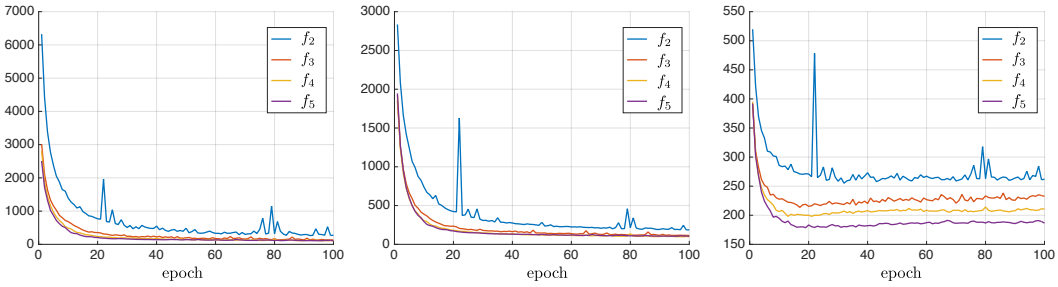
Figure 3.7: Results on MNIST.

the algorithms use the same stochastic gradient to do updates. Each curve is averaged over 20 runs.

As shown in Fig. 3.9, EPG and MD with KL regularization have comparable performances, significantly outperforming SPG. However, EPG in its nature is a policy gradient method, which has a cheap update in each iteration, while MD needs to solve an optimization problem to do one update.

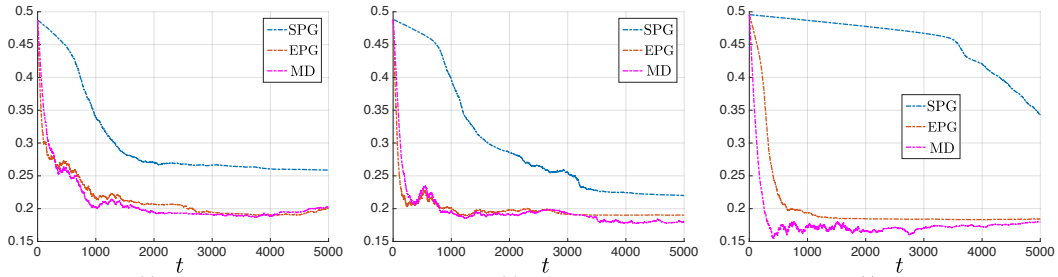
## 3.7 Summary

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




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

Figure 3.8: Results of EPG with different $p$ values on MNIST.




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

Figure 3.9: Sub-optimality $(\pi^* - \pi_{\theta_t})^\top r$ on single-state MDPs using stochastic gradients.

in machine learning. The key technical innovation is to use the concept of non-uniform Łojasiewicz (NŁ) coefficient to characterize different NŁ inequalities. This goes beyond the classic convex "matching loss" theory (Auer et al., 1996; Kivinen and Warmuth, 1998) and guarantees better optimization results.

# Chapter 4

# Non-uniform Analysis

In Chapters 2 and 3, the non-uniform Łojasiewicz inequality was introduced to study the policy gradient optimization (PG) in RL. In this chapter, inspired by the PG results as well as other non-convex optimization problems in machine learning, I unify and generalize the previous non-uniform properties, and propose a non-uniform analysis for general optimization, then apply it to both PG in RL and generalized linear model training (GLM) in SL.

The results in this chapter appeared in Mei et al. (2021b).

## 4.1 Introduction

The optimization of non-convex objective functions is a topic of key interest in modern-day machine learning. Recent, intriguing results show that simple gradient-based optimization can achieve *globally* optimal solutions in certain non-convex problems arising in machine learning, such as in reinforcement learning (RL) (Chapter 2), supervised learning (SL) (Hazan et al., 2015), and deep learning (Allen-Zhu et al., 2019). While gradient-based algorithms remain the method of choice in machine learning, the convergence of such algorithms to global minimizers has still only been established in restrictive settings where one can assert two strong assumptions about the objective function: *(i)* that the objective is smooth, and *(ii)* that the objective satisfies a gradient dominance over sub-optimality such as the Łojasiewicz inequality. We will find it beneficial to recall the definitions of these properties. For the remainder of this chapter let $\Theta = \mathbb{R}^d$.

**Definition 4** (Smoothness)**.** *The function $f : \Theta \to \mathbb{R}$ is $\beta$-smooth ($\beta > 0$) if it is differentiable and for all $\theta, \theta' \in \Theta$,*

$$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2. \tag{4.1}$$

**Definition 5** (Kurdyka (1998), Łojasiewicz (1963), and Polyak (1963))**.** *The differentiable function $f : \Theta \to \mathbb{R}$ satisfies the $(C, \xi)$-Łojasiewicz inequality if for all $\theta \in \Theta$,*

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C \cdot \left( f(\theta) - \inf_{\theta \in \Theta} f(\theta) \right)^{1-\xi}, \tag{4.2}$$

*where $C > 0$ and $\xi \in [0, 1]$.*

In particular, if an objective function $f$ satisfies both assumptions, gradient-based optimization can be shown to converge to a global minimizer by noting first that uniform smoothness Eq. (4.1) ensures the gradient updates achieve monotonic improvement with an appropriate step size, i.e., we have, for $\theta_{t+1} \leftarrow \theta_t - \frac{1}{\beta} \cdot \nabla f(\theta_t)$,

$$f(\theta_{t+1}) \leq f(\theta_t) - \frac{1}{2\beta} \cdot \|\nabla f(\theta_t)\|_2^2 . \tag{4.3}$$

while the Łojasiewicz inequality Eq. (4.2) ensures the gradient does not vanish before a global minimizer is reached. Several global convergence results have recently been achieved in the machine learning literature by exploiting assumptions of this kind. For example, in reinforcement learning it has recently been shown that policy gradient (PG) methods converge to a globally optimal policy (Chapter 2); in supervised learning it has been shown that gradient descent (GD) methods converge to global minimizers of certain non-convex problems (Hazan et al., 2015); and in deep learning theory it has been shown that (stochastic) GD can converge to a global minimizer with an over-parameterized neural network (Allen-Zhu et al., 2019).

However, previous work that relies on the two *uniform* conditions in Definitions 4 and 5 assumes *universal constants* $\beta$ and $C$, which ignores important problem structure and limits both the applicability of the results and the strength of the results that can be obtained.

In this chapter, we expand the class of problems for which gradient-based optimization is globally convergent, develop novel gradient-based methods that better exploit local structure, and improve the convergence rate analysis. We achieve these results by first defining then investigating a new set of *non-uniform* smoothness and Łojasiewicz inequalities, which generalize the classical definitions and allow a refined characterization of the space of objectives. Given these refined notions, we then tailor novel gradient-based algorithms that improve previous methods for these new problem classes, and extend the analysis to exploit these new forms of non-uniformity, achieving significantly stronger convergence rates in many cases. Importantly, these improvements are achieved in non-convex optimization problems that arise in relevant machine learning problems.

The remainder of this chapter is organized as follows.

**First**, in Section 4.2 we illustrate how natural optimization problems, including those in machine learning, exhibit interesting *local* structure that cannot be adequately captured by the uniform smoothness and Łojasiewicz inequalities.

**Second**, Section 4.3 introduces the the Non-uniform Smoothness (NS) property and the Non-uniform Łojasiewicz (NŁ) inequality, based on which Section 4.5 provides non-uniform analyses.

**Finally**, Sections 4.6 and 4.7 then present new results for policy gradient and generalized linear models respectively.

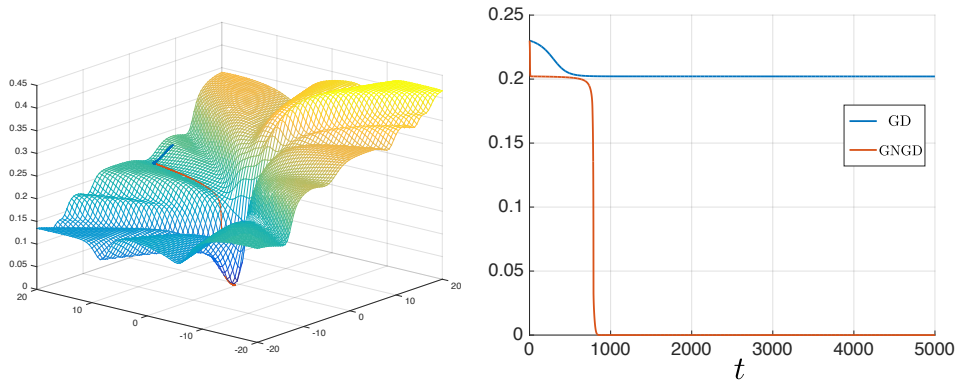
## 4.2 Motivation

To illustrate the significance of non-uniformity in machine learning problems, we consider examples motivated by recent theoretical (Wilson et al., 2019; Zhang et al., 2019) (and Chapter 2) and empirical studies (Cohen et al., 2021).

Regarding smoothness, it is clear that a uniform smoothness constant $\beta$ cannot always adequately characterize an objective over its entire domain. For example, the convex function $f : x \mapsto x^4$ cannot be informatively characterized by a uniform smoothness constant $\beta$ because its Hessian $f'' : x \mapsto 12 \cdot x^2$ has

the property that $f''(x) \to \infty$ as $|x| \to \infty$, and $f''(x) \to 0$ as $|x| \to 0$. Varying smoothness of this kind has motivated the study of alternative definitions to explain, for example, the effectiveness of gradient clipping in training neural networks and normalization in optimization (Wilson et al., 2019; Zhang et al., 2019). Meanwhile Cohen et al. (2021) present neural network training results that cannot be well explained using the standard smoothness condition of Definition 4.

Regarding the Łojasiewicz inequality, our study of policy gradient optimization in Chapter 2 has shown that, with the standard softmax parameterization, the expected return objective cannot satisfy *any* Łojasiewicz inequality with a universal constant $C$ (Remark 1), which removes the possibility of using (Definition 5) to prove convergence. By introducing a *non-uniform* version of the Łojasiewicz inequality (Lemmas 3 and 8), we were able to show a global convergence rate for the same problem.




(a) MSE landscape in GLM (b) Sub-optimality $\mathcal{L}(\theta_t)$

Figure 4.1: Non-uniform landscape of non-convex function.

Fig. 4.1 illustrates another example of a non-convex objective, which arises in supervised learning. Subfigure (a) visualizes the mean squared error (MSE) of a generalized linear model (GLM) (Hazan et al., 2015), which is not only non-convex but also highly non-uniform. As a "teaser", subfigure (b) compares the convergence behavior of two algorithms: standard gradient descent (GD), which suffers from slow convergence on the plateaus due to the non-uniformity of the objective, and an alternative algorithm (GNGD), soon to be introduced. This figure previews how proper handling of non-uniformity in

the optimization landscape can enable significant acceleration of optimization progress, including a quick escape from landscape plateaus.

## 4.3 Non-uniform Properties

The main results in this chapter depend on two core concepts, Non-uniform Smoothness (NS) and Non-uniform Łojasiewicz (NŁ) inequality. The NS property is a new, intuitive generalization of smoothness. The NŁ inequality is a new proposal that generalizes previous Łojasiewicz inequalities as well as special NŁ inequalities in Chapter 2. Our key contribution is to show that the *combination* of these two non-uniform concepts is particularly powerful, applicable to important non-convex objectives in machine learning, and allows the development of improved algorithms and analysis.

### 4.3.1 Non-uniform Smoothness (NS)

The first main concept we leverage is a new generalized notion of smoothness that depends on the parameters non-uniformly.

**Definition 6** (Non-uniform Smoothness (NS))**.** *The function $f : \Theta \to \mathbb{R}$ satisfies $\beta(\theta)$ non-uniform smoothness if $f$ is differentiable and for all $\theta, \theta' \in \Theta$,*

$$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2, \tag{4.4}$$

*where $\beta$ is a positive valued function: $\beta : \Theta \to (0, \infty)$.*

We will refer to $\beta(\theta)$ in Definition 6 as the *NS coefficient*. This alternative definition generalizes and unifies several smoothness concepts from the recent literature. First, NS clearly reduces to Eq. (4.1) with $\beta(\theta) = \beta$. However, NS also generalizes the notion of $(L_0, L_1)$ smoothness from Zhang et al. (2019) by using $\beta(\theta) = L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$. By using $\beta(\theta) = c \cdot \|\nabla f(\theta)\|_2^{\frac{p-2}{p-1}}$, NS also reduces to the notion of strong smoothness of order $p$ proposed in Wilson et al. (2019). Finally, with $\beta(\theta) = c/\|\theta\|_p^2$, NS reduces to a special form of non-uniform smoothness considered in Lemma 19. We will show later that

NS also covers other previously unstudied smoothness variants. Below we will demonstrate the key benefits of Definition 6 in terms of its *generality*, *better convergence results*, and *practical implications* in conjunction with the NŁ inequality.

### 4.3.2 Non-uniform Łojasiewicz (NŁ) Inequality

The second main concept we leverage is a new generalized Łojasiewicz inequality introduced in Lemma 3 in Chapter 2:

**Definition 7** (Non-uniform Łojasiewicz (NŁ))**.** *The differentiable function* $f : \Theta \to \mathbb{R}$ *satisfies the* $(C(\theta), \xi)$ *non-uniform Łojasiewicz inequality if for all* $\theta \in \Theta$*,*

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi}, \tag{4.5}$$

*where* $\xi \in (-\infty, 1]$*, and* $C(\theta) : \Theta \to \mathbb{R} > 0$ *holds for all* $\theta \in \Theta$*. In this definition, either* $\theta^* = \arg\min_{\theta \in \Theta} f(\theta)$*, or* $f(\theta^*)$ *is replaced with* $\inf_\theta f(\theta)$ *if the global optimum is not achieved within the domain* $\Theta$*.*

Definition 7 extends the classical "uniform" Łojasiewicz inequalities in optimization literature, such as the Polyak-Łojasiewicz (PŁ) inequality with $C(\theta) = C > 0$ and $\xi = 1/2$ (Łojasiewicz, 1963; Polyak, 1963); and the Kurdyka-Łojasiewicz (KŁ) inequality[1] by setting $C(\theta) = C > 0$ (Kurdyka, 1998). We refer to $\xi$ as the *NŁ degree* (Definition 1) and $C(\theta)$ as the *NŁ coefficient* (Definition 3). Generally speaking, a larger NŁ degree $\xi$ and NŁ coefficient $C(\theta)$ indicate faster convergence for gradient based algorithms. Chapter A provides an overview of remarkable non-convex functions that satisfy the NŁ inequality for various $\xi$ and $C(\theta)$. As stated, our main contribution is to show how, when *combined* with NS, NŁ becomes a powerful tool for both algorithm design and analysis, which is a novel direction of investigation.

[1] The KŁ inequality is violated at bad local optima, since vanishing gradient norm cannot dominate non-zero sub-optimality gap. Therefore Definition 7 actually recovers global KŁ inequality.

## 4.4 Geometry-aware Gradient Descent

A key benefit of the non-uniform definitions is that we can introduce step-size rules that make gradient descent adapt to the local "geometry" of the optimization objective. First consider the classical gradient decent update.

**Definition 8** (Gradient Descent (GD))**.**

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla f(\theta_t). \tag{4.6}$$

The key challenge with deploying GD is choosing the step size $\eta$; if $\eta$ is too large, instability ensues, if too small, progress becomes slow. Recall from the Eq. (4.3) that $\eta = 1/\beta$ is a canonical choice for assuring convergence in *uniformly* $\beta$ smooth objectives. This suggests that in the presence of *non-uniform* smoothness $\beta(\theta)$ given in NS, the stepsize should be adapted to $1/\beta(\theta)$. This leads to a new variant of normalized gradient descent.

**Definition 9** (Geometry-aware Normalized GD (GNGD))**.**

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \frac{\nabla f(\theta_t)}{\beta(\theta_t)}. \tag{4.7}$$

Key to making this approach practical will be efficient ways to measure (or bound) $\beta(\theta)$. Below we will show how in the context of NS and NŁ properties, GNGD can be made both practical and extremely efficient at solving various global optimization problems in machine learning. These results also broaden our fundamental knowledge of the set of objectives that admit efficient global optimization.

## 4.5 Non-uniform Analysis

### 4.5.1 Main Theorem

Our first main contribution in this chapter is an analysis for GD and GNGD based in the presence of non-uniform properties. For minimization problems, we assume $\inf_\theta f(\theta) > -\infty$ ($\sup_\theta f(\theta) < \infty$ for maximization problems).

**Theorem 16.** *Suppose $f : \Theta \to \mathbb{R}$ satisfies NS with $\beta(\theta)$ and the NŁ inequality with $(C(\theta), \xi)$. Suppose $C \coloneqq \inf_{t\geq 1} C(\theta_t) > 0$ for GD and GNGD. Let $\delta(\theta) \coloneqq f(\theta) - f(\theta^*)$ be the sub-optimality gap. The following hold:*

**(1a)** *if $\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}$ with $\xi \in (-\infty, 1/2)$, then the conclusions of (1b) hold;*

**(1b)** *if $\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$ with $\xi \in (-\infty, 1/2)$, then GD with $\eta \in O(1)$ achieves $\delta(\theta_t) \in \Theta(1/t^{\frac{1}{1-2\xi}})$, and GNGD achieves $\delta(\theta_t) \in O(e^{-c' \cdot t})$.*

**(2a)** *if $\beta(\theta) \leq L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$, then the conclusions of (2b) hold;*

**(2b)** *if $\beta(\theta) \leq L_0 \cdot \frac{\|\nabla f(\theta)\|^2}{\delta(\theta)^{2-2\xi}} + L_1 \cdot \|\nabla f(\theta)\|_2$, then GD and GNGD both achieve $\delta(\theta_t) \in O(1/t^{\frac{1}{1-2\xi}})$ when $\xi \in (-\infty, 1/2)$, and $O(e^{-c' \cdot t})$ when $\xi = 1/2$. GNGD has strictly better constant than GD ($1 > C > C^2$).*

**(3a)** *if $\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$ with $\xi \in (1/2, 1)$, then the conclusions of (3b) hold;*

**(3b)** *if $\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}$ with $\xi \in (1/2, 1)$, then GD with $\eta \in \Theta(1)$ does not converge, while GNGD achieves $\delta(\theta_t) \in O(e^{-c' \cdot t})$.*

**Remark 9.** *The cases (1)-(3) cover all three possibilities of $\beta(\theta^*)$. Since $\theta^*$ is the global minimum, $\nabla^2 f(\theta^*)$ is positive semi-definite (negative if $\theta^*$ is maximum) if it exists.*

*(1) If $\nabla^2 f(\theta^*) = \mathbf{0}$, then $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$, which means the landscape around $\theta^*$ is flat.*

*(2) If $\nabla^2 f(\theta^*)$ has at least one strictly positive (negative) eigenvalue, then $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.*

*The cases (1)-(2) also cover the situations where the Hessian $\nabla^2 f(\theta^*)$ does not exist but one can find a finite $\beta(\theta^*) > 0$ to upper bound the l.h.s. of Definition 6.*

*(3) The case (3) is for blow-up type non-existence of $\nabla^2 f(\theta^*)$, where $\beta(\theta^*)$ is unbounded.*

**Remark 10.** *In Theorem 16, the NŁ coefficient $C \coloneqq \inf_{t\geq 1} C(\theta_t)$ is related to the early optimization and plateau escaping behavior studied in Chapter 3.*

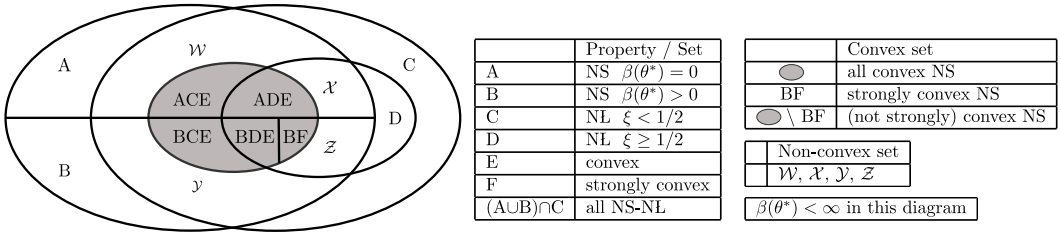



Figure 4.2: Different function classes for $\beta(\theta^*) < \infty$. We use a label notation where, e.g., C denotes the set of all functions that satisfy property C, and ACE $\coloneqq$ A $\cap$ C $\cap$ E. The two largest ellipsoids correspond to A $\cup$ B and C. We study the following four non-convex function classes in (A $\cup$ B) $\cap$ $C$, i.e., $\mathcal{W} \coloneqq$ AC $\setminus$ (AD $\cup$ ACE), $\mathcal{X} \coloneqq$ AD $\setminus$ ADE, $\mathcal{Y} \coloneqq$ BC $\setminus$ (BD $\cup$ BCE), and $\mathcal{Z} \coloneqq$ BD $\setminus$ (BDE $\cup$ BF).

*It remains open to study whether GNGD can be combined with the alternative escort parameterizations in Definition 2 to further improve C.*

Note that (1b) recovers the strong smoothness of order $p$ with $p = 1/\xi$ in Wilson et al. (2019), and (2a) recovers the $(L_0, L_1)$ smoothness of Zhang et al. (2019). The results here consider more general NŁ functions and establish faster rates of convergence. The other cases have not been studied in literature to our knowledge. In Sections 4.6 and 4.7 below we study practical machine learning examples that are covered by cases (1) and (2) in Theorem 16. Other cases of different $\beta(\theta)$ and $\xi$ are discussed below for completeness.

### 4.5.2 Function Classes

Before applying these results to problems in machine learning, we first provide a refined characterization of function classes organized by their NS and NŁ properties. This also clarifies the relation between the non-uniform properties and standard notions of convexity and smoothness; see Fig. 4.2.

**Proposition 7.** *The following hold for an objective $f$:*

**(1)** *$D \subseteq C$. If $f$ satisfies NŁ with degree $\xi$, it satisfies NŁ with degree $\xi' < \xi$;*

**(2)** *$F \subseteq D$. A strongly convex $f$ satisfies NŁ with $\xi = 1/2$;*

**(3)** *$F \cap A = \emptyset$. A strongly convex $f$ cannot satisfy NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$;*

**(4)** *$E \subseteq C$. A (not strongly) convex $f$ satisfies NŁ with $\xi = 0$.*

The next proposition provides concrete examples for each convex function class in $(A \cup B) \cap C$ in Fig. 4.2.

**Proposition 8.** *The following results hold:*

**(1)** *$ACE \neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.*

**(2)** *$ADE \neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.*

**(3)** *$BCE \neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.*

**(4)** *$BDE \neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.*

**(5)** *$BF \neq \emptyset$. There exists at least one strongly convex function which satisfies NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.*

A more interesting result considers examples in the classes of non-convex functions $(A \cup B) \cap C$ in Fig. 4.2. The non-uniform analysis above largely still applies to these problems, even when standard convex analysis cannot apply.

**Proposition 9.** *The following results hold:*

**(1)** *$\mathcal{W} \coloneqq AC \setminus (AD \cup ACE) \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.*

**(2)** *$\mathcal{X} \coloneqq AD \setminus ADE \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.*

**(3)** *$\mathcal{Y} \coloneqq BC \setminus (BD \cup BCE) \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.*

**(4)** *$\mathcal{Z} \coloneqq BD \setminus (BDE \cup BF) \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.*

We next apply the techniques to a class of convex functions, achieving results that cannot be explained by classical convex-smooth analysis.

**Proposition 10.** *The convex function $f : x \mapsto |x|^p$ with $p > 1$ satisfies the NŁ inequality with $\xi = 1/p$ and the NS property with $\beta(x) \leq c_1 \cdot \delta(x)^{1-2\xi}$.*

Consider any $p > 2$, such as $p = 4$, where it follows that $f$ satisfies NŁ with degree $\xi = 1/4 < 1/2$. According to (1a) in Theorem 16, GD will achieve $\delta(x_t) \in \Theta(1/t^2)$, while GNGD attains $\delta(x_t) \in O(e^{-c \cdot t})$ (where $c > 0$). Note that standard convex analysis can only give a $O(1/t)$ rate on (not strongly) convex smooth functions. The $\Theta(1/t^2)$ rate for GD here follows from using NŁ degree $\xi = 1/4$, which improves on $\xi = 0$ from mere convexity ((4) in Proposition 7). The $O(e^{-c \cdot t})$ rate has also been observed for this example by exploiting strong smoothness ((1b), as noted) (Wilson et al., 2019). Fig. 4.2 provides a more general understanding of when this happens.

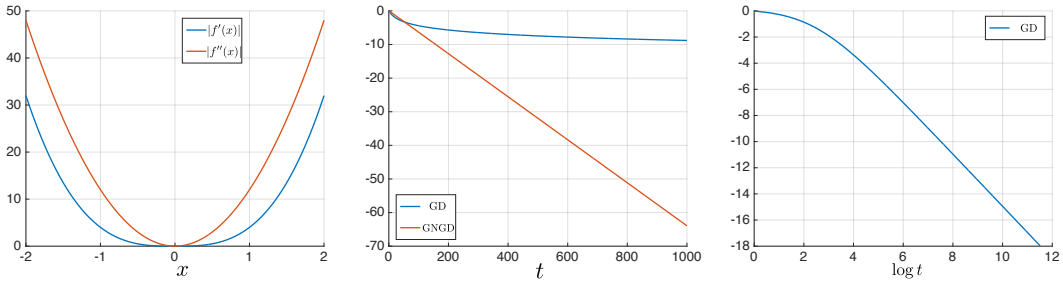
### 4.5.3 Existing Lower Bounds

**$\Omega(1/t^2)$ lower bound for convexity-smoothness.** Note that GNGD satisfies

$$x_{t+1} = x_1 - \sum_{i=1}^{t} \frac{\eta}{\beta(x_i)} \cdot \nabla f(x_i) \in \text{Span}\left\{x_1, \nabla f(x_1), \ldots, \nabla f(x_t)\right\}, \tag{4.8}$$

which is a first-order oracle (Nesterov, 2003). Thus there exists a worst-case objective in the convex-smooth class that forces $\delta(x_t) \in \Omega(1/t^2)$ for $t \in O(n)$, where $n$ is the parameter dimension (Bubeck et al., 2015; Nemirovski and Yudin, 1983; Nesterov, 2003). This is not a contradiction, since the lower bound is established by constructing a convex smooth function with a *constant* $\beta > 0$ (Bubeck et al., 2015), and $\beta(x) \to \beta > 0$ as $x, x' \to x^*$ in Definition 6. Hence, the $\Omega(1/t^2)$ result covers *some* functions in BCE in Fig. 4.2. Meanwhile $f : x \mapsto |x|^p$ with $p > 2$ satisfies $\beta(x) \to 0$ as $x, x' \to 0$ in Definition 6 (ACE in Fig. 4.2), which implies that the standard convex-smooth class consists of two subclasses. One subclass (BCE) admits first-order sub-linear lower bounds, while the other (ACE) allows linear convergence using first-order methods. This illustrates the *necessity* of non-uniformity in subdividing the NS class as A $\cup$ B in Fig. 4.2. This partition also inspires geometry-aware GD.

As shown in Proposition 10, with $p \in (1, 2)$, $f : x \mapsto |x|^p$ satisfies NŁ inequality with $\xi = 1/p \in (0, 1/2)$. As shown in Fig. 4.3(a), the spectral radius of Hessian approaches 0 as $x \to 0$, which is the case (1) in Theorem 16.




(a) gradient and Hessian (b) $\log \delta(x_t)$ (c) $\log \delta(x_t)$, slope $\approx -1.9996$

Figure 4.3: GD and GNGD on $f : x \mapsto |x|^p$, $p = 4$.

Subfigure (c) shows that the standard GD with constant learning rate $\eta = 0.01$ achieves sublinear rate about $O(1/t^2)$, while subfigure (b) shows that GNGD with $\eta = 0.01$ enjoys linear rate $O(e^{-c \cdot t})$, verifying Theorem 16.

**$\Omega(1/\sqrt{t})$ lower bound for $(L_0, L_1)$-smoothness.** For $\beta(\theta) = L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$ ((2a) in Theorem 16) with $L_0, L_1 \geq 1$, standard normalized GD is subject to a $\Omega(1/\sqrt{t})$ lower bound (Zhang et al., 2019). However, in Section 2.3, we will show that normalized policy gradient (PG) method achieves a linear rate of $O(e^{-c \cdot t})$. Again, this is not a contradiction for similar reasons. With $L_0 \geq 1$, $\beta(\theta) \to L_0 > 0$ as $\theta, \theta' \to \theta^*$, the $\Omega(1/\sqrt{t})$ lower bound will hold for *some* functions in BCE $\cup\, \mathcal{Y}$ in Fig. 4.2. While in Section 2.3 the objective satisfies $L_0 = 0$ and $L_1 > 0$, hence $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$ (ACE $\cup\, \mathcal{W}$ in Fig. 4.2). This shows a similar separation of rates for first-order methods will also occur based on NS conditions. Furthermore, in Section 4.7, we will show that both GD and GNGD achieve a $O(e^{-c \cdot t})$ rate for GLM, but here the objective is in $\mathcal{Z}$ in Fig. 4.2 so the lower bounds do not apply.

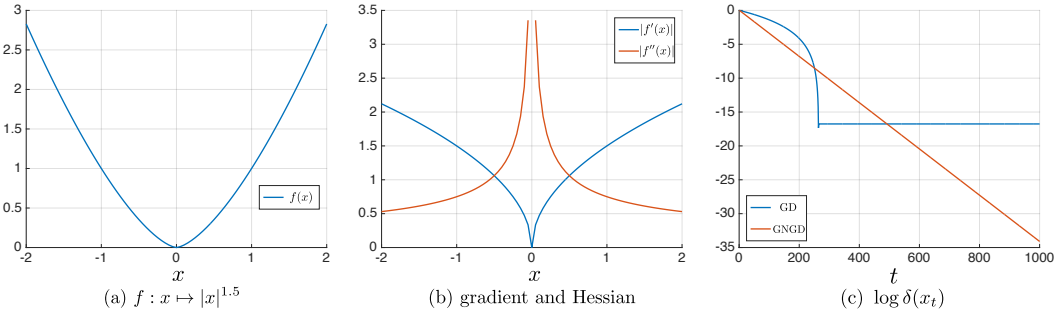
### 4.5.4 Unbounded Hessian

Consider any $p \in (1, 2)$, such as $p = 3/2$ where $f$ satisfies $\xi = 2/3$. According to Theorem 16(3a), GD diverges since the Hessian is unbounded near 0. This makes it *necessary* to introduce geometry-aware normalization to ensure convergence, which is verified in Fig. 4.4. This has practical implications for RL, for example ensuring exploration using state distribution entropy, which has

unbounded Hessian near probability simplex boundary (Hazan et al., 2019) and alternative escort probability transforms when $p \in (1, 2)$ (Definition 2).

As shown in Proposition 10, with $p \in (1, 2)$, $f : x \mapsto |x|^p$ satisfies NŁ inequality with $\xi = 1/p \in (1/2, 1)$, which is the case (3) in Theorem 16. The function $f$ is differentiable, and the Hessian $|f''(x)| = p \cdot (p-1) \cdot |x|^{p-2} \to \infty$, as $x \to 0$, which indicates GD with $\eta \in \Theta(1)$ does not converge.

Fig. 4.4(a) shows the image of $f : x \mapsto |x|^{1.5}$. As shown in subfigure (b), the gradient of $f$ exists at $x = 0$, and the Hessian $|f''(x)| \to \infty$ as $x \to 0$. The results of GD with $\eta = 0.005$ and GNGD are presented in subfigure (c). The sub-optimality of GD update decreased for some time, and then it increased later. This is due to the Hessian is unbounded near $x = 0$, and thus constant learning rates cannot guarantee monotonic progresses for GD. On the other hand, GNGD with $\eta = 0.01$ enjoys $O(e^{-c \cdot t})$ convergence rate, verifying the results in the case (3) in Theorem 16.




(a) $f : x \mapsto |x|^{1.5}$ (b) gradient and Hessian (c) $\log \delta(x_t)$

Figure 4.4: GD and GNGD on $f : x \mapsto |x|^p$, $p = 1.5$.

## 4.6 Geometry-aware Normalized Policy Gradient

Our second main contribution is to show that the expected return objective considered in direct policy optimization in RL falls under the function class $\mathcal{W}$ in Fig. 4.2, in particular satisfying case (1) of Theorem 16 with NŁ degree $\xi = 0$. The key point is that value functions in Markov decision processes (MDPs) satisfy NS properties with coefficient being the PG norm (Lemmas 21 and 24).

This novel finding not only provides a much more precise characterization than existing standard smoothness results in Lemma 7, but also enables PG with normalization to use the NŁ inequalities (Lemmas 3 and 8) differently than for standard PG ($\|\nabla f(\theta)\|_2$ vs. $\|\nabla f(\theta)\|_2^2$), which leads to faster convergence as well as plateau escaping.

### 4.6.1 Convergence Rate: One-state MDPs

We first illustrate some key insights for one-state MDPs with $K$ actions and $\gamma = 0$. The value function Eq. (2.1) reduces to expected reward Eq. (2.12),

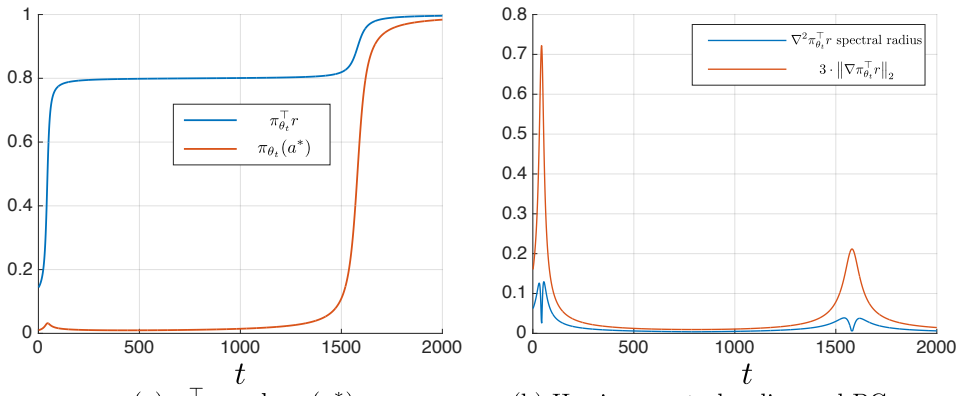
$$\max_{\theta:\mathcal{A}\to\mathbb{R}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} [r(a)]. \tag{4.9}$$

where $r \in [0,1]^K$, $\theta \in \mathbb{R}^K$, and $\pi_\theta = \text{softmax}(\theta)$. In Chapter 2, we have shown that even though $\max_\theta \pi_\theta^\top r$ is a non-concave maximization, global convergence can be achieved with a $O(1/t)$ rate using uniform smoothness and the NŁ inequality of Lemma 3. Let $a^*$ be the optimal action. Denote $\pi^* = \arg\max_{\pi\in\Delta} \pi^\top r$. Then,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \tag{4.10}$$

Note that Lemma 3 is not improvable in terms of the coefficients $C(\theta) = \pi_\theta(a^*)$ and $\xi = 0$ as shown in Remark 1 and Lemma 17 respectively. However, this result is based on only using a *uniform* smoothness coefficient $\beta = 5/2$ in Lemma 2, which even empirical evidence suggests can be significantly refined. To illustrate, we run standard policy gradient (PG) on a 3-action one-state MDP. As shown in Fig. 4.5(a), PG first goes through a long suboptimal plateau, and then eventually escapes to approach $\pi^*$. Fig. 4.5(b) presents the spectral radius of the Hessian and the PG norm $3 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$ as functions of time $t$. It is evident that the smoothness behaves non-uniformly: it is close to zero at the suboptimal plateau and near $\pi^*$, highly aligned with the PG norm. Compared to any universal constant $\beta$, the PG norm characterizes the non-uniform landscape information far more precisely. We formalize this observation by proving the following key result:

**Lemma 21** (NS). *Denote* $\theta_\zeta := \theta + \zeta \cdot (\theta' - \theta)$ *with some* $\zeta \in [0,1]$. *For any* $r \in [0,1]^K$, $\theta \mapsto \pi_\theta^\top r$ *satisfies* $\beta(\theta_\zeta)$ *non-uniform smoothness with* $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$.




(a) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$ (b) Hessian spectral radius and PG norm

Figure 4.5: PG results on $r = (1.0, 0.8, 0.1)^\top$.

Comparing Lemma 21 with (1b) in Theorem 16, we have $\xi = 0$, and GNGD requires normalizing $\beta(\theta_\zeta)$, which is the PG norm of $\theta_\zeta$ rather than $\theta$. However, $\zeta$ is unknown. Fortunately, the next lemma shows that, if we still normalize the PG norm of $\theta$, the $\beta(\theta_\zeta)$ in Lemma 21 can be upper bounded by $\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$, given the learning rate is small enough:

**Lemma 22.** *Let* $\theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta} \Big/ \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$. *Denote* $\theta_\zeta := \theta + \zeta \cdot (\theta' - \theta)$ *with some* $\zeta \in [0,1]$. *We have, for all* $\eta \in (0, 1/3)$,

$$\left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2 \leq \frac{1}{1 - 3\eta} \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 . \tag{4.11}$$

Next, the NŁ coefficient $\pi_\theta(a^*)$ is bounded away from 0, which provides constants in the convergence rate results.

**Lemma 23** (Non-vanishing NŁ coefficient). *Using normalized policy gradient method, we have* $\inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$.

To this point, we demonstrate that the non-concave function $\pi_\theta^\top r$ satisfies (1b) in Theorem 16 with $\xi = 0$ *in each iteration of normalized PG*[2]: Lemmas 21

[2]This essentially means we prove that a uniform Łojasiewicz inequality holds for the

and 22 show that the NS coefficient $\beta(\theta_{\zeta_t}) \leq c_1 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$, while Lemmas 3 and 23 guarantee $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \geq c_2 \cdot (\pi^* - \pi_{\theta_t})^\top r$. Therefore, combining Lemmas 3 and 21 to 23, we prove the global linear convergence rate $O(e^{-c \cdot t})$ of normalized PG:

**Theorem 17.** *Using normalized PG* $\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$, *with* $\eta = 1/6$, *for all* $t \geq 1$, *we have,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq (\pi^* - \pi_{\theta_1})^\top r \cdot e^{-\frac{c \cdot (t-1)}{12}}, \tag{4.12}$$

*where* $c = \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ *is from Lemma 23, and* $c$ *is a constant that depends on* $r$ *and* $\theta_1$, *but not on the time* $t$.

**Remark 11.** *If* $\pi_{\theta_1}$ *is uniform, i.e.,* $\pi_{\theta_1}(a) = 1/K$, $\forall\ a \in [K]$, *then we have* $c \geq 1/K$ *in Theorem 17. This can be proved by showing that* $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$, *similar to Proposition 2*

### 4.6.2 Geometry-aware Normalized PG (GNPG)

Next, we generalize from one-state to finite MDPs, using the GNPG[3] on value function, as shown in Algorithm 2.

**Algorithm 2** Geometry-aware Normalized Policy Gradient

**Input:** Learning rate $\eta > 0$.
Initialize parameter $\theta_1(s, a)$ for all $(s, a)$.
**while** $t \geq 1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2$.
**end while**

### 4.6.3 Convergence Rate: General MDPs

For general finite MDPs, we assume "sufficient exploration" for the initial state distribution $\mu$, which is from Assumption 2. The initial state distribution

entire sequence $\{\theta_t\}_{t \geq 1}$, but this does not imply that the NŁ condition is unnecessary. As shown in Remark 1, Łojasiewicz-type inequalities with constant $C > 0$ cannot hold. It can only become uniform after specifying an initialization $\theta_1$ and an algorithm (in this case, PG). Otherwise, uniform Łojasiewicz cannot hold since initialization can make the NŁ coefficient $\pi_\theta(a^*)$ arbitrarily close to 0.

[3]We use GNPG as the name of Algorithm 2, since NPG is usually used to refer to the natural PG algorithm in RL literature (Kakade, 2002).

satisfies

$$\min_s \mu(s) > 0. \tag{4.13}$$

Given Assumption 2, in Theorem 4, we prove a $O(1/t)$ rate using uniform smoothness and the NŁ inequality of Lemma 8. We have, $\forall\ \theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot [V^*(\rho) - V^{\pi_\theta}(\rho)], \tag{4.14}$$

where $S \coloneqq |\mathcal{S}|$ is the total number of states, and $a^*(s)$ is the action that $\pi^*$ selects in state $s$. Here, the NŁ degree $\xi = 0$ is not improvable as shown in Lemma 18. In one-state MDPs with $S = 1$, Lemma 8 recovers Lemma 3 with the same NŁ coefficient $C(\theta) = \pi_\theta(a^*)$, indicating that $C(\theta)$ in Lemma 8 might also be unimprovable. On the other hand, the uniform smoothness considered in Lemma 7, i.e., $\beta = 8/(1-\gamma)^3$ is too conservative, particularly when $\gamma$ is close to 1. Our next key result shows that the policy value also satisfies a stronger NS property, with the NS coefficient being the PG norm, generalizing Lemma 21:

**Lemma 24** (NS)**.** *Let Assumption 2 hold and denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. $\theta \mapsto V^{\pi_\theta}(\mu)$ satisfies $\beta(\theta_\zeta)$ non-uniform smoothness with*

$$\beta(\theta_\zeta) = \left[3 + \frac{2 \cdot (C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma}\right] \cdot \sqrt{S} \cdot \left\|\frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta}\right\|_2, \tag{4.15}$$

*where $C_\infty \coloneqq \max_\pi \left\|\frac{d_\mu^\pi}{\mu}\right\|_\infty \leq \frac{1}{\min_s \mu(s)} < \infty$.*

In one-state MDPs with $\gamma = 0$ and $S = 1$, we have $C_\infty = 1 - \gamma$. Thus Lemma 24 reduces to Lemma 21 with the same NS coefficient $\beta(\theta_\zeta) = 3 \cdot \left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2$. Similar to Lemma 22, if we use Algorithm 2 with small enough learning rate, then $\beta(\theta_\zeta)$ in Lemma 24 is upper bounded by the PG norm of $\theta$:

**Lemma 25.** *Let $\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}}$ and $\theta' = \theta + \eta \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \Big/ \left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2$. Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. We have,*

$$\left\|\frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta}\right\|_2 \leq 2 \cdot \left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2. \tag{4.16}$$

Next, the NŁ coefficient $\min_s \pi_\theta(a^*(s)|s)$ in Lemma 8 is lower bounded away from 0:

**Lemma 26** (Non-vanishing NŁ coefficient)**.** *Let Assumption 2 hold and let* $\{\theta_t\}_{t\geq 1}$ *be generated by Algorithm 2. We have,* $c := \inf_{s\in\mathcal{S},t\geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$.

Now we have the non-concave function $V^{\pi_\theta}(\rho)$ satisfies (1b) in Theorem 16 with $\xi = 0$ *in each iteration of Algorithm 2*. Therefore, combining Lemmas 8 and 24 to 26, we prove the global linear convergence rate $O(e^{-c\cdot t})$ of Algorithm 2:

**Theorem 18.** *Let Assumption 2 hold and let* $\{\theta_t\}_{t\geq 1}$ *be generated using Algorithm 2 with learning rate*

$$\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))}\cdot\frac{1}{\sqrt{S}}, \tag{4.17}$$

*where* $C_\infty := \max_\pi \left\|\frac{d_\mu^\pi}{\mu}\right\|_\infty$. *Denote* $C'_\infty := \max_\pi \left\|\frac{d_\rho^\pi}{\mu}\right\|_\infty$. *Let* $c$ *be the positive constant from Lemma 26. We have, for all* $t \geq 1$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{(V^*(\mu) - V^{\pi_{\theta_1}}(\mu))\cdot C'_\infty}{1-\gamma}\cdot e^{-C\cdot(t-1)}, \tag{4.18}$$

*where*

$$C = \frac{(1-\gamma)^2\cdot\gamma\cdot c}{12\cdot(1-\gamma)\cdot\gamma + 8\cdot(C_\infty - (1-\gamma))}\cdot\frac{1}{S}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-1}. \tag{4.19}$$

Not only the $O(e^{-c\cdot t})$ rate in Theorem 18 is faster than $O(1/t)$ for standard PG without normalization, but also the constant is better than Theorem 4. The strictly better dependence $c$ ($\gg c^2$ in PG) is related to faster escaping plateaus as shown in Chapter 3.

**Remark 12.** *The conclusion of GNPG has better constants than PG* ($c \gg c^2$) *arises from upper bounds (Theorems 4 and 18), which is also supported by empirical evidence. According to Theorem 11, there exists a lower bound that shows* $c$ *cannot be removed for PG under one-state MDP settings. For finite MDPs, very recently, Li et al. (2021) show that for softmax PG (without normalization),* $c$ *can be very small in terms of the number of states. It remains open to consider whether* $c$ *is reasonably large for GNPG.*

**Remark 13.** *To our knowledge, existing PG variants can achieve linear convergence $O(e^{-c\cdot t})$ only if using at least one of the following techniques: (a)* ***regularization****; In Theorem 6, we prove that entropy regularized PG enjoys $O(e^{-c\cdot t})$ convergence toward the regularized optimal policy. (b)* ***natural gradient****; Cen et al. (2020) prove that entropy regularized natural PG achieves linear convergence toward regualrized optimal policy. (c)* ***exact line-search****; Bhandari and Russo (2020) prove that without parameterization, PG variants with exact line-search achieve linear rates by approximating policy iteration.*

*Among the above techniques, regularization changes the problem to regularized MDPs, while natural PG and line-search require solving expensive optimization problems to do updates, since each update is an* $\arg\max$.

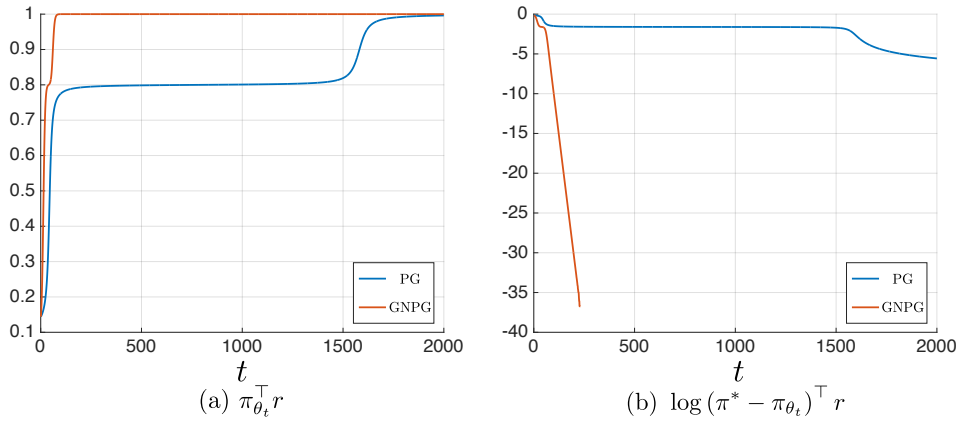
On the contrary, Algorithm 2 enjoys global $O(e^{-c\cdot t})$ rate *(i)* without using regularization, since Algorithm 2 directly works on the original MDPs; *(ii)* without solving optimization problems in each iteration, and the normalized PG update is cheap. The strong results rely on the NS and NŁ properties, and also the geometry-aware normalization that takes advantage of the non-uniform properties.

**Remark 14.** *According to Theorem 10, standard softmax PG of Algorithm 1 with bounded learning rate follows $\Omega(1/t)$ lower bound, which is consistent with the case (1) in Theorem 16. Algorithm 2 achieves faster linear convergence rates, indicating that the adaptive update stepsize $\eta/\left\|\nabla V^{\pi_{\theta_t}}(\rho)\right\|_2$ is asymptotically unbounded, since $\left\|\nabla V^{\pi_{\theta_t}}(\rho)\right\|_2 \to 0$ as $t \to \infty$.*

### 4.6.4 Empirical Verification

We compare PG and GNPG on the one-state MDP problem as shown in Fig. 4.6. Fig. 4.6(a) shows that GNPG escapes from the sub-optimal plateau significantly faster than PG, while Fig. 4.6(b) shows that GNPG follows linear convergence $O(e^{-c\cdot t})$ of sub-optmality, verifying the theoretical results.

Fig. 4.7 shows the results for PG and GNPG beyond one-state MDPs. The environment is a synthetic tree with height $h$ and branching factor $b$. The total




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

Figure 4.6: PG and GNPG on $r = (1.0, 0.8, 0.1)^\top$.

number of states is

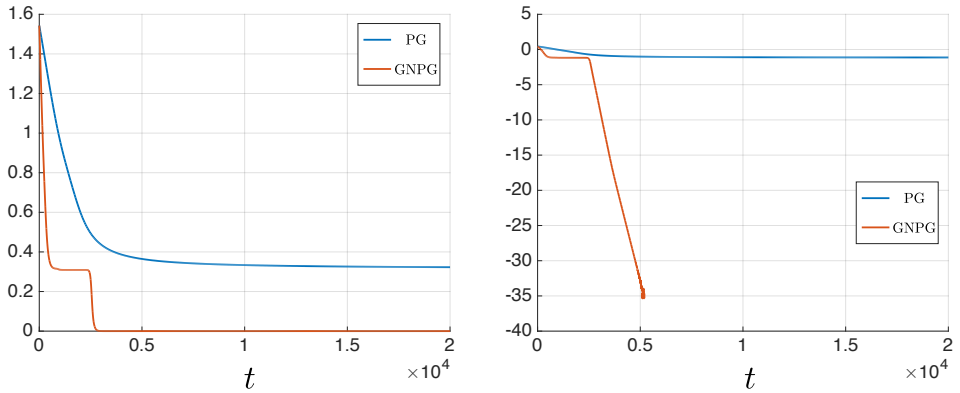
$$S = \sum_{i=0}^{h-1} b^i. \tag{4.20}$$

The discount factor $\gamma = 0.99$, and we set $\mu = \rho$ (e.g., in Algorithm 2 and Theorem 18), where $\rho(s_0) = 1$ for the root state $s_0$. For PG, in each iteration, we calculate the policy gradient (Lemma 1) to do one update. For GNPG, Algorithm 2 is used.

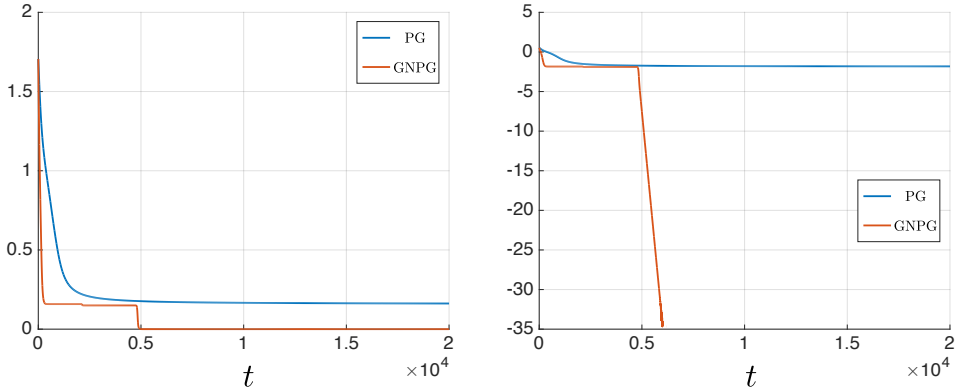
Subfigures (a) and (b) show the results for $h = b = 4$, and $S = 85$. The learning rate is $\eta = 0.02$ for PG and GNPG. Subfigures (c) and (d) show the results for $h = 5$ and $b = 4$, and $S = 341$. The learning rate is $\eta = 0.05$ for PG and GNPG.

## 4.7 Generalized Linear Models

Next, we investigate the generalized linear model (GLM) with quasi-maximum likelihood estimate (quasi-MLE), which applied widely in supervised learning. We show that the mean squared error (MSE) of GLM is in the non-convex function class $\mathcal{Z}$ in Fig. 4.2, and it satisfies the case (2) in Theorem 16 with $\xi = 1/2$. As a result, both GD and GNGD achieve global linear convergence rates $O(e^{-c \cdot t})$, significantly improving the best existing results of $O(1/\sqrt{t})$ (Hazan et al., 2015). We also provide new understandings of using normalization in GLM based on our non-uniform analysis.




(a) $V^*(\rho) - V^{\pi_{\theta_t}}(\rho)$ (b) $\log\left(V^*(\rho) - V^{\pi_{\theta_t}}(\rho)\right)$




(c) $V^*(\rho) - V^{\pi_{\theta_t}}(\rho)$ (d) $\log\left(V^*(\rho) - V^{\pi_{\theta_t}}(\rho)\right)$

Figure 4.7: Results for PG and GNPG on tree MDPs. In (a) and (b), $S = 85$. In (c) and (d), $S = 341$.

### 4.7.1 Basic Settings and Notations

Given a training data set $\mathcal{D} = \{(x_i, y_i)\}_{i\in[N]}$, which consists of $N$ data points, there is a feature map $x_i \mapsto \phi(x_i) \in \mathbb{R}^d$ for each pair $(x_i, y_i) \in \mathcal{D}$. We denote $\phi_i \coloneqq \phi(x_i)$ for conciseness. For each data point $x_i$, we have $y_i \in [0, 1]$ as the ground truth likelihood. Following Hazan et al. (2015), our model is parameterized by a weight vector $\theta \in \mathbb{R}^d$ as ,

$$\pi_i = \sigma(\phi_i^\top \theta) = \frac{1}{1 + \exp\{-\phi_i^\top \theta\}}, \tag{4.21}$$

where $\sigma : \mathbb{R} \to (0, 1)$ is the sigmoid activation. The problem is to minimize the mean squared error (MSE),

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta \in \mathbb{R}^d} \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - y_i)^2. \tag{4.22}$$

We assume $y_i = \pi_i^* := \sigma(\phi_i^\top \theta^*)$, where $\theta^* \in \mathbb{R}^d$, and $\|\theta^*\|_2 < \infty$, which means the target $y_i$ is realizable and non-deterministic. According to Hazan et al. (2015), the MSE in Eq. (4.22) is not quasi-convex (thus not convex). Fortunately, Hazan et al. (2015) manage to show that Eq. (4.22) satisfies a weaker Strictly-Locally-Quasi-Convex (SLQC) property, based on which they prove the following result:

**Theorem 19** (Hazan et al. (2015))**.** *With diminishing learning rate* $\eta_t \in \Theta(1/\sqrt{t})$*, the normalized gradient descent (NGD) update* $\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \Big/ \left\| \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \right\|_2$ *satisfies,*

$$\delta(\theta_t) := \mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \in O(1/\sqrt{t}), \tag{4.23}$$

*where* $\theta^* := \arg\min_\theta \mathcal{L}(\theta)$ *is the global optimal solution.*

### 4.7.2 Fast Convergence using Non-uniform Analysis

Based on the $O(1/\sqrt{t})$ rate for NGD in Theorem 19, Hazan et al. (2015) propose to normalize gradient norm in MSE minimization. However, there is no lower bound for other methods including GD on GLM, and thus it is not clear if there exists a faster rate for GLM optimization.

Surprisingly, we prove that both GD and GNGD actually achieve much faster rates of $O(e^{-c \cdot t})$ using the non-uniform analysis. Our first key finding is to show that the MSE in GLM satisfies a new NŁ inequality with $\xi = 1/2$:

**Lemma 27** (NŁ)**.** *Denote*

$$u(\theta) := \min_{i \in [N]} \{\pi_i \cdot (1 - \pi_i)\}, \text{ and} \tag{4.24}$$

$$v := \min_{i \in [N]} \{\pi_i^* \cdot (1 - \pi_i^*)\}. \tag{4.25}$$

*We have,*

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \geq C(\theta, \phi) \cdot \left[ \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}, \tag{4.26}$$

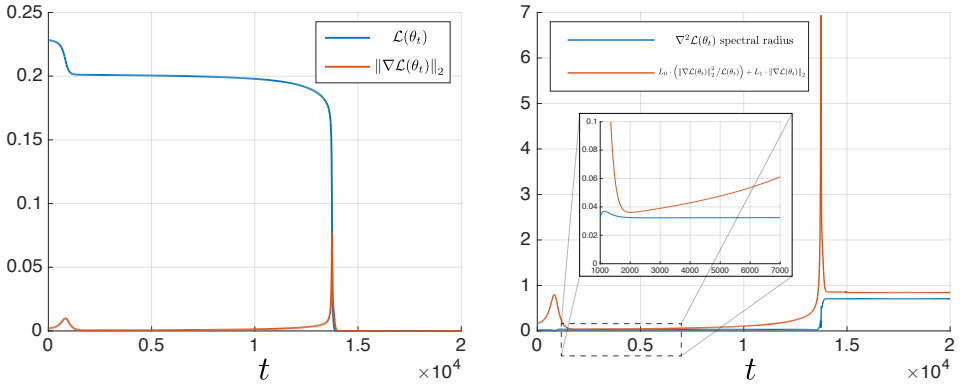
*holds for all* $\theta \in \mathbb{R}^d$*, where*

$$C(\theta, \phi) = 8 \cdot u(\theta) \cdot \min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi}, \tag{4.27}$$

*and* $\lambda_\phi$ *is the smallest positive eigenvalue of* $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$*.*

**Remark 15.** *It is not clear if results similar to Lemma 27 hold without assuming: (i) realizable optimal prediction* $y_i = \pi_i^* := \sigma(\phi_i^\top \theta^*)$*; (ii) non-deterministic optimal prediction* $\|\theta^*\|_2 < \infty$*. We leave it as an open question to study non-uniformity of GLM without the above assumptions.*

In Lemma 27, $\lambda_\phi$ is determined by the feature $\phi$, and $u(\theta)$ shows that the gradient is vanishing when $\pi_i$ is near deterministic, which is consistent with the fact that the sigmoid saturates and provides uninformative gradient as the parameter magnitude becomes large.




(a) $\mathcal{L}(\theta_t)$ and $\|\nabla\mathcal{L}(\theta_t)\|_2$ (b) Hessian spectral radius and NS

Figure 4.8: Experiments on GLM using GD.

We run GD on one example with $N = 10$ and $d = 2$. As shown in Fig. 4.8, the gradient norm $\|\nabla\mathcal{L}(\theta_t)\|_2$ is close to zero at plateaus and near optimum. However, unlike the PG, the spectral radius of the Hessian $\nabla^2\mathcal{L}(\theta_t)$ is only close to zero at plateaus, while it approaches positive constant near optimum. This indicates a different NS condition other than Lemmas 21 and 24 is needed,

since only gradient norm $\to 0$ cannot upper bound the spectral radius of Hessian $\to \beta > 0$. With some calculations, we prove the following key results:

**Lemma 28** (Smoothness and NS)**.** *$\mathcal{L}(\theta)$ satisfies $\beta$ smoothness with*

$$\beta = \frac{3}{8} \cdot \max_{i \in [N]} \|\phi_i\|_2^2, \tag{4.28}$$

*and $\beta(\theta)$ NS with*

$$\beta(\theta) = L_1 \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 + L_0 \cdot \left( \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta) \right). \tag{4.29}$$

At the optimal solution $\theta^*$, the spectral radius of the Hessian $\frac{\partial^2 \mathcal{L}(\theta^*)}{\partial (\theta^*)^2}$ is strictly positive. Therefore, the MSE objective of Eq. (4.22) is in the non-convex function class $\mathcal{Z}$ in Fig. 4.2, and it satisfies the case (2) in Theorem 16 with $\xi = 1/2$. Combining Lemmas 27 and 28 and applying Theorem 16, we have the following global linear convergence result:

**Theorem 20.** *With $\eta = 1/\beta$, GD update satisfies for all $t \geq 1$,*

$$\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_1) \cdot e^{-C^2 \cdot (t-1)}. \tag{4.30}$$

*With $\eta \in \Theta(1)$, GNGD update satisfies for all $t \geq 1$,*

$$\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_1) \cdot e^{-C \cdot (t-1)}, \tag{4.31}$$

*where $C \in (0, 1)$, i.e., GNGD is strictly faster than GD.*

Theorem 20 significantly improves the $O(1/\sqrt{t})$ rate in Theorem 19. The key difference is that we discovered a new NŁ inequality of Lemma 27 that is satisfied by GLMs.

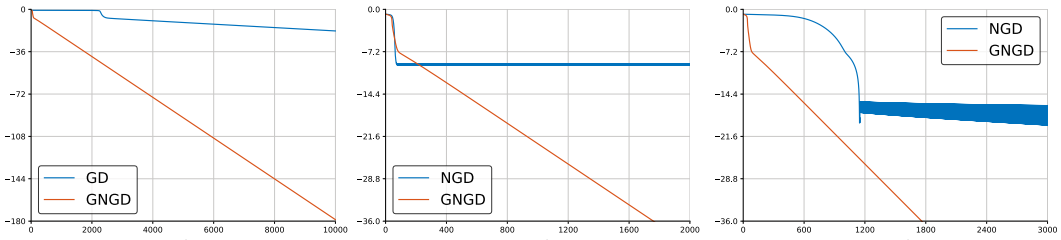
In Theorem 20, we have $C = \inf_{t \geq 1} C(\theta_t, \phi)$, which is very close to zero if $\pi_i$ is near deterministic, and GD suffers sub-optimality plateaus as shown in Fig. 4.1. GNGD has strictly (orders of magnitudes) better constant dependence $C \gg C^2$, and escapes plateaus significantly faster than GD. Intuitively, for the GLM in Fig. 4.1, $C$ in Theorem 16 is lower bounded reasonably if $\theta_1$ is initialized within some finite distance of the central valley containing $\theta^*$.

Combining the NŁ and NS properties (Lemmas 27 and 28), we provide new understandings of using normalization in GLM: *(i)* **First**, using standard NGD (Hazan et al., 2015) for all $t \geq 1$ is not a good choice. By examining the asymptotic behaviour as $\theta \to \theta^*$, we have $\beta(\theta) \to \beta > 0$. However, the normalization $\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2$ in standard NGD gives incremental updates with adaptive stepsize $\to \infty$. To guarantee convergence, it is necessary to use $\eta_t \to 0$, which counteracts normalization and slows down the learning, since it might not be easy to find a learning rate scheme. This is consistent with the $O(e^{-c \cdot t})$ result for GD with $\eta > 0$ and without normalization in Theorem 20. *(ii)* **Second**, using geometry-aware normalization $\beta(\theta_t)$ is a better choice than normalizing the gradient norm $\|\nabla \mathcal{L}(\theta_t)\|_2$. We elaborate this point by investigating *both the asymptotic and the early-stage behaviours* using NS-NŁ. Since $\beta(\theta_t) \to \beta > 0$ asymptotically, GNGD is approaching GD as $\theta_t \to \theta^*$, which makes GNGD enjoy the same $O(e^{-c \cdot t})$ rate. On the other hand, at early-stage optimization (e.g., close to initialization in Fig. 4.1), when $\theta_t$ is far from $\theta^*$, we have thus $\beta(\theta_t) \leq c \cdot \left\| \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \right\|_2$. Then GNGD is close to NGD, which guarantees strictly better progresses than GD. This is because of the progress of GNGD in each iteration at this time is about $\|\nabla \mathcal{L}(\theta_t)\|_2$, while the progress of GD is $\|\nabla \mathcal{L}(\theta_t)\|_2^2$, and the gradient norm is close to 0 on plateaus. Using NŁ of Lemma 27, GNGD will have strictly better constant dependence $C$ than $C^2$ in GD.

### 4.7.3 Empirical Verification

Theorem 20 proves linear convergence rates $O(e^{-c \cdot t})$ for both GD and GNGD on GLM. We compare GD, NGD (Hazan et al., 2015), and GNGD on GLM, as shown in Fig. 4.9.

Subfigure (a) presents the results of GD with $\eta = 0.09$ and GNGD with $\eta = 0.09$. Both GD and GNGD achieve linear $O(e^{-c \cdot t})$ rates, verifying Theorem 20. GD suffers from the plateaus at the early-stage optimization, which is consistent with Fig. 4.1 and the explanations after Theorem 20. On the other hand, the slopes indicate that GNGD converges strictly faster than GD, which justifies the constant dependences ($C \geq C^2$) in Theorem 20. Subfigure




(a) $\log \delta(\theta_t)$ (b) $\log \delta(\theta_t)$ (c) $\log \delta(\theta_t)$

Figure 4.9: Convergence rates for GD, NGD, and GNGD on GLM.

(b) shows that standard NGD (Hazan et al., 2015) with constant learning rate $\eta = 0.09$ does not converge. The NGD update keeps oscillating, which verifies our argument of using standard normalization for all $t \geq 1$ is not a good idea. Subfigure (c) presents the NGD using adaptive learning rate $\eta_t = \frac{0.09}{\sqrt{t}}$, which has faster convergence than NGD with constant $\eta$. However, GNGD still significantly outperforms NGD with $\eta_t = \frac{0.09}{\sqrt{t}}$, verifying the $O(e^{-c \cdot t})$ in Theorem 20 and $O(1/\sqrt{t})$ in Theorem 19.

## 4.8 Summary

The main contributions of this chapter concern a general characterization and analysis based on non-uniform properties, which are not only sufficiently general to cover concrete examples, but also significantly improve convergence rates over previous work and even over classical lower bounds. Importantly, the techniques apply to important applications in machine learning that involve non-convex optimization problems.

# Chapter 5

# Understanding Stochasticity in Policy Optimization

The results in Chapters 2 to 4 apply to true gradient settings. This chapter extends the analysis and investigates the algorithms in stochastic gradient settings. A new analytical tool, which I refer to as the committal rate, is introduced and reveals interesting properties of stochastic policy optimization.

The results in this chapter appeared in the paper Mei et al. (2021a), which has been submitted for review.

## 5.1 Introduction

Policy optimization is a central problem in reinforcement learning (RL) that provides a foundation for both policy-based and actor-critic RL methods. As shown in Chapters 2 to 4, recent findings indicate that policy gradient methods can indeed be guaranteed to converge to globally optimal solutions at least in the tabular setting, even if the policy value function is non-concave.

In particular, the standard softmax PG method with a constant learning rate has been shown to converge to a globally optimal policy at a $\Theta(1/t)$ rate for finite MDPs (Chapter 2), albeit with challenging problem and initialization dependent constants (Chapter 3). Several techniques have been developed to further improve standard PG and achieve better rates and constants. For example, adding *entropy regularization* has been shown to produce faster $O(e^{-c\cdot t})$ convergence ($c > 0$) to the optimal regularized policy (Chapter 3 and Cen et al.

(2020), Lan (2021), and Mei et al. (2020b)). By exploiting natural geometries based on Bregman divergences, *natural PG (NPG) or mirror descent (MD)* have been shown to achieve better constants than standard PG (Agarwal et al., 2019; Cen et al., 2020) and faster $O(e^{-c \cdot t})$ rates, with (Cen et al., 2020; Lan, 2021) and without regularization (Khodadadian et al., 2021). Alternative policy parameterizations, such as the escort parameterization, have been shown to improve the constants achieved by softmax and yield faster plateau escaping (Chapter 3). A *geometry-aware normalized PG (GNPG)* approach has been proposed to exploit the non-uniformity of the value function, achieving even faster $O(e^{-c \cdot t})$ rates with improved constants (Chapter 4).

A key observation is that each of these four techniques—*(i)* entropy regularization, *(ii)* NPG (or MD), *(iii)* alternative escort policy parameterization, and *(iv)* GNPG—accelerate the convergence of standard softmax PG by better exploiting the **geometry** of the optimization landscape. In particular, entropy regularization makes the regularized objective behave more like a quadratic (Lemma 15), which significantly improves the near-linear character of the softmax policy value (Lemma 18). Natural PG (or MD) perform non-Euclidean updates in the parameter space, which is quite different from the Euclidean geometry characterizing standard softmax PG updates. The escort policy parameterization induces an alternative policy-parameter relation (Lemma 20). GNPG exploits the non-uniform smoothness in the optimization landscape via a simple gradient normalization operation (Lemma 24).

However, these advantages have only been established for the true gradient setting. A natural question therefore is whether geometry can also be exploited to accelerate convergence to global optimality in *stochastic* gradient settings. In this chapter, we show that in a certain fundamental sense, the answer is *no*. That is, there exists a fundamental trade-off between leveraging geometry to accelerate convergence and overcoming the noise introduced by stochastic gradients (possibly infinite); in particular, no uninformed algorithm can improve the $O(1/t)$ convergence rate without incurring a positive probability of failure (i.e. diverging or converging to a sub-optimal stationary point).

The conditions used in vanilla stochastic gradient convergence analysis, *i.e.*,

unbiased and variance-bounded gradient estimator (Nemirovski et al., 2009), has been exploited to attempt to explain such a trade-off in policy gradients (Abbasi-Yadkori et al., 2019; Lan, 2021). However, the bounded variance requires the sample policy to be bounded away from zero everywhere, which is impractical. Meanwhile, a variant of NPG can converge even with unbounded variance (Chung et al., 2020). These gaps raise the question that if not the bounded variance, then what is the key factor to ensure the convergence of stochastic policy optimization algorithms? Motivated by this question, we introduce the concept *committal rate* to characterize the behavior policies, which significantly affect whether convergence to a correct solution can be guaranteed in the stochastic on-policy setting. In particular, we make the following contributions.

- *First*, we illustrate the anomaly that the preferability of policy optimization algorithms (softmax PG vs. NPG and GNPG) changes dramatically depending on whether true versus on-policy stochastic gradients are considered, and reveal the impracticality and unnecessity of a bounded variance requirement in Section 5.2;
- *Second*, we introduce the concept of the *committal rate* in Section 5.3 to characterize how quickly a sampled action's probability approaches 1, which provide us tools for analyzing the stochasticity effect in convergences;
- *Third*, we use the committal rate to study general stochastic policy optimization behaviors rigorously and reveal the inherent geometry-convergence trade-off in Section 5.4;
- *Finally*, we explain the sensitivity to random initialization in practical policy optimization algorithms. From these results, we then develop an ensemble method that can achieve fast convergence to global optima with high probability in Section 5.5.

## 5.2 Understanding Algorithm Preferability in On-line Policy Optimization

To illustrate the key aspects of policy optimization methods and their comparative preferability, it suffices to consider deterministic, single-state, finite-action Markov decision processes (MDPs). The main results below extend to general finite MDPs, but for clarity of exposition we restrict attention to one-state MDPs.

A deterministic, single-state, finite-action MDP can be simply be specified by an action space is $[K] \coloneqq \{1, 2, \dots K\}$ and a $K$-dimensional reward vector $r \in \mathbb{R}^K$. The problem is to maximize the expected reward of a parametric policy $\pi_\theta$ as in Eq. (2.12),

$$\max_{\theta:[K]\to\mathbb{R}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} [r(a)], \tag{5.1}$$

where $\pi_\theta$ is parameterized by $\theta$ using the standard softmax transform as in Eq. (2.6),

$$\pi_\theta(a) = \frac{\exp\{\theta(a)\}}{\sum_{a'} \exp\{\theta(a')\}}, \quad \forall a \in [K]. \tag{5.2}$$

Without loss of generality, we assume there exists a unique optimal action

$$a^* = \arg\max_{a\in[K]} r(a), \tag{5.3}$$

hence there exists a unique optimal deterministic policy $\pi^*$ such that

$$\pi^{*\top} r = \sup_{\theta\in\mathbb{R}^K} \pi_\theta^\top r = r(a^*). \tag{5.4}$$

We make the following assumption on the reward.

**Assumption 3** (Positive reward)**.** $r(a) \in (0, 1]$, $\forall a \in [K]$.

### 5.2.1 Exact Gradient Setting

As shown in Proposition 1, Eq. (2.12) is a non-concave maximization over the policy parameter $\theta$. Nevertheless, it has recently become better understood how policy gradient (PG) methods still converge to global optima for Eq. (2.12)

when exact gradients are used. To illustrate the main considerations, we focus on the following three representative algorithms that have recently been proved to achieve convergence to globally optimal solutions but at different rates: softmax policy gradient (PG) in Chapter 2, natural PG (NPG), and geometry-aware normalized PG (GNPG) in Chapter 4, while similar conclusions can be drawn for other variants (Chung et al., 2020; Denisov and Walton, 2020).

**Standard Softmax PG**

The standard softmax PG method is specified by Update 1,

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \tag{5.5}$$

where

$$\frac{d\pi_\theta^\top r}{d\theta} = \left(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top\right) r, \tag{5.6}$$

and thus

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r), \quad \forall a \in [K]. \tag{5.7}$$

As shown in Chapter 2, the convergence of this update to a globally optimal policy, given exact gradients, can be established by considering the following non-uniform Łojasiewicz (NŁ) inequality of Lemma 3,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \tag{5.8}$$

By considering smoothness of the optimization landscape, Chapter 2 then shows that the progress in each iteration of PG can be lower bounded by the squared norm of the gradient, $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2$, which leads to a $O(1/t)$ rate.

**Proposition 11** (PG upper bound, Theorem 2)**.** *Using Update 1 with $\eta = 2/5$, we have*

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(c^2 \cdot t), \tag{5.9}$$

*for all $t \geq 1$, such that $c = \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ is a constant that depends on $r$ and $\theta_1$, but it does not depend on the time $t$. In particular, if $\pi_{\theta_1}(a) = 1/K$ $\forall a$ then $c \geq 1/K$.*

**Proposition 12** (PG lower bound, Theorem 9)**.** *For sufficiently large $t \geq 1$, Update 1 with $\eta \in (0, 1]$ exhibits*

$$(\pi^* - \pi_{\theta_t})^\top r \geq \Delta^2 / (6 \cdot t)\,, \tag{5.10}$$

*where $\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0$ is the reward gap of $r$.*

**Remark 16.** *According to Theorem 11, the constant dependence of PG follows a $\Omega(1/c)$ lower bound for one-state MDPs, while $c$ can be exponentially small in terms of the number of states for general finite MDPs (Li et al., 2021).*

To summarize, using $\eta \in O(1)$, softmax PG achieves convergence to a global optima, but with a $\Theta(1/t)$ rate that exhibits poor constant dependence.

### Natural PG (NPG)

An alternative method, natural PG (NPG) (Kakade, 2002), provides the prototype for many practical policy optimization methods, such as Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) (Schulman et al., 2015; Schulman et al., 2017). NPG is based on the following update.

**Update 4** (Natural PG (NPG), true gradient)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot r, \;\; \textit{and} \tag{5.11}$$

$$\pi_{\theta_{t+1}} = \text{softmax}(\theta_{t+1}). \tag{5.12}$$

For softmax policies, it turns out that Update 4 is identical to mirror descent (MD) with a Kullback-Leibler (KL) divergence. Therefore a standard MD analysis shows that Update 4 achieves convergence to a global optimum at a rate of $O(1/t)$ (Agarwal et al., 2019). Very recently, work concurrent to this thesis (Khodadadian et al., 2021) has shown that Update 4 actually enjoys a much faster $O(e^{-c \cdot t})$ rate. In fact, here too we can establish the same $O(e^{-c \cdot t})$ rate, but using a simpler argument based on the following variant of the NŁ inequality for natural gradients. These results are new.

**Lemma 29** (Natural NŁ inequality, continuous)**.** *We have,*

$$\left\langle \frac{d\pi_\theta^\top r}{d\theta}, r \right\rangle \geq \pi_\theta(a^*) \cdot \Delta \cdot (\pi^* - \pi_\theta)^\top r. \tag{5.13}$$

**Lemma 30** (Natural NŁ, discrete)**.** *Let* $\pi'(a) \coloneqq \frac{\pi(a) \cdot e^{\eta \cdot r(a)}}{\sum_{a'} \pi(a') \cdot e^{\eta \cdot r(a')}}$, $\forall\ a \in [K]$, *where* $\eta > 0$. *Then,*

$$(\pi' - \pi)^\top r \geq \left[1 - \frac{1}{\pi(a^*) \cdot (e^{\eta \cdot \Delta} - 1) + 1}\right] \cdot (\pi^* - \pi)^\top r. \tag{5.14}$$

In particular, by using a non-Euclidean update and analysis, the progress of each iteration of NPG can be lower bounded by the larger bound $\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, r \right\rangle$ instead of the weaker bound $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2$ established for standard PG. Based on this inequality, one can easily establish a much faster $O(e^{-c \cdot t})$ convergence to a globally optimal solution for NPG, making it far preferable to PG if true gradients are available.

**Theorem 21** (NPG upper bound)**.** *Using Update 4 with any* $\eta > 0$, *we have, for all* $t \geq 1$,

$$(\pi^* - \pi_{\theta_t})^\top r \leq (\pi^* - \pi_{\theta_1})^\top r \cdot e^{-c \cdot (t-1)}, \tag{5.15}$$

*where* $c \coloneqq \log\left(\pi_{\theta_1}(a^*) \cdot \left(e^{\eta \cdot \Delta} - 1\right) + 1\right) > 0$ *for any* $\eta > 0$.

### Geometry-aware Normalized PG (GNPG)

The Geometry-aware Normalized PG (GNPG) update is investigated in Chapter 4 to accelerate the convergence of PG by exploiting local smoothness properties of the optimization landscape. GNPG is specified by

**Update 5** (Geometry-aware Normalized PG (GNPG), true gradient)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2. \tag{5.16}$$

The analysis in Chapter 4 focuses on exploiting non-uniform smoothness (NS) rather than improving the NŁ inequality as for NPG above. According to Lemma 21, the spectral radius of Hessian matrix $\frac{d^2 \pi_\theta^\top r}{d\theta^2}$ is upper bounded by $3 \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$.

Given this NS property, Chapter 4 then shows how the progress in each iteration of GNPG can be lower bounded by the larger quantity $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$ instead of the weaker $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2$ bound for standard PG. Then, using the same NŁ inequality as for PG, Theorem 17 shows that GNPG also converges to a globally optimal solution at rate $O(e^{-c \cdot t})$. Again, one naturally concludes that GNPG is preferable to PG if exact gradients are used.

**Proposition 13** (GNPG upper bound, Theorem 17)**.** *Using Update 5 with* $\eta = 1/6$*, we have, for all* $t \geq 1$*,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq (\pi^* - \pi_{\theta_1})^\top r \cdot e^{-\frac{c \cdot (t-1)}{12}}, \tag{5.17}$$

*where* $c = \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ *does not depend on* $t$*. If* $\pi_{\theta_1}(a) = 1/K$*,* $\forall a$*, then* $c \geq 1/K$*.*

### 5.2.2 On-policy Stochastic Gradient Setting: Anomalies

Although the above results show that exploiting geometric information can allow linear convergence to an optimal solution given true gradients—obviously $O(e^{-c \cdot t})$ represents an exponential speedup over the $\Omega(1/t)$ lower bound established for standard PG—it is critical to understand whether such advantages can also be obtained in the more natural stochastic gradient setting. Given the previous results, it would seem natural to prefer accelerated algorithms over PG in practice, and there is some evidence that such thinking has become mainstream based on the popularity of TRPO and PPO over PG. However, by more closely examining the behavior of these algorithms when true gradients are replaced by on-policy stochastic estimates, serious shortcomings begin to emerge, as empirically observed in Chung et al. (2020), and it is far from obvious that similar advantages from the true gradient case might be recoverable in the more practical stochastic scenario.

We begin by examining the behavior of the previous algorithms in the context of on-policy stochastic gradients. To enable this analysis, first note that each of the above PG methods, Updates 1, 4 and 5, can be adapted to the

stochastic setting by using on-policy importance sampling (IS) to provide an unbiased estimate of the true reward. We do not make assumptions like each action is sufficiently explored, since $\pi_{\theta_t}$ is the behaviour policy as well as the policy to be optimized. It is possible that $\pi_{\theta_t}$ approaches a near deterministic policy, ruling out positive results based on such assumptions (Abbasi-Yadkori et al., 2019).

**Definition 10** (On-policy IS)**.** *At iteration $t$, sample one action $a_t \sim \pi_{\theta_t}$. The IS reward estimator $\hat{r}_t$ is constructed as $\hat{r}_t(a) = \frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)} \cdot r(a)$ for all $a \in [K]$.*

**Remark 17.** *We consider sampling one action in each iteration for convenience, and the results hold for sampling a constant $B > 0$ mini-batch of actions in each iteration.*

### Softmax PG

**Update 6** (Softmax PG, on-policy stochastic gradient)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}, \tag{5.18}$$

*where*

$$\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)} = \pi_{\theta_t}(a) \cdot (\hat{r}_t(a) - \pi_{\theta_t}^\top \hat{r}_t), \quad \forall a \in [K]. \tag{5.19}$$

Using IS estimation, the softmax PG is unbiased and its variance is upper bounded by constant.

**Lemma 31.** *Let $\hat{r}$ be the IS estimator using on-policy sampling $a \sim \pi_\theta(\cdot)$. The stochastic softmax PG estimator is unbiased and bounded, i.e.,*

$$\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)} \left[ \frac{d\pi_\theta^\top \hat{r}}{d\theta} \right] = \frac{d\pi_\theta^\top r}{d\theta}, \text{ and} \tag{5.20}$$

$$\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)} \left\| \frac{d\pi_\theta^\top \hat{r}}{d\theta} \right\|_2^2 \leq \frac{K}{2}. \tag{5.21}$$

These observations imply that stochastic softmax PG converges to a global optimum almost surely, since the stochastic update follows the true gradient update with controlled noise, which was also proved by Chung et al. (2020).

**Theorem 22.** *Using Update 6, $(\pi^* - \pi_{\theta_t})^\top r \to 0$ as $t \to \infty$ with probability 1.*

### NPG

Similarly, we can use on-policy IS estimation to adapt NPG to the stochastic setting.

**Update 7** (NPG, on-policy stochastic gradient)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \hat{r}_t, \text{ and} \tag{5.22}$$

$$\pi_{\theta_{t+1}} = \text{softmax}(\theta_{t+1}). \tag{5.23}$$

Although the NPG is unbiased, its variance can be possibly unbounded in the on-policy setting, as established by the following direct calculation.

**Lemma 32.** *For NPG, we have,* $\mathbb{E}_{a\sim\pi_\theta(\cdot)}[\hat{r}] = r$, *and* $\mathbb{E}_{a\sim\pi_\theta(\cdot)}\|\hat{r}\|_2^2 = \sum_{a\in[K]} \frac{r(a)^2}{\pi_\theta(a)}$.

Note that if $\pi_\theta(a) \to 0$, the variance becomes unbounded, which predicts trouble if someone tried to use the standard analysis for stochastic gradient methods[1] (e.g., Nemirovski et al. (2009)). In fact, we provide a more direct result showing that stochastic NPG has a positive probability of converging to a sub-optimal deterministic policy.

**Theorem 23.** *Using Update 7, we have:* ***(i)*** *with positive probability, as* $t \to \infty$, $\sum_{a\neq a^*} \pi_{\theta_t}(a) \to 1$*;* ***(ii)*** $\forall a \in [K]$, *with positive probability,* $\pi_{\theta_t}(a) \to 1$, *as* $t \to \infty$.

This result extends the result of Chung et al. (2020) who considered the two-action ($K = 2$) case only. The intuition is that the stochastic NPG accumulates too much probability on sampled sub-optimal actions and cannot recover due to the "vicious circle" between sampling and updating (Chung et al., 2020).

### GNPG

Finally, we consider the stochastic version of GNPG.

[1] Standard treatment of stochastic approximation algorithms does deal with unbounded noise in a controlled way to still get positive results (Benveniste et al., 2012), which means that bounded variance is far from being necessary.

**Update 8** (GNPG, on-policy stochastic gradient)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2. \tag{5.24}$$

Unfortunately, this estimator involves a ratio of random variables, and its bias can be large. As for NPG we can show that stochastic GNPG fails with positive probability in the stochastic case.

**Theorem 24.** *Using Update 8, we have:* ***(i)*** *with positive probability, as* $t \to \infty$, $\sum_{a \neq a^*} \pi_{\theta_t}(a) \to 1$*;* ***(ii)*** $\forall a \in [K]$*, with positive probability,* $\pi_{\theta_t}(a) \to 1$*, as* $t \to \infty$*.*

### 5.2.3 Motivating the On-policy Stochastic Setting

We summarize the preferability of alternative optimization strategies in the exact versus on-policy stochastic gradient settings in Table 5.1: there appears to be a major reversal in going from one scenario to the other.

| | Softmax PG | NPG | GNPG |
|---|---|---|---|
| True gradient | converges $\Theta(1/t)$ | converges $O(e^{-c \cdot t})$ | converges $O(e^{-c \cdot t})$ |
| Stochastic on-policy | converges w.p. 1 | fails w.p. $> 0$ | fails w.p. $> 0$ |

Table 5.1: Convergence properties of softmax PG, NPG and GNPG in the alternative settings.

Of course it is possible to study the convergence of algorithms when the gradient estimates are assumed to be unbiased and have bounded variance as in the analysis of vanilla SGD (Nemirovski et al., 2009), and in some other work for policy gradients (Abbasi-Yadkori et al., 2019; Lan, 2021; Zhang et al., 2021; Zhang et al., 2020a; Zhang et al., 2020b). However, **first**, such conditions are only *sufficient* conditions which are difficult to be satisfied, because a bounded variance assumption requires that the probabilities induced by a behaviour policy are bounded away from 0 everywhere (Chung et al., 2020), which is impractical for large state and action spaces and impossible when they are infinite. **Second**, PPO (Schulman et al., 2017) and TRPO

(Schulman et al., 2015) use on-policy sampling *without* any explicit correction to fulfill the exploratory behaviour requirement, while still solving practical problems (Andrychowicz et al., 2020). **Third**, the on-policy setting under a practical requirement is technically more challenging. In off-policy settings, the sampling procedure is independent of the parameter update (Chung et al., 2020), which makes the analysis much more straightforward with some extra assumptions on the estimate of value functions or gradients (Ren et al., 2021), while these become coupled in the on-policy setting and a more subtle analysis is required. **Finally**, Chung et al. (2020) investigate baselines and show that variance reduction techniques are not able to overcome unbounded variance, while NPG can still achieve global convergence almost surely with a judicious choice of baseline even though its variance remains *unbounded* (see Update 9 for details). This means bounded variance is not necessary for convergence, and some other factors rather than variance account for the convergence behaviours of stochastic policy optimization algorithms.

This leave us an important question to be answered to bridge the gap between theory and practice,

*What are the key factors determining the convergence of stochastic policy optimization?*

We propose the *committal rate* to characterize the behavior of algorithms to answer this question.

## 5.3 Committal Rate of Stochastic Policy Optimization Algorithms

Although the baseline study (Chung et al., 2020) only focuses on two- and three-action bandits primarily, it develops a useful intuition that stochastic policy optimization in practical settings consists of separate "sampling" and "updating" steps that become coupled in the on-policy setting. Building from this observation, and seeking to explain the outcomes in Section 5.2, we formalize the following "committal rate" function of a policy optimization algorithm. The main idea is to decouple the "sampling" and "updating" by fixing

sampling one action and characterizing the aggressiveness of an update in a deterministic way. Thus, in what follows, by a policy optimization algorithm $\mathcal{A}$ we mean a mapping from all sequences of pairs of action-reward pairs to the set of parameter vectors.

**Definition 11** (Committal Rate)**.** *Fix a reward function $r \in (0,1]^K$ and an initial parameter vector $\theta_1 \in \mathbb{R}^K$. Consider a policy optimization algorithm $\mathcal{A}$. Let action $a$ be the sampled action **forever** after initialization and let $\theta_t$ be the resulting parameter vector obtained by using $\mathcal{A}$ on the first $t$ observations. The committal rate of algorithm $\mathcal{A}$ on action $a$ (given $r$ and $\theta_1$) is then defined as*

$$\kappa(\mathcal{A}, a) = \sup\left\{\alpha \geq 0 : \limsup_{t\to\infty} t^\alpha \cdot [1 - \pi_{\theta_t}(a)] < \infty\right\}. \tag{5.25}$$

Note that in the definition we have suppressed the dependence of $\kappa$ on the rewards and the initial parameter vector. Definition 11 accounts for **how aggressive an update rule is**: An algorithm with committal rate $\alpha$ will make $\pi_{\theta_t}(a)$ approach 1 at the polynomial rate of $1/t^\alpha$ provided that the sampling rule only chooses action $a$. Thus, a larger value of $\kappa(\mathcal{A}, a)$ indicates an algorithm that quickly commits to the action $a$. For example, if $\pi_{\theta_t}(a) = 1 - 1/(t \cdot \log(t))$, then $\kappa(\mathcal{A}, a) = 1$. Similarly, if $\pi_{\theta_t}(a) = 1 - 1/e^t$, then $\kappa(\mathcal{A}, a) = \infty$, which means $\pi_{\theta_t}(a)$ approaches 1 extremely quickly. On the other hand, if $1 - \pi_{\theta_t}(a) \in \Omega(1)$, then $\kappa(\mathcal{A}, a) = 0$, implying that $\pi_{\theta_t}$ never becomes committal, since $\pi_{\theta_t}(a)$ never approaches 1.

Our next results shows that a small committal rate with respect to sub-optimal actions is necessary for almost sure convergence to a globally optimal policy.

**Theorem 25.** *Consider a policy optimization method $\mathcal{A}$. Fix $r \in (0,1]^K$ an action $a \in [K]$ which is sub-optimal under $r$ so that $\kappa(\mathcal{A}, a) > 1$. Fix $\theta_1 \in \mathbb{R}^K$ so that $\pi_{\theta_1}(a) > 0$ and let $\{\theta_t\}_{t\geq 1}$ be the parameter sequence obtained by using $\mathcal{A}$ with online sampling, i.e., when $a_t \sim \pi_{\theta_t}(\cdot)$. Then, the event $\mathcal{E} = \{a_t = a \text{ holds for all } t \geq 1\}$ happens with positive probability, and it also holds that $\pi_{\theta_t}$ converges to a sub-optimal deterministic policy with positive probability.*

Theorem 25 shows that $\max_{a:r(a)<r(a^*)} \kappa(\mathcal{A}, a) \leq 1$ is a necessary condition for ensuring the almost sure convergence of the policies obtained using $\mathcal{A}$ and online sampling to a global optimum. In words, slow reaction to constantly sampling sub-optimal actions is necessary for the success of policy optimization methods when they are used with online sampling.

Using this result, we can now interrogate the committal rates of the previously listed algorithms.

**Theorem 26.** *Let Assumption 3 holds. For the stochastic updates NPG and GNPG from Updates 7 and 8 we obtain* $\kappa(\mathit{NPG}, a) = \infty$ *and* $\kappa(\mathit{GNPG}, a) = \infty$ *for all* $a \in [K]$ *respectively.*

Theorem 26 explains why stochastic NPG and GNPG have a non-zero failure probability in the on-policy stochastic setting: they do not obey a necessary condition for almost sure global convergence. Intuitively, these algorithms can fail by prematurely allocating too much probability to a sub-optimal action: each sampling of an action $a \in [K]$ increments its parameter by $\Theta(1)$, so if $a$ is sampled $t$ times successively, then we have $1 - \pi_{\theta_t}(a) \in O(e^{-c \cdot t})$, which means $\kappa(\mathcal{A}, a) = \infty$. According to Theorem 25, there is a positive probability that a single sub-optimal action can receive a long enough sampling run to ensure the other actions will never again be sampled.

By contrast, we can compare these outcomes to the committal rate of the softmax PG algorithm.

**Theorem 27.** *Let* $r(a) > 0$ *and* $\pi_{\theta_1}(a) > 0$. *Softmax PG obtains* $\kappa(\mathit{PG}, a) = 1$ *for all* $a \in [K]$.

Theorems 25 and 27 provide (partial) explanations of the observations in Section 5.2: stochastic NPG and GNPG can fail while PG almost surely converges to a global optimum, but their committal rates lie on different sides of the necessary condition. Since $\kappa(\mathrm{PG}, a) = 1$ for softmax PG, it follows that $\prod_{t=1}^{\infty} \pi_{\theta_t}(a) = 0$ (see Lemma 62), hence it is not possible to sample sub-optimal actions forever, and the optimal action $a^*$ always has a sufficient chance to be sampled, which ensures learning.

Next, we consider NPG using special baselines (Chung et al., 2020), which enjoys almost sure global convergence but has unbounded variance, and this result cannot be explained by the standard variance-based analysis using bounded variance assumption and decaying learning rates.

**Update 9** (NPG with oracle baseline)**.**

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \left(\hat{r}_t - \hat{b}_t\right), \tag{5.26}$$

*where*

$$\hat{b}_t(a) = \left(\frac{\mathbb{I}\left\{a_t = a\right\}}{\pi_{\theta_t}(a)} - 1\right) \cdot b, \quad \forall a \in [K], \tag{5.27}$$

*and* $b \in (r(a^*) - \Delta, r(a^*))$.

**Theorem 28.** *Using Update 9,* $\left(\pi^* - \pi_{\theta_t}\right)^\top r \to 0$ *as* $t \to \infty$ *with probability* 1.

For Update 9, the variance is still unbounded (Chung et al., 2020), while the learning rate is not decaying. Thus the convergence is **not** due to the analysis based on bounded variance with decaying learning rate. However, the necessary condition in Theorem 25 is satisfied. We have, $\pi_{\theta_{t+1}}(a_t) < \pi_{\theta_t}(a_t)$, if $a_t \neq a^*$, i.e., whenever a sub-optimal action is selected its probability decreases, while the optimal action's probability always increases after any update. Therefore, we have $\kappa(\mathcal{A}, a^*) = \infty$ and $\kappa(\mathcal{A}, a) = 0$ for all $a \neq a^*$. This example means committal rate works for cases beyond the bounded variance condition used widely in the optimization and reinforcement learning communities.

## 5.4 The Geometry-Convergence Trade-off in Stochastic Policy Optimization

Theorem 27 raises the question of whether $\kappa(\mathcal{A}, a) = 1$ for all sub-optimal actions $a \in [K]$ is sufficient to ensure an algorithm $\mathcal{A}$ converges to an optimal policy almost surely. Unfortunately, this is not the case, and the complete picture of global optimality in stochastic policy optimization is more complex and requires detailed study of different iteration behaviors.

### 5.4.1 Iteration Behaviours

**Remark 18.** *The condition that $\kappa(\mathcal{A}, a) \leq 1$ for all sub-optimal actions $a \in [K]$ is **not** sufficient for ensuring almost sure convergence to global optimality. In addition to "convergence to a sub-optimal policy with positive probability" and "convergence to a globally optimal policy with probability 1" there exist other possible optimization behaviours, such as "not converging to any policy".*

In particular, consider the following update behaviors.

**Staying.** For the stationary update $\mathcal{A} : \theta_{t+1} \leftarrow \theta_t$ we obtain $\kappa(\mathcal{A}, a) = 0 \leq 1$ for all $a \in [K]$, yet $\pi_{\theta_t} = \pi_{\theta_1}$ does not converge to the optimal policy nor any sub-optimal deterministic policy.

**Wandering** (NPG with a large baseline)**.** Consider

$$\mathcal{A} : \theta_{t+1} \leftarrow \theta_t + \eta \cdot \left(\hat{r}_t - \hat{b}_t\right), \tag{5.28}$$

with $\hat{b}_t(a) = \left(\frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} - 1\right) \cdot b$ for all $a \in [K]$. If $b > r(a^*)$, then we have $\pi_{\theta_{t+1}}(a_t) < \pi_{\theta_t}(a_t)$, i.e., a selected action's probability will decrease after updating, hence $\kappa(\mathcal{A}, a) = 0$ for all $a \in [K]$. However, $\pi_{\theta_t}(a) \not\to 1$ as $t \to \infty$ for all $a \in [K]$, therefore $\pi_{\theta_t}$ will wander within the simplex forever.

The above examples show that not converging to a sub-optimal policy does not necessarily imply converging to an optimal policy almost surely, and a stronger condition is needed to eliminate unreasonable behaviors like $\theta_{t+1} \leftarrow \theta_t$. We leave it as an open question to identify necessary and sufficient conditions for almost sure convergence to a global optimum.

### 5.4.2 Geometry-Convergence Trade-off

In Section 5.2 we saw that NPG and GNGP can use true gradients to significantly accelerate PG by better exploiting geometry. However, in the stochastic setting, any estimated geometry might be inaccurate, and intuitively, accelerated methods risk leveraging inaccurate information too aggressively. On the one hand, if progress is sufficiently fast (i.e., with a large committal rate),

then an algorithm might never recover from aggressive yet inaccurate updates (Theorem 25). On the other hand, large progress is necessary for fast convergence. The tension between these observations suggest that there might be an inherent trade-off between exploiting geometry and avoiding premature convergence in stochastic policy optimization. We formalize this intuition with the following results. For the first result, we need to restrict to the class of policy optimization methods that do not decrease the probability of the optimal action whenever that action is chosen: In particular, a policy optimization method is said to be *optimality-smart* if for any $t \geq 1$, $\pi_{\tilde{\theta}_t}(a^*) \geq \pi_{\theta_t}(a^*)$ holds where $\tilde{\theta}_t$ is the parameter vector obtained when $a^*$ is chosen in every time step, starting at $\theta_1$, while $\theta_t$ is *any* parameter vector that can be obtained with $t$ updates (regardless of the action sequence chosen), but also starting from $\theta_1$.

**Theorem 29.** *Let $\mathcal{A}$ be optimality-smart and pick a bandit instance. If $\mathcal{A}$ together with on-policy sampling leads to $\{\theta_t\}_{t\geq 1}$ such that $\{\pi_{\theta_t}\}_{t\geq 1}$ converges to a globally optimal policy at a rate $O(1/t^\alpha)$ with positive probability, for $\alpha > 0$, then $\kappa(\mathcal{A}, a^*) \geq \alpha$.*

This theorem implies that a large committal rate for the optimal action is necessary for achieving fast convergence to the globally optimal policy, since the sub-optimality dominates how close the optimal action's probability is to 1, i.e.,

$$(\pi^* - \pi_{\theta_t})^\top r = \sum_{a \neq a^*} \pi_{\theta_t}(a) \cdot (r(a^*) - r(a)) \tag{5.29}$$

$$\geq (1 - \pi_{\theta_t}(a^*)) \cdot \Delta. \tag{5.30}$$

Therefore $(\pi^* - \pi_{\theta_t})^\top r \in O(1/t^\alpha)$ implies $1 - \pi_{\theta_t}(a^*) \in O(1/t^\alpha)$. Combining this result with Theorem 25 formally establishes the following inherent trade-off between exploiting geometry to accelerate convergence versus achieving global optimality almost surely (aggressiveness vs. stability).

**Theorem 30** (Geometry-Convergence trade-off)**.** *If an algorithm $\mathcal{A}$ is optimality-smart, and $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one $a \neq a^*$, then $\mathcal{A}$ with on-policy sampling can only exhibit at most one of the following two behaviors:*

**(i)** *$\mathcal{A}$ converges to a globally optimal policy almost surely;*

**(ii)** *$\mathcal{A}$ converges to a deterministic policy at a rate faster than $O(1/t)$ with positive probability.*

In other words, if $\mathcal{A}$ has a chance to converge to a global optimum, then either $\mathcal{A}$ converges to the globally optimal policy with probability 1 ($\mathcal{A}$ is stable) but at a rate no better than $O(1/t)$, or it achieves a faster than $O(1/t)$ convergence rate ($\mathcal{A}$ is aggressive) but fails to converge to the globally optimal policy with some positive probability. This trade-off between the geometry and convergence is faced by any stochastic policy optimization algorithm that is not informed by external oracle information that allows it to distinguish optimal and sub-optimal actions based on on-policy samples.

**Remark 19.** *Theorem 30 means an algorithm can achieve at most one of the mentioned two results. It is possible that an algorithm achieves neither (e.g., staying or wandering).*

### 5.4.3 Exploiting External Information

In Theorem 30, the condition of $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action $a \in [K]$ is necessary for the trade-off to hold. If this condition can somehow be bypassed, for example, by providing problem specific information, then it is possible to simultaneously achieve faster rates and almost sure convergence to a global optimum. For example, consider the NPG with oracle baseline of Update 9. As mentioned before, we have $\kappa(\mathcal{A}, a^*) = \infty$ and $\kappa(\mathcal{A}, a) = 0$ for all $a \neq a^*$, breaking the mentioned condition, which allows $\mathcal{A}$ to enjoy almost sure global convergence as well as a $O(e^{-c \cdot t})$ rate. Of course, such a fortuitous outcome required a very specific baseline that is aware of both the optimal reward and the reward gap. Without introducing external mechanisms that inform an on-policy algorithm it appears that such information cannot be recovered sufficiently quickly from sample data alone (Tucker et al., 2018). Nevertheless, it remains an open question to prove that this is not possible, or whether some other strategy might allow an on-policy stochas-

tic policy optimization algorithm to avoid the condition of Theorem 30 and achieve both fast rates and almost sure global convergence.

| Property I | $\kappa(\mathcal{A}, a) > 1$ | $\kappa(\mathcal{A}, a) \leq 1$, for all sub-optimal action $a \in [K]$ | | |
|---|---|---|---|---|
| Algorithm | NPG<br>GNPG | Softmax PG<br>SAMBA | Staying<br>Wandering<br>(NPG + large baseline) | NPG + oracle baseline |
| Property II | $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$, for at least one sub-optimal action $a \in [K]$ | | | $\kappa(\mathcal{A}, a^*) \neq \kappa(\mathcal{A}, a)$ |

Figure 5.1: Different algorithmic behaviours subdivided by two properties of committal rate. SAMBA does not use parametric policies and is discussed below.

Fig. 5.1 summarizes all the iteration behaviours we studied in this chapter, organized by two properties of committal rate: *(i)* possible failure if $\kappa(\mathcal{A}, a) > 1$ for at least one sub-optimal action $a$; and *(ii)* an inherent geometry-convergence trade-off if $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action $a$. It remains open to study where other algorithms suit themselves in this diagram.

## 5.5 Initialization Sensitivity and Ensemble Methods

In this section, we will keep exploiting the newly introduced concept, committal rate, to further reveal mystery observed in practice about the sensitivity of the initialization in policy optimization (Henderson et al., 2018). With the understanding of this unavoidable phenomenon, we introduce ensemble method and quantitatively characterize the successful rate in terms of number of trials.

### 5.5.1 Initialization Sensitivity

It has been observed empirically that RL algorithms are sensitive to initialization in practice: the same algorithm can produce remarkably different performance given different random seeds (Henderson et al., 2018). Chapter 3 has attempted to explain initialization sensitivity due to the softmax transform, but such results only hold for true gradients and apply to standard PG methods.

Using the committal rate theory developed above, we can provide a new explanation and additional understanding of the initialization sensitivity of practical policy optimization algorithms. Most well-performing policy optimization algorithms in practice, such as TRPO and PPO (Schulman et al., 2015; Schulman et al., 2017), are based on NPG, which exploits geometry to accelerate PG in true gradient settings. However, according to Theorem 30, such fast convergence must incur a positive probability of failing to reach a global optimum, even in bandit settings. Therefore, the need to attempt multiple random seeds to achieve success is an unavoidable consequence of using these algorithms according to this theory.

### 5.5.2 Ensemble Methods

The committal rate theory also explains why ensemble methods (Jung et al., 2020; Parker-Holder et al., 2020; Wiering and Van Hasselt, 2008), *i.e.*, running a policy optimization algorithm in multiple parallel threads and picking the best performing one, can provably work well. This is because a fast algorithm for the true gradient setting can have a positive probability of success or failure across different initializations while always converging quickly. In which case, multiple independent runs can then be used to reduce the failure probability to any desired positive value, while retaining efficiency (if full parallelism can be maintained).

**Theorem 31.** *With probability* $1-\delta$*, the best single run among* $O(\log(1/\delta))$ *independent runs of NPG (GNPG) converges to a globally optimal policy at an* $O(e^{-c\cdot t})$ *rate.*

As shown in Chapter 3, softmax PG can get stuck on long plateaus for even true gradient settings, which means almost sure global convergence does not necessarily imply good practical performance. Therefore, it is a reasonable choice to perform well with the compromise of small failure probability. Here we consider simply best selection, and it remains open to study whether and how practical training tricks, such as proximal update (Lazić et al., 2021;

Schulman et al., 2017) and regularization (Mnih et al., 2016), stabilize training by increasing the success probability under stochastic settings.

## 5.6 Discussions

We leave it open to identify sufficient and necessary conditions for almost sure global convergence in Section 5.4. We make the following conjecture with some intuitions.

**Conjecture 1.** *Given a stochastic policy optimization algorithm $\mathcal{A}$, if $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one sub-optimal action $a$, then $\kappa(\mathcal{A}, a^*) \in (0, 1]$ is a sufficient and necessary condition for global convergence to $\pi^*$ with **polynomial convergence rate** of $O(1/t^\alpha)$, where $\alpha > 0$.*

The necessary condition part is guaranteed by Theorem 25. For the sufficient condition part, Theorem 29 can potentially be strengthened with the claim that $\kappa(\mathcal{A}, a^*) \geq \alpha$ is a sufficient and necessary condition for global convergence with rate $O(1/t^\alpha)$ ($\alpha > 0$). The observation here is that under Assumption 3, we have $r(a^*) - r(a) \leq 1$, which leads to,

$$(\pi^* - \pi_{\theta_t})^\top r = \sum_{a \neq a^*} \pi_{\theta_t}(a) \cdot (r(a^*) - r(a)) \leq 1 - \pi_{\theta_t}(a^*). \tag{5.31}$$

This suggests that if $\kappa(\mathcal{A}, a^*) \geq \alpha$ (i.e., $1 - \pi_{\theta_t}(a^*) \in O(1/t^\alpha)$), then $(\pi^* - \pi_{\theta_t})^\top r \in O(1/t^\alpha)$. However, a gap here is $\kappa(\mathcal{A}, a^*) \geq \alpha$ means "$1 - \pi_{\theta_t}(a^*) \in O(1/t^\alpha)$ if we fix sampling $a^*$ forever", and it is not clear if this says something about "$1 - \pi_{\theta_t}(a^*) \in O(1/t^\alpha)$ if we run the algorithm $\mathcal{A}$ using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$".

### 5.6.1 Lower Bounds in Bandit Literature

In the bandit literature (Lattimore and Szepesvári, 2020), there exist $\Omega(\log T)$ and $\Omega(\sqrt{T})$ lower bounds for stochastic and adversarial reward settings respectively, which implies that the convergence speed in terms of sub-optimality ("average regret" in bandit) cannot be faster than $O(1/t)$. However, the lower bound construction there is information-theoretic, which holds for adversarial

(changing reward signal) and stochastic (the reward needs to be estimated accurately enough) settings. Theorem 30 holds for a simpler optimization setting: the reward is fixed and deterministic, but the policy gradient is estimated using on-policy sampling. Therefore, the difficulty and trade-off are from the restriction on the action-selection scheme (balancing the aggressiveness of the update and the stability), not from estimating or tracking the reward signal.

### 5.6.2 General MDPs

The one-state MDP results already show the main findings, since a large portion of this section is about constructing counterexamples showing that the stochastic policy optimization algorithms do not perform well as in the true gradient setting. A counterexample for one-state MDPs is also a counterexample for general MDPs. Therefore, there is no loss of generality by establishing negative results using one-state MDPs.

## 5.7 Summary

This chapter introduces the committal rate theory, which not only explains why faster policy optimization algorithms in the true gradient setting become dominated by slower counterparts in the on-policy stochastic setting, but also reveals an inherent geometry-convergence trade-off in stochastic policy optimization. The theory also explains empirical observations of sensitivity to random initialization for practical policy optimization algorithms as well as the effectiveness of ensemble methods.

# Chapter 6

# Conclusions and Future Directions

This dissertation introduces a new **non-uniform analysis** for non-convex optimization in reinforcement learning and machine learning. The main pillars of this analysis are two new non-uniform properties: the non-uniform Łojasiewicz inequality (NŁ) and the non-uniform smoothness (NS) property. Below I summarize the contributions and logic behind the development of the non-uniform analysis as well as some future directions.

Chapter 2 set out to study the global convergence rate of policy gradient (PG) methods with the softmax parameterization. The value function maximization problem is non-concave with respect to the parameters, and existing uniform Łojasiewicz inequalities with universal constants cannot be satisfied in this case. I therefore introduce the non-uniform Łojasiewicz inequality (NŁ) as a necessary analysis tool. Using NŁ inequalities to help analyze PG methods, I successfully resolved a number of longstanding open problems:

- By showing that policy value functions satisfy a NŁ inequality, I establish the first finite time $O(1/t)$ upper bound on the convergence rate of softmax PG methods.

- By proving that entropy regularized value functions satisfy a better NŁ inequality, I show that entropy regularized PG achieves a global linear convergence rate of $O(e^{-c \cdot t})$ (where $c > 0$).

- By proving that, without regularization, policy value functions satisfy a

reversed Łojasiewicz inequality, I establish an $\Omega(1/t)$ convergence rate lower bound for softmax PG.

- The above results provide a new understanding of entropy regularization, showing how it accelerates the convergence speed of PG optimization.
- A deeper explanation for how entropy accelerates PG convergence speed is further provided by the NŁ degree, which is a key quantity in the NŁ inequality.

This work leaves open a number of interesting questions: While some lower bounds are established, there remain gaps between the lower and upper bounds. Another interesting direction is to extend the results to alternative (e.g., restricted) policy parametrizations, or to study policy gradient when the gradient must be estimated from data. One also expects that non-uniform Łojasiewicz inequalities and the non-uniform Łojasiewicz degree could also be put to good use in other areas of non-convex optimization.

Chapter 3 focused on explaining the gap between the above theoretical results and practical performance of PG methods. In theory, PG has a $\Theta(1/t)$ convergence rate, while in practice, it exhibits extreme sensitivity to different initializations. Using the concept of NŁ coefficient, another key quantity in the NŁ inequality, I locate the source of the issue, which arises from using the softmax transform with gradient descent (ascent) methods. In particular, the logic behind the main contributions in this chapter are as follows:

- By proving that in the worst case, the progress of PG methods in each iteration is upper bounded by the NŁ coefficient, I first show that softmax PG is guaranteed to suffer from initialization sensitivity, which is the "softmax gravity well" phenomenon.
- By using an alternative escort policy transformation to improve the NŁ coefficient, I show that escort PG methods strictly improve softmax PG in terms of faster escape times from landscape plateaus.
- By showing that vanishing NŁ coefficients lead to decreasing NŁ degrees, I discover and explain why convergence degrades from $O(e^{-c \cdot t})$ to $O(1/t)$

behaviour when using the softmax transform for cross entropy minimization in supervised learning, i.e., the "softmax damping" phenomenon.

- I show that a specific choice of escort transform results in non-vanishing NŁ coefficients, and therefore preserves the NŁ degree and the fast linear convergence rate of $O(e^{-c \cdot t})$.

Uncovering these two phenomena challenges the common practice of using the softmax transformation in machine learning. However, there are other factors to consider when assessing such transformations for machine learning problems, such as the "temperature" of the softmax, or how different transforms might impact the generalization ability of 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. Since our underlying explanation using the concept of (NŁ) coefficient and its interplay with the NŁ degree matches empirical observations, we also expect the NŁ coefficient to be useful in understanding other problems in machine learning.

Chapter 4 introduced a new non-uniform smoothness (NS) property, which was inspired from the special case of escort PG and other relevant optimization research in machine learning. The combination of the NS property and NŁ inequality is extremely successful in terms of inspiring new algorithm design, being sufficiently general to cover different function classes, improving previous results beyond even classical lower bounds, and being applicable to fundamental non-convex optimization problems in machine learning. To summarize,

- I introduce a new NS property that generalizes and unifies previous special cases. The NS property inspires a new first-order method called geometry-aware normalized gradient descent (GNGD), which exploits non-uniform landscape information.

- I propose a non-uniform analysis of gradient descent (GD) and GNGD when the NS and NŁ properties are satisfied. GNGD overcomes the

classical $\Omega(1/t^2)$ lower bound in convex-smooth analysis by exploiting the additional non-uniform properties.

- By showing that value functions satisfy the NS property, and that the NS coefficient is the PG norm in this case, I prove that geometry-aware normalized PG (GNPG) achieves a global linear convergence rate of $O(e^{-c \cdot t})$ without using regularization or introducing an arg max update.

- By showing that the mean squared error (MSE) in generalized linear model training (GLM) satisfies the NŁ and NS properties, I prove that both GD and GNGD achieve $O(e^{-c \cdot t})$ convergence in this case, significantly improving previous results that establish only $O(1/\sqrt{t})$ convergence rates. Combining NS and NŁ also provides a better understanding of how to use geometry-aware normalization.

This general characterization and analysis based on non-uniform properties applies to important non-convex optimization problems in machine learning. One future direction is to further push the analysis to other domains with more complex function approximators, such as neural networks (Allen-Zhu et al., 2019). Another valuable direction for future work is to incorporate stochastic gradient (Karimi et al., 2016) and other adaptive gradient-based methods (Kingma and Ba, 2014) in the analysis. Finally, it would be interesting to apply other non-uniform properties beyond those mentioned in this chapter.

Chapter 5 finally extends the analysis to stochastic policy optimization. The novel findings inspire the introduction of a new committal rate theory:

- By investigating several policy optimization algorithms in the stochastic setting, I uncover an anomaly that the preferability of policy optimization algorithms changes dramatically depending on whether true versus on-policy stochastic gradients are used.

- I introduce the concept of committal rate to characterize the interplay between convergence rates and almost sure global convergence, which is then used to explain the anomaly introduced above.

- In particular, I use the committal rate to reveal an inherent geometry-convergence trade-off: an uninformed algorithm can either converge to a globally optimal policy with probability 1 but at a rate no faster than $O(1/t)$, or it can achieve faster than $O(1/t)$ convergence but necessarily with a positive probability of diverging or converging to a sub-optimal policy.

- I use this committal rate theory to explain why practical policy optimization algorithms are sensitive to random initialization. From these results, I then develop an ensemble method that can achieve fast convergence to global optima with high probability, allowing a positive but controllable probability of failure.

An interesting direction for future study is to investigate the necessary and sufficient conditions for almost sure global convergence, which could be weaker than the bounded variance assumption. Another important direction is to investigate whether other techniques might be integrated into on-policy stochastic optimization to break the condition of Theorem 30 and bypass the geometry-convergence trade-off, to achieve both almost sure global convergence and a faster than $O(1/t)$ rate. One also expects that some generalized versions of committal rate would be meaningful in stochastic reward settings.

# Bibliography


Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., & Weisz, G. (2019). Politex: Regret bounds for policy iteration using expert prediction. *International Conference on Machine Learning*, 3692–3702.
(Cit. on pp. 89, 95, 97.)

Agarwal, A., Kakade, S. M., Lee, J. D., & Mahajan, G. (2019). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*.
(Cit. on pp. 1, 9, 10, 14, 20–22, 88, 92, 126, 136, 141, 156.)

Ahmed, Z., Le Roux, N., Norouzi, M., & Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. *International Conference on Machine Learning*, 151–160.
(Cit. on pp. 10, 11.)

Allen-Zhu, Z., Li, Y., & Song, Z. (2019). A convergence theory for deep learning via over-parameterization. *International Conference on Machine Learning*, 242–252.
(Cit. on pp. 1, 61, 62, 112.)

Andrychowicz, O. M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A. et al. (2020). Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, *39*(1), 3–20.
(Cit. on p. 98.)

Auer, P., Herbster, M., & Warmuth, M. K. (1996). Exponentially many local minima for single neurons. *Advances in neural information processing systems*, 316–322.
(Cit. on p. 60.)

Bárta, T. (2017). Rate of convergence to equilibrium and Łojasiewicz-type estimates. *Journal of Dynamics and Differential Equations*, *29*(4), 1553–1568.
(Cit. on p. 31.)

Beck, C., & Schögl, F. (1995). *Thermodynamics of chaotic systems: An introduction*. Cambridge University Press.
(Cit. on p. 44.)

Benveniste, A., Métivier, M., & Priouret, P. (2012). *Adaptive algorithms and stochastic approximations* (Vol. 22). Springer Science & Business Me-

dia.
(Cit. on p. 96.)

Bhandari, J., & Russo, D. (2019). Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*.
(Cit. on pp. 1, 9.)

Bhandari, J., & Russo, D. (2020). A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*.
(Cit. on p. 79.)

Boyd, S., Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.
(Cit. on p. 1.)

Bridle, J. S. (1989). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. *Advances in neural information processing systems*.
(Cit. on p. 38.)

Bubeck, S. et al. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, *8*(3-4), 231–357.
(Cit. on p. 71.)

Cen, S., Cheng, C., Chen, Y., Wei, Y., & Chi, Y. (2020). Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*.
(Cit. on pp. 79, 87, 88.)

Chen, M., Gummadi, R., Harris, C., & Schuurmans, D. (2019). Surrogate objectives for batch policy optimization in one-step decision making. *Advances in Neural Information Processing Systems*, 8825–8835.
(Cit. on pp. 43, 58.)

Chung, W., Thomas, V., Machado, M. C., & Roux, N. L. (2020). Beyond variance reduction: Understanding the true impact of baselines on policy optimization. *arXiv preprint arXiv:2008.13773*.
(Cit. on pp. 89, 91, 94–98, 101, 329, 358.)

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., & Talwalkar, A. (2021). Gradient descent on neural networks typically occurs at the edge of stability. *International Conference on Learning Representations*.
(Cit. on pp. 63, 64.)

de Brébisson, A., & Vincent, P. (2015). An exploration of softmax alternatives belonging to the spherical loss family. *International Conference on Learning Representations*.
(Cit. on p. 38.)

Denisov, D., & Walton, N. (2020). Regret analysis of a markov policy gradient algorithm for multi-arm bandits. *arXiv preprint arXiv:2007.10229*.
(Cit. on pp. 91, 350.)

Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.
(Cit. on p. 1.)

Fazel, M., Ge, R., Kakade, S., & Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. *International Conference on Machine Learning*, 1467–1476.
(Cit. on p. 9.)

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.
(Cit. on p. 37.)

Ge, R., Lee, J. D., & Ma, T. (2016). Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*.
(Cit. on p. 1.)

Golub, G. H. (1973). Some modified matrix eigenvalue problems. *SIAM Review*, *15*(2), 318–334.
(Cit. on p. 196.)

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning*, 1861–1870.
(Cit. on pp. 23, 29, 201.)

Hazan, E., Kakade, S., Singh, K., & Van Soest, A. (2019). Provably efficient maximum entropy exploration. *International Conference on Machine Learning*, 2681–2691.
(Cit. on p. 73.)

Hazan, E., Levy, K., & Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, *28*, 1594–1602.
(Cit. on pp. 61, 62, 64, 80–82, 85, 86.)

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).
(Cit. on p. 105.)

Hennes, D., Morrill, D., Omidshafiei, S., Munos, R., Perolat, J., Lanctot, M., Gruslys, A., Lespiau, J.-B., Parmas, P., Duenez-Guzman, E. et al. (2019). Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*.
(Cit. on p. 44.)

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
(Cit. on p. 54.)

Jung, W., Park, G., & Sung, Y. (2020). Population-guided parallel policy search for reinforcement learning. *arXiv preprint arXiv:2001.02907*.
(Cit. on p. 106.)

Kakade, S., & Langford, J. (2002). Approximately optimal approximate reinforcement learning. *ICML*, *2*, 267–274.
(Cit. on p. 194.)

Kakade, S. M. (2002). A natural policy gradient. *Advances in neural information processing systems*, 1531–1538.
(Cit. on pp. 14, 76, 92.)

Karimi, H., Nutini, J., & Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 795–811.
(Cit. on p. 112.)

Kawaguchi, K. (2016). Deep learning without poor local minima. *arXiv preprint arXiv:1605.07110*.
(Cit. on p. 1.)

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., & Maguluri, S. T. (2021). On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*.
(Cit. on pp. 10, 88, 92.)

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
(Cit. on pp. 56, 112.)

Kivinen, J., & Warmuth, M. K. (1998). Relative loss bounds for multidimensional regression problems. *Advances in neural information processing systems*, 287–293.
(Cit. on p. 60.)

Knopp, K. (1947). *Theory and application of infinite series*. Hafner Publishing Company, New York.
(Cit. on p. 360.)

Kurdyka, K. (1998). On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, *48*(3), 769–783.
(Cit. on pp. 62, 66.)

Laha, A., Chemmengath, S. A., Agrawal, P., Khapra, M. M., Sankaranarayanan, K., & Ramaswamy, H. G. (2018). On controllable sparse alternatives to softmax. *Advances in Neural Information Processing Systems*, 6423–6433.
(Cit. on p. 38.)

Lan, G. (2021). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*.
(Cit. on pp. 88, 89, 97.)

Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.
(Cit. on p. 107.)

Lazić, N., Hao, B., Abbasi-Yadkori, Y., Schuurmans, D., & Szepesvári, C. (2021). Optimization issues in kl-constrained approximate policy iteration. *arXiv preprint arXiv:2102.06234*.
(Cit. on p. 106.)

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.
(Cit. on p. 1.)

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). A tutorial on energy based learning. *Predicting structured data*. MIT Press.
(Cit. on p. 38.)

Li, G., Wei, Y., Chi, Y., Gu, Y., & Chen, Y. (2021). Softmax policy gradient methods can take exponential time to converge. *arXiv preprint arXiv:2102.11270*.
(Cit. on pp. 78, 92.)

Łojasiewicz, S. (1963). Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, *117*, 87–89.
(Cit. on pp. 16, 62, 66.)

Mei, J., Dai, B., Xiao, C., Szepesvari, C., & Schuurmans, D. (2021a). Understanding the effect of stochasticity in policy optimization. *arXiv preprint arXiv:*
(cit. on pp. 7, 87.)

Mei, J., Gao, Y., Dai, B., Szepesvari, C., & Schuurmans, D. (2021b). Leveraging non-uniformity in first-order non-convex optimization. *arXiv preprint arXiv:2105.06072*.
(Cit. on pp. 4, 6, 7, 61, 125.)

Mei, J., Xiao, C., Dai, B., Li, L., Szepesvári, C., & Schuurmans, D. (2020a). Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, *33*.
(Cit. on pp. 5, 7, 37, 124.)

Mei, J., Xiao, C., Huang, R., Schuurmans, D., & Müller, M. (2019). On principled entropy exploration in policy optimization. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3130–3136.
(Cit. on pp. 7, 23, 29.)

Mei, J., Xiao, C., Szepesvari, C., & Schuurmans, D. (2020b). On the global convergence rates of softmax policy gradient methods. *International Conference on Machine Learning*, 6820–6829.
(Cit. on pp. 5, 7, 8, 88, 123, 124, 289.)

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *International conference on machine learning*, 1928–1937.
(Cit. on pp. 23, 29, 107.)

Nachum, O., Norouzi, M., Xu, K., & Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. *Advances in Neural Information Processing Systems*, 2775–2785.
(Cit. on pp. 23, 26, 29, 179, 199, 201–203.)

Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, *19*(4), 1574–1609.
(Cit. on pp. 89, 96, 97.)

Nemirovski, A. S., & Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.
(Cit. on p. 71.)

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course* (Vol. 87). Springer Science & Business Media.
(Cit. on p. 71.)

Nesterov, Y. (2018). *Lectures on convex optimization* (Vol. 137). Springer.
(Cit. on pp. 1, 23, 51.)

Nesterov, Y., & Polyak, B. T. (2006). Cubic regularization of Newton method and its global performance. *Mathematical Programming*, *108*(1), 177–205.
(Cit. on p. 31.)

Neu, G., Jonsson, A., & Gómez, V. (2017). A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*.
(Cit. on p. 23.)

Norouzi, M., Bengio, S., Jaitly, N., Schuster, M., Wu, Y., Schuurmans, D. et al. (2016). Reward augmented maximum likelihood for neural structured prediction. *Advances In Neural Information Processing Systems*, *29*, 1723–1731.
(Cit. on p. 54.)

O'Donoghue, B., Osband, I., & Ionescu, C. (2020). Making sense of reinforcement learning and probabilistic inference. *arXiv preprint arXiv:2001.00805*.
(Cit. on p. 29.)

Parker-Holder, J., Pacchiano, A., Choromanski, K., & Roberts, S. (2020). Effective diversity in population-based reinforcement learning. *arXiv preprint arXiv:2002.00632*.
(Cit. on p. 106.)

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, *3*(4), 643–653.
(Cit. on pp. 62, 66.)

Ren, T., Li, J., Dai, B., Du, S. S., & Sanghavi, S. (2021). Nearly horizon-free offline reinforcement learning. *arXiv preprint arXiv:2103.14077*.
(Cit. on p. 98.)

Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.
(Cit. on p. 1.)

Schaul, T., Borsa, D., Modayil, J., & Pascanu, R. (2019). Ray interference: A source of plateaus in deep reinforcement learning. *arXiv preprint arXiv:1904.11455*.
(Cit. on p. 44.)

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. *International conference on machine learning*, 1889–1897.
(Cit. on pp. 8, 14, 40, 92, 98, 106.)

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
(Cit. on pp. 8, 92, 97, 106, 107.)

Shani, L., Efroni, Y., & Mannor, S. (2020). Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 5668–5675.
(Cit. on pp. 9, 10.)

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., & Riedmiller, M. (2014). Deterministic policy gradient algorithms. *International Conference on Machine Learning*, 387–395.
(Cit. on p. 14.)

Sun, R.-Y. (2020). Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, *8*(2), 249–294.
(Cit. on p. 1.)

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (second). MIT Press.
(Cit. on pp. 1, 8, 37.)

Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 1057–1063.
(Cit. on pp. 8, 14.)

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.
(Cit. on p. 54.)

Tsallis, C., Mendes, R. S., & Plastino, A. R. (1998). The role of constraints within generalized nonextensive statistics. *Physica A: Statistical Mechanics and its Applications*, *261*(3-4), 534–554.
(Cit. on p. 44.)

Tucker, G., Bhupatiraju, S., Gu, S., Turner, R., Ghahramani, Z., & Levine, S. (2018). The mirage of action-dependent baselines in reinforcement learning. *International Conference on Machine Learning*, 5015–5024.
(Cit. on p. 104.)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.
(Cit. on p. 37.)

Walton, N. (2020). A short note on soft-max and policy gradients in bandits problems. *arXiv preprint arXiv:2007.10297*.
(Cit. on p. 17.)

Wiering, M. A., & Van Hasselt, H. (2008). Ensemble algorithms in reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics,*

*Part B (Cybernetics)*, *38*(4), 930–936.
(Cit. on p. 106.)

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3-4), 229–256.
(Cit. on p. 14.)

Williams, R. J., & Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, *3*(3), 241–268.
(Cit. on pp. 23, 29.)

Wilson, A., Mackey, L., & Wibisono, A. (2019). Accelerating rescaled gradient descent: Fast optimization of smooth functions. *arXiv preprint arXiv:1902.08825*.
(Cit. on pp. 63–65, 69, 71.)

Xiao, C., Huang, R., Mei, J., Schuurmans, D., & Müller, M. (2019). Maximum entropy monte-carlo planning. *Advances in Neural Information Processing Systems*, *32*, 9520–9528.
(Cit. on pp. 23, 198.)

Xiao, C., Mei, J., & Müller, M. (2018). Memory-augmented monte carlo tree search. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).
(Cit. on p. 23.)

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.
(Cit. on p. 1.)

Zhang, J., He, T., Sra, S., & Jadbabaie, A. (2019). Why gradient clipping accelerates training: A theoretical justification for adaptivity. *International Conference on Learning Representations*.
(Cit. on pp. 63–65, 69, 72.)

Zhang, J., Ni, C., Yu, Z., Szepesvari, C., & Wang, M. (2021). On the convergence and sample efficiency of variance-reduced policy gradient method. *arXiv preprint arXiv:2102.08607*.
(Cit. on p. 97.)

Zhang, J., Kim, J., O'Donoghue, B., & Boyd, S. (2020a). Sample efficient reinforcement learning with reinforce. *arXiv preprint arXiv:2010.11364*.
(Cit. on p. 97.)

Zhang, K., Koppel, A., Zhu, H., & Basar, T. (2020b). Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, *58*(6), 3586–3612.
(Cit. on p. 97.)

Zhou, Y., Wang, Z., & Liang, Y. (2018). Convergence of cubic regularization for nonconvex optimization under KL property. *Advances in Neural Information Processing Systems*, 3760–3769.
(Cit. on p. 31.)

Ziebart, B. D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy.
(Cit. on p. 23.)

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. *Aaai*, *8*, 1433–1438.
(Cit. on p. 23.)

Zou, D., Cao, Y., Zhou, D., & Gu, Q. (2020). Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, *109*(3), 467–492.
(Cit. on p. 1.)

# Appendix A

# Non-convex (Non-concave) Examples for NŁ Inequality

We list some non-convex (or non-concave in maximization problems) functions which satisfy NŁ inequalities here from literature. See corresponding references for details.

**Expected reward, softmax parameterization.** As shown in Lemma 3 and Mei et al. (2020b, Lemma 3),

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. \tag{A.1}$$

**Value function, softmax parameterization.** As shown in Lemma 8 and Mei et al. (2020b, Lemma 8),

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

**Entropy regularized expected reward, softmax parameterization.** As shown in Proposition 5 and Mei et al. (2020b, Proposition 5),

$$\left\| \frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} \right\|_2 \geq C(\theta) \cdot \left[ \pi_\tau^{*\top} (r - \tau \log \pi_\tau^*) - \pi_\theta^\top (r - \tau \log \pi_\theta) \right]^{\frac{1}{2}}, \tag{A.3}$$

where

$$C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a). \tag{A.4}$$

**Entropy regularized value function, softmax parameterization.** As shown in Lemma 15 and Mei et al. (2020b, Lemma 15),

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq C(\theta) \cdot \left[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}, \tag{A.5}$$

where

$$C(\theta) = \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}}. \tag{A.6}$$

**Expected reward, escort parameterization.** As shown in Lemma 19 and Mei et al. (2020a, Lemma 3),

$$\left\| \frac{d \pi_\theta^\top r}{d\theta} \right\|_2 \geq \frac{p}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r. \tag{A.7}$$

**Value function, escort parameterization.** As shown in Lemma 44 and Mei et al. (2020a, Lemma 7),

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq C(\theta, p) \cdot \left[ V^*(\rho) - V^{\pi_\theta}(\rho) \right], \tag{A.8}$$

where

$$C(\theta, p) = \frac{p}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \frac{\min_s \pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s \|\theta(s, \cdot)\|_p}. \tag{A.9}$$

**Entropy regularized value function, escort parameterization.** As shown in Lemma 48 and Mei et al. (2020a, Lemma 12),

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq C(\theta, p) \cdot \left[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}, \tag{A.10}$$

where

$$C(\theta, p) = \frac{p \cdot \sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \frac{\min_{s,a} \pi_\theta(a|s)^{1-1/p}}{\max_s \|\theta(s, \cdot)\|_p} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}}. \tag{A.11}$$

**Cross entropy, escort parameterization.** As shown in Lemma 50 and Mei et al. (2020a, Lemma 17),

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

**Generalized linear models, sigmoid activation, mean squared error.** As shown in Lemma 27 and Mei et al. (2021b, Lemma 9),

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \geq 8 \cdot u(\theta) \cdot \min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi} \cdot \left[ \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}. \quad \text{(A.13)}$$

# Appendix B

# Proofs for Chapter 2: Global Convergence Rates of Softmax Policy Gradient

## B.1 Proofs for Section 2.3: Softmax Parametrization

### B.1.1 Preliminaries

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

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

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

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s'\sim d_\mu^{\pi_\theta}} \left[ \sum_a \frac{\partial \pi_\theta(a|s')}{\partial \theta} \cdot Q^{\pi_\theta}(s',a) \right]. \tag{B.2}$$

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[ \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right] \tag{B.3}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left( \frac{d\pi(\cdot|s)}{d\theta(s,\cdot)} \right)^\top Q^{\pi_\theta}(s,\cdot) \tag{B.4}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot). \qquad \text{(using Eq. (2.8))} \tag{B.5}$$

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[ Q^{\pi_\theta}(s,a) - \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right] \tag{B.6}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot [Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)] \tag{B.7}$$

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

where the second equation is using

$$V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a). \qquad \square$$

### B.1.2 Proofs for Softmax Parametrization in Bandits

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

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

$$\frac{1}{2} \cdot \left(\pi_{\theta_1}^\top r + \pi_{\theta_2}^\top r\right) = \frac{1}{2} \cdot \left(\frac{2}{3} + \frac{259}{500}\right) = \frac{1777}{3000} = \frac{14216}{24000}. \tag{B.9}$$

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

$$\pi_{\bar{\theta}}^\top r = \frac{71}{120} = \frac{14200}{24000}. \tag{B.10}$$

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

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{5}{4} \cdot \|\theta' - \theta\|_2^2. \tag{B.11}$$

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

$$S = \frac{d}{d\theta}\left\{\frac{d\pi_\theta^\top r}{d\theta}\right\} \tag{B.12}$$

$$= \frac{d}{d\theta}\left\{H(\pi_\theta)r\right\} \qquad \text{(using Eq. (2.8))} \tag{B.13}$$

$$= \frac{d}{d\theta}\left\{(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top)r\right\}. \tag{B.14}$$

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

$$S_{i,j} = \frac{d\{\pi_\theta(i)\cdot(r(i)-\pi_\theta^\top r)\}}{d\theta(j)} \tag{B.15}$$

$$= \frac{d\pi_\theta(i)}{d\theta(j)}\cdot(r(i)-\pi_\theta^\top r) + \pi_\theta(i)\cdot\frac{d\{r(i)-\pi_\theta^\top r\}}{d\theta(j)} \tag{B.16}$$

$$= (\delta_{ij}\pi_\theta(j) - \pi_\theta(i)\pi_\theta(j))\cdot(r(i)-\pi_\theta^\top r) \tag{B.17}$$

$$\quad - \pi_\theta(i)\cdot(\pi_\theta(j)r(j) - \pi_\theta(j)\pi_\theta^\top r) \tag{B.18}$$

$$= \delta_{ij}\pi_\theta(j)\cdot(r(i)-\pi_\theta^\top r) \tag{B.19}$$

$$\quad - \pi_\theta(i)\pi_\theta(j)\cdot(r(i)-\pi_\theta^\top r) \tag{B.20}$$

$$\quad - \pi_\theta(i)\pi_\theta(j)\cdot(r(j)-\pi_\theta^\top r), \tag{B.21}$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i=j, \\ 0, & \text{otherwise} \end{cases} \tag{B.22}$$

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

$$\left|y^\top S y\right| = \left|\sum_{i=1}^K\sum_{j=1}^K S_{i,j}y(i)y(j)\right| \tag{B.23}$$

$$= \Bigg|\sum_i \pi_\theta(i)(r(i)-\pi_\theta^\top r)y(i)^2 \tag{B.24}$$

$$\quad - 2\sum_i \pi_\theta(i)(r(i)-\pi_\theta^\top r)y(i)\sum_j \pi_\theta(j)y(j)\Bigg| \tag{B.25}$$

$$= \left|\left(H(\pi_\theta)r\right)^\top (y\odot y) - 2\cdot\left(H(\pi_\theta)r\right)^\top y\cdot\left(\pi_\theta^\top y\right)\right| \tag{B.26}$$

$$\le \left\|H(\pi_\theta)r\right\|_\infty\cdot\left\|y\odot y\right\|_1 + 2\cdot\left\|H(\pi_\theta)r\right\|_1\cdot\left\|y\right\|_\infty\cdot\left\|\pi_\theta\right\|_1\cdot\left\|y\right\|_\infty, \tag{B.27}$$

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

$$\begin{aligned}
\|H_{i,:}(\pi_\theta)\|_1 &= \pi_\theta(i) - \pi_\theta(i)^2 + \pi_\theta(i) \cdot \sum_{j \neq i} \pi_\theta(j) && \text{(B.28)}\\
&= \pi_\theta(i) - \pi_\theta(i)^2 + \pi_\theta(i) \cdot (1 - \pi_\theta(i)) && \text{(B.29)}\\
&= 2 \cdot \pi_\theta(i) \cdot (1 - \pi_\theta(i)) && \text{(B.30)}\\
&\le 1/2. \qquad \text{(using that } x \cdot (1-x) \le 1/4 \text{ holds for } x \in [0,1]) && \text{(B.31)}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|H(\pi_\theta) r\|_1 &= \sum_i \pi_\theta(i) \cdot \left| r(i) - \pi_\theta^\top r \right| && \text{(B.32)}\\
&\le \max_i \left| r(i) - \pi_\theta^\top r \right| && \text{(B.33)}\\
&\le 1. \qquad \left(\text{using } r \in [0,1]^K\right) && \text{(B.34)}
\end{aligned}$$

Therefore we have,

$$\begin{aligned}
\left| y^\top S(r,\theta) y \right| &\le \|H(\pi_\theta) r\|_\infty \cdot \|y\|_2^2 + 2 \cdot \|H(\pi_\theta) r\|_1 \cdot \|y\|_2^2 && \text{(B.35)}\\
&= \max_i \left| \left(H_{i,:}(\pi_\theta)\right)^\top r \right| \cdot \|y\|_2^2 + 2 \cdot \|H(\pi_\theta) r\|_1 \cdot \|y\|_2^2 && \text{(B.36)}\\
&\le \max_i \|H_{i,:}(\pi_\theta)\|_1 \cdot \|r\|_\infty \cdot \|y\|_2^2 + 2 \cdot 1 \cdot \|y\|_2^2 && \text{(B.37)}\\
&\le (1/2 + 2) \cdot \|y\|_2^2 = 5/2 \cdot \|y\|_2^2, && \text{(B.38)}
\end{aligned}$$

finishing the proof. □

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

$$\left\| \frac{d \pi_\theta^\top r}{d\theta} \right\|_2 \ge \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r \,. \tag{B.39}$$

When there are multiple optimal actions, we have

$$\left\| \frac{d \pi_\theta^\top r}{d\theta} \right\|_2 \ge \frac{1}{\sqrt{|\mathcal{A}^*|}} \cdot \left[ \sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*) \right] \cdot (\pi^* - \pi_\theta)^\top r, \tag{B.40}$$

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

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \left( \sum_{a^* \in \mathcal{A}^*} \left[ \pi_\theta(a^*) \cdot (r(a^*) - \pi_\theta^\top r) \right]^2 \right)^{\frac{1}{2}} \tag{B.41}$$

$$\geq \frac{1}{\sqrt{|\mathcal{A}^*|}} \sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*) \cdot (r(a^*) - \pi_\theta^\top r) \qquad \text{(by Cauchy-Schwarz)} \tag{B.42}$$

$$= \frac{1}{\sqrt{|\mathcal{A}^*|}} \cdot \left[ \sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*) \right] \cdot (\pi^* - \pi_\theta)^\top r. \qquad \square$$

For the remaining results in this section, for simplicity, we assume that $\mathcal{A}^* = \{a^*\}$, i.e., there is a unique optimal action $a^*$.

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(t \cdot c_t^2), \qquad \text{and} \tag{B.43}$$

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \leq \min \left\{ \sqrt{5T}/c_T,\ (5 \log T)/c_T^2 + 1 \right\}. \tag{B.44}$$

*Proof.* According to Lemma 2,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \leq \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{B.45}$$

which implies

$$\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r \leq -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{B.46}$$

$$= -\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{5}{4} \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \text{(using Update 1)} \tag{B.47}$$

$$= -\frac{1}{5} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \text{(using } \eta = 2/5\text{)} \tag{B.48}$$

$$\leq -\frac{1}{5} \cdot \left[ \pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \text{(by Lemma 3)} \tag{B.49}$$

$$\leq -\frac{c_t^2}{5} \cdot \left[ (\pi^* - \pi_{\theta_t})^\top r \right]^2, \qquad \text{(by the definition of } c_t\text{)} \tag{B.50}$$

which is equivalent to

$$(\pi^* - \pi_{\theta_{t+1}})^\top r - (\pi^* - \pi_{\theta_t})^\top r \leq -\frac{c_t^2}{5} \cdot \left[ (\pi^* - \pi_{\theta_t})^\top r \right]^2. \tag{B.51}$$

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

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

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

$$\begin{aligned}
\delta_{t+1} &\le f_t(\delta_t) && \text{(by Eq. (B.51))} && \text{(B.52)}\\
&\le f_t\left(\frac{5}{c_t^2} \cdot \frac{1}{t}\right) && \left(\text{using } \delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t} \le \frac{5}{2 \cdot c_t^2},\ t \ge 2\right) && \text{(B.53)}\\
&= \frac{5}{c_t^2} \cdot \left(\frac{1}{t} - \frac{1}{t^2}\right) && && \text{(B.54)}\\
&\le \frac{5}{c_t^2} \cdot \frac{1}{t+1} && && \text{(B.55)}\\
&\le \frac{5}{c_{t+1}^2} \cdot \frac{1}{t+1}, && \text{(using } c_t \ge c_{t+1} > 0) && \text{(B.56)}
\end{aligned}$$

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

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

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \frac{5 \log T}{c_T^2} + 1. \tag{B.57}$$

On the other hand, rearranging Eq. (B.51) and summing up

$$\delta_t^2 \le \frac{5}{c_t^2} \cdot (\delta_t - \delta_{t+1}) \le \frac{5}{c_T^2} \cdot (\delta_t - \delta_{t+1}), \tag{B.58}$$

from $t = 1$ to $T$,

$$\begin{aligned}
\sum_{t=1}^{T} \delta_t^2 &\le \frac{5}{c_T^2} \sum_{t=1}^{T} (\delta_t - \delta_{t+1}) && \text{(B.59)}\\
&= \frac{5}{c_T^2} \cdot (\delta_1 - \delta_{T+1}) && \text{(B.60)}\\
&\le \frac{5}{c_T^2}. \qquad (\text{since } \delta_{T+1} \ge 0,\ \delta_1 \le 1) && \text{(B.61)}
\end{aligned}$$

Therefore, by Cauchy-Schwarz,

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

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

*Proof.* Let

$$c = \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right) \tag{B.62}$$

and

$$\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0 \tag{B.63}$$

denote the reward gap of $r$. We will prove that $\inf_{t\geq 1} \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(a^*)$, where $t_0 = \min\{t : \pi_{\theta_t}(a^*) \geq \frac{c}{c+1}\}$. Note that $t_0$ depends only on $\theta_1$ and $c$, and $c$ depends only on the problem. Define the following regions,

$$\mathcal{R}_1 = \left\{\theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)},\ \forall a \neq a^*\right\}, \tag{B.64}$$

$$\mathcal{R}_2 = \left\{\theta : \pi_\theta(a^*) \geq \pi_\theta(a),\ \forall a \neq a^*\right\}, \tag{B.65}$$

$$\mathcal{N}_c = \left\{\theta : \pi_\theta(a^*) \geq \frac{c}{c+1}\right\}. \tag{B.66}$$

We make the following three-part claim.

**Claim 1.** *The following hold:*

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

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

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

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

$$\mathcal{R}_1(a) = \left\{\theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)}\right\}. \tag{B.67}$$

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

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

$$\theta_{t+1}(a^*) = \theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \tag{B.68}$$

$$\geq \theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \tag{B.69}$$

$$= \theta_{t+1}(a), \tag{B.70}$$

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

$$\pi_{\theta_{t+1}}(a^*) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) \geq \pi_{\theta_{t+1}}(a) \cdot \left(r(a) - \pi_{\theta_{t+1}}^\top r\right), \tag{B.71}$$

which is equivalent to $\frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a^*)} \geq \frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a)}$, i.e., $\theta_{t+1} \in \mathcal{R}_1(a)$.
Case (b): Suppose now that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$. First note that for any $\theta$ and $a \neq a^*$, $\theta \in \mathcal{R}_1(a)$ holds if and only if

$$r(a^*) - r(a) \geq \left(1 - \frac{\pi_\theta(a^*)}{\pi_\theta(a)}\right) \cdot \left(r(a^*) - \pi_\theta^\top r\right). \tag{B.72}$$

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

$$\pi_\theta(a^*) \cdot \left(r(a^*) - \pi_\theta^\top r\right) \geq \pi_\theta(a) \cdot \left(r(a) - \pi_\theta^\top r\right) \tag{B.73}$$

$$= \pi_\theta(a) \cdot \left(r(a^*) - \pi_\theta^\top r\right) - \pi_\theta(a) \cdot \left(r(a^*) - r(a)\right), \tag{B.74}$$

which, after rearranging, is equivalent to Eq. (B.72). Hence, it suffices to show that Eq. (B.72) holds for $\theta_{t+1}$ provided it holds for $\theta_t$.

From the latter condition, we get

$$r(a^*) - r(a) \geq \left(1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\}\right) \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right). \tag{B.75}$$

After an update of the parameters, according to the ascent lemma for smooth function (Lemma 33), $\pi_{\theta_{t+1}}^\top r \geq \pi_{\theta_t}^\top r$, i.e.,

$$0 < r(a^*) - \pi_{\theta_{t+1}}^\top r \leq r(a^*) - \pi_{\theta_t}^\top r\,. \tag{B.76}$$

On the other hand,

$$\theta_{t+1}(a^*) - \theta_{t+1}(a) = \theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} - \theta_t(a) - \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \tag{B.77}$$

$$\geq \theta_t(a^*) - \theta_t(a), \tag{B.78}$$

which implies that

$$1 - \exp\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\} \leq 1 - \exp\{\theta_t(a^*) - \theta_t(a)\}. \tag{B.79}$$

Furthermore, by our assumption that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$, we have

$$1 - \exp\{\theta_t(a^*) - \theta_t(a)\} = 1 - \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} > 0. \tag{B.80}$$

Putting things together, we get

$$\begin{aligned}
&(1 - \exp\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\}) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) && \text{(B.81)}\\
&\quad \leq (1 - \exp\{\theta_t(a^*) - \theta_t(a)\}) \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right) && \text{(B.82)}\\
&\quad \leq r(a^*) - r(a), && \text{(B.83)}
\end{aligned}$$

which is equivalent to

$$\left(1 - \frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)}\right) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) \leq r(a^*) - r(a), \tag{B.84}$$

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

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

$$\begin{aligned}
\pi_{\theta_{t+1}}(a^*) &= \frac{\exp\{\theta_{t+1}(a^*)\}}{\sum_a \exp\{\theta_{t+1}(a)\}} && \text{(B.85)}\\
&= \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right\}} && \text{(B.86)}\\
&\geq \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}} \qquad \left(\text{using } \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \geq \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right) && \text{(B.87)}\\
&= \frac{\exp\{\theta_t(a^*)\}}{\sum_a \exp\{\theta_t(a)\}} = \pi_{\theta_t}(a^*). && \text{(B.88)}
\end{aligned}$$

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

$$\frac{d\pi_\theta^\top r}{d\theta(a^*)} = \pi_\theta(a^*) \cdot \big(r(a^*) - \pi_\theta^\top r\big) \tag{B.89}$$

$$> \pi_\theta(a) \cdot \big(r(a) - \pi_\theta^\top r\big) \tag{B.90}$$

$$\big(\text{using } r(a^*) - \pi_\theta^\top r > 0 \text{ and } r(a^*) > r(a)\big) \tag{B.91}$$

$$= \frac{d\pi_\theta^\top r}{d\theta(a)}. \tag{B.92}$$

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

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

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

$$\frac{d\pi_\theta^\top r}{d\theta(a^*)} - \frac{d\pi_\theta^\top r}{d\theta(a)} = \frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} \tag{B.93}$$

$$= \pi_\theta(1) \cdot \big(r(1) - \pi_\theta^\top r\big) - \pi_\theta(2) \cdot \big(r(2) - \pi_\theta^\top r\big) \tag{B.94}$$

$$= 2 \cdot \pi_\theta(1) \cdot \big(r(1) - \pi_\theta^\top r\big) + \sum_{i=3}^{K} \pi_\theta(i) \cdot \big(r(i) - \pi_\theta^\top r\big) \tag{B.95}$$

$$\text{(see below)} \tag{B.96}$$

$$= \left(2 \cdot \pi_\theta(1) + \sum_{i=3}^{K} \pi_\theta(i)\right) \cdot \big(r(1) - \pi_\theta^\top r\big) \tag{B.97}$$

$$- \sum_{i=3}^{K} \pi_\theta(i) \cdot (r(1) - r(i)) \tag{B.98}$$

$$\geq \left(2 \cdot \pi_\theta(1) + \sum_{i=3}^{K} \pi_\theta(i)\right) \cdot \big(r(1) - \pi_\theta^\top r\big) - \sum_{i=3}^{K} \pi_\theta(i) \tag{B.99}$$

$$\geq \left(2 \cdot \pi_\theta(1) + \sum_{i=3}^{K} \pi_\theta(i)\right) \cdot \frac{\Delta}{K} - \sum_{i=3}^{K} \pi_\theta(i), \tag{B.100}$$

where the second equation is because

$$\pi_\theta(2) \cdot \big(r(2) - \pi_\theta^\top r\big) + \sum_{i \neq 2} \pi_\theta(i) \cdot \big(r(i) - \pi_\theta^\top r\big) = 0, \tag{B.101}$$

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

$$r(1) - \pi_\theta^\top r = \sum_{i=1}^{K} \pi_\theta(i) \cdot r(1) - \sum_{i=1}^{K} \pi_\theta(i) \cdot r(i) \tag{B.102}$$

$$= \sum_{i=2}^{K} \pi_\theta(i) \cdot (r(1) - r(i)) \tag{B.103}$$

$$\geq \sum_{i=2}^{K} \pi_\theta(i) \cdot \Delta \geq \max_{a \neq a^*} \{\pi_\theta(a)\} \cdot \Delta \tag{B.104}$$

$$\geq \frac{\Delta}{K}. \qquad \left(\text{using } \max_{a \neq a^*} \{\pi_\theta(a)\} = \max_{a} \{\pi_\theta(a)\} \geq \frac{1}{K}\right) \tag{B.105}$$

Plugging $\sum_{i=3}^{K} \pi_\theta(i) = 1 - \pi_\theta(1) - \pi_\theta(2)$ into Eq. (B.93) and rearranging the resulting expression we get

$$\frac{d\pi_\theta^\top r}{d\theta(a^*)} - \frac{d\pi_\theta^\top r}{d\theta(a)} \tag{B.106}$$

$$\geq \pi_\theta(1) \cdot \left(1 + \frac{\Delta}{K}\right) - \left(1 - \frac{\Delta}{K}\right) + \pi_\theta(2) \cdot \left(1 - \frac{\Delta}{K}\right) \tag{B.107}$$

$$\geq \pi_\theta(2) \cdot \left(1 - \frac{\Delta}{K}\right) \geq 0, \tag{B.108}$$

$$\text{(using } \theta \in \mathcal{N}_c,\ \pi_\theta(1) \geq c/(c+1)) \tag{B.109}$$

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

**Claim c)** We claim that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. For this, we wish to use the asymptotic convergence results of Agarwal et al. (2019, Theorem 5.1), which states this, but the stepsize there is $\eta \leq 1/5$ while here we have $\eta = 2/5$. We claim that their asymptotic result still hold with the larger $\eta$. In fact, the restriction on $\eta$ comes from that they can only prove the ascent lemma (Lemma 33) for $\eta \leq 1/5$. Other than this, their proof does not rely on the choice of $\eta$. Since we can prove the ascent lemma with $\eta \leq 2/5$ (and in particular with $\eta = 2/5$), their result continues to hold even with $\eta = 2/5$.

Thus, $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. Hence, there exists $t_0 \geq 1$, such that $\pi_{\theta_{t_0}}(a^*) \geq \frac{c}{c+1}$ , which means $\theta_{t_0} \in \mathcal{N}_c \subset \mathcal{R}_1$. According to the first part in our proof, i.e., once $\theta_t$ is in $\mathcal{R}_1$, following gradient update $\theta_{t+1}$ will be in $\mathcal{R}_1$,

and $\pi_{\theta_t}(a^*)$ is increasing in $\mathcal{R}_1$, we have $\inf_t \pi_{\theta_t}(a^*) = \min_{1\le t\le t_0} \pi_{\theta_t}(a^*)$. $t_0$ depends on initialization and $c$, which only depends on the problem. □

**Proposition 2.** For any initialization there exist $t_0 \ge 1$ such that for any $t \ge t_0$, $t \mapsto \pi_{\theta_t}(a^*)$ is increasing. In particular, when $\pi_{\theta_1}$ is the uniform distribution, $t_0 = 1$.

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

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \le 5/(c^2 \cdot t), \tag{B.110}$$

where $c = \inf_{t\ge 1} \pi_{\theta_t}(a^*) > 0$ is a constant that depends on $r$ and $\theta_1$, but it does not depend on the time $t$.

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \le 5K^2/t, \qquad \text{and} \tag{B.111}$$

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \min\left\{K\sqrt{5T},\ 5K^2 \log T + 1\right\}. \tag{B.112}$$

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{5}{c_t^2} \cdot \frac{1}{t}, \tag{B.113}$$

we have $(\pi^* - \pi_{\theta_t})^\top r \leq 5K^2/t$, $\forall t \geq 1$. The remaining results follow from Eq. (B.43) and $c_T \geq 1/K$. □

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

$$t_0 = \min \left\{ t \geq 1 \, : \, \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \geq \frac{r(2) - r(3)}{2 \cdot (r(1) - r(2))} \right\}. \tag{B.114}$$

In general, for $K$-action bandit cases, let $r(1) > r(2) > \cdots > r(K)$, we have,

$$t_0 = \min \left\{ t \geq 1 \, : \, \pi_\theta(1) \geq \frac{\sum_{j \neq 1, j \neq i} \pi_\theta(j) \cdot (r(i) - r(j))}{2 \cdot (r(1) - r(i))}, \tag{B.115}$$

$$\text{for all } i \in \{2, 3, \ldots K-1\} \right\}. \tag{B.116}$$

*Proof.* **3-action case.** Recall the definition of $\mathcal{R}_1$ from the proof for Lemma 5:

$$\mathcal{R}_1 = \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)}, \; \forall a \neq a^* \right\}. \tag{B.117}$$

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

$$r(2) - \pi_\theta^\top r = -\pi_\theta(1) \cdot (r(1) - r(2)) + \pi_\theta(3) \cdot (r(2) - r(3)) \tag{B.118}$$

$$= \pi_\theta(3) \cdot (r(1) - r(2)) \cdot \left[ -\frac{\pi_\theta(1)}{\pi_\theta(3)} + \frac{r(2) - r(3)}{r(1) - r(2)} \right] \tag{B.119}$$

$$\leq 0, \qquad \left( \frac{\pi_\theta(1)}{\pi_\theta(3)} \geq \frac{r(2) - r(3)}{r(1) - r(2)} \right) \tag{B.120}$$

which implies,

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} = \pi_\theta(1) \cdot \left( r(1) - \pi_\theta^\top r \right) - \pi_\theta(2) \cdot \left( r(2) - \pi_\theta^\top r \right) \tag{B.121}$$

$$\geq 0 - 0 = 0. \qquad \left( r(1) - \pi_\theta^\top r > 0 \right) \tag{B.122}$$

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

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(3)} = \pi_\theta(1) \cdot \big(r(1) - \pi_\theta^\top r\big) - \pi_\theta(3) \cdot \big(r(3) - \pi_\theta^\top r\big) \tag{B.123}$$

$$\geq 0 - 0 = 0. \tag{B.124}$$

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

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

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} = \pi_\theta(1) \cdot \big(r(1) - \pi_\theta^\top r\big) - \pi_\theta(2) \cdot \big(r(2) - \pi_\theta^\top r\big) \tag{B.125}$$

$$= 2 \cdot \pi_\theta(1) \cdot \big(r(1) - \pi_\theta^\top r\big) + \pi_\theta(3) \cdot \big(r(3) - \pi_\theta^\top r\big) \tag{B.126}$$

$$\geq \pi_\theta(3) \cdot \left[\frac{r(2) - r(3)}{r(1) - r(2)} \cdot \big(r(1) - \pi_\theta^\top r\big) + \big(r(3) - \pi_\theta^\top r\big)\right] \tag{B.127}$$

$$\left(\text{using } \frac{\pi_\theta(1)}{\pi_\theta(3)} \geq \frac{r(2) - r(3)}{2 \cdot (r(1) - r(2))}\right) \tag{B.128}$$

$$\geq \pi_\theta(3) \cdot \left[\frac{r(2) - r(3)}{r(1) - r(2)} \cdot (r(1) - r(2)) + \big(r(3) - \pi_\theta^\top r\big)\right] \tag{B.129}$$

$$= \pi_\theta(3) \cdot \big(r(2) - \pi_\theta^\top r\big) \geq 0, \tag{B.130}$$

where the second equation is according to

$$\pi_\theta(1) \cdot \big(r(1) - \pi_\theta^\top r\big) + \pi_\theta(2) \cdot \big(r(2) - \pi_\theta^\top r\big) + \pi_\theta(3) \cdot \big(r(3) - \pi_\theta^\top r\big) \tag{B.131}$$

$$= \pi_\theta^\top r - \pi_\theta^\top r = 0, \tag{B.132}$$

and the second inequality is because of

$$r(1) - \pi_\theta^\top r = (1 - \pi_\theta(1)) \cdot r(1) - (\pi_\theta(2) \cdot r(2) + \pi_\theta(3) \cdot r(3)) \tag{B.133}$$

$$= \pi_\theta(2) \cdot (r(1) - r(2)) + \pi_\theta(3) \cdot (r(1) - r(3)) \tag{B.134}$$

$$= (\pi_\theta(2) + \pi_\theta(3)) \cdot (r(1) - r(2)) + \pi_\theta(3) \cdot (r(2) - r(3)) \tag{B.135}$$

$$> (\pi_\theta(2) + \pi_\theta(3)) \cdot (r(1) - r(2)) + \pi_\theta(1) \cdot (r(1) - r(2)) \tag{B.136}$$

$$\left(\text{using } \frac{\pi_\theta(1)}{\pi_\theta(3)} < \frac{r(2) - r(3)}{r(1) - r(2)}\right) \tag{B.137}$$

$$= r(1) - r(2), \tag{B.138}$$

and the last inequality is from

$$r(2) - \pi_\theta^\top r = \pi_\theta(3) \cdot (r(1) - r(2)) \cdot \left[-\frac{\pi_\theta(1)}{\pi_\theta(3)} + \frac{r(2) - r(3)}{r(1) - r(2)}\right] \tag{B.139}$$

$$> 0. \qquad \left(\frac{\pi_\theta(1)}{\pi_\theta(3)} < \frac{r(2) - r(3)}{r(1) - r(2)}\right) \tag{B.140}$$

Now we have $\frac{d\pi_\theta^\top r}{d\theta(1)} \geq \frac{d\pi_\theta^\top r}{d\theta(2)}$. According to Eq. (B.123), we have $\frac{d\pi_\theta^\top r}{d\theta(1)} \geq \frac{d\pi_\theta^\top r}{d\theta(3)}$. Therefore we have $\theta \in \mathcal{R}_1$.

**$K$-action case.** Suppose for each action $i \in \{2, 3, \dots K-1\}$, $\pi_\theta(1) \geq \frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{2\cdot(r(1)-r(i))}$. There are two cases.

Case (a): If $\pi_\theta(1) \geq \frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}$, then we have,

$$r(i) - \pi_\theta^\top r = -\pi_\theta(1)\cdot(r(1)-r(i)) + \sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j)) \tag{B.141}$$

$$\leq 0, \qquad \left(\pi_\theta(1) \geq \frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}\right) \tag{B.142}$$

which implies, for all $i \in \{2, 3, \dots K-1\}$,

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(i)} = \pi_\theta(1)\cdot\left(r(1) - \pi_\theta^\top r\right) - \pi_\theta(i)\cdot\left(r(i) - \pi_\theta^\top r\right) \tag{B.143}$$

$$\geq 0 - 0 = 0. \qquad \left(r(1) - \pi_\theta^\top r > 0\right) \tag{B.144}$$

Similar with Eq. (B.123), since $r(1) > \pi_\theta^\top r$, and $r(K) < \pi_\theta^\top r$, we have

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(K)} = \pi_\theta(1)\cdot\left(r(1) - \pi_\theta^\top r\right) - \pi_\theta(K)\cdot\left(r(K) - \pi_\theta^\top r\right) \tag{B.145}$$

$$\geq 0 - 0 = 0. \tag{B.146}$$

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

Case (b): If $\frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{2\cdot(r(1)-r(i))} \leq \pi_\theta(1) < \frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}$, then we have, for all $i \in \{2, 3, \dots K-1\}$,

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(i)} = \pi_\theta(1)\cdot\left(r(1) - \pi_\theta^\top r\right) - \pi_\theta(i)\cdot\left(r(i) - \pi_\theta^\top r\right) \tag{B.147}$$

$$= 2\cdot\pi_\theta(1)\cdot\left(r(1) - \pi_\theta^\top r\right) + \sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot\left(r(j) - \pi_\theta^\top r\right) \tag{B.148}$$

$$\geq \frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}\cdot\left(r(1) - \pi_\theta^\top r\right) \tag{B.149}$$

$$+ \sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot\left(r(j) - \pi_\theta^\top r\right) \tag{B.150}$$

$$\geq \frac{\sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}\cdot(r(1)-r(i)) \tag{B.151}$$

$$+ \sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot\left(r(j) - \pi_\theta^\top r\right) \tag{B.152}$$

$$= \sum_{j\neq 1, j\neq i} \pi_\theta(j)\cdot\left(r(i) - \pi_\theta^\top r\right) \geq 0, \tag{B.153}$$

where the second equation is according to

$$\pi_\theta(1)\cdot\left(r(1)-\pi_\theta^\top r\right)+\pi_\theta(i)\cdot\left(r(i)-\pi_\theta^\top r\right) \tag{B.154}$$

$$+\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot\left(r(j)-\pi_\theta^\top r\right) \tag{B.155}$$

$$=\pi_\theta^\top r-\pi_\theta^\top r=0, \tag{B.156}$$

and the first inequality is by $r(1)-\pi_\theta^\top r>0$ and,

$$\pi_\theta(1)\geq\frac{\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(i)-r(j))}{2\cdot(r(1)-r(i))}, \tag{B.157}$$

and the second inequality is because of

$$r(1)-\pi_\theta^\top r=\pi_\theta(i)\cdot(r(1)-r(i))+\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(1)-r(j)) \tag{B.158}$$

$$=\sum_{j\neq 1}\pi_\theta(j)\cdot(r(1)-r(i))+\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(i)-r(j)) \tag{B.159}$$

$$>\sum_{j\neq 1}\pi_\theta(j)\cdot(r(1)-r(i))+\pi_\theta(1)\cdot(r(1)-r(i)) \tag{B.160}$$

$$\left(\text{using }\frac{\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}>\pi_\theta(1)\right) \tag{B.161}$$

$$=r(1)-r(i), \tag{B.162}$$

and the last inequality is from $\frac{\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}>\pi_\theta(1)>0$ and,

$$r(i)-\pi_\theta^\top r=-\pi_\theta(1)\cdot(r(1)-r(i))+\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(i)-r(j)) \tag{B.163}$$

$$>0. \qquad \left(\pi_\theta(1)<\frac{\sum_{j\neq 1,j\neq i}\pi_\theta(j)\cdot(r(i)-r(j))}{r(1)-r(i)}\right) \tag{B.164}$$

Now we have $\frac{d\pi_\theta^\top r}{d\theta(1)}\geq\frac{d\pi_\theta^\top r}{d\theta(i)}$, for all $i\in\{2,3,\dots K-1\}$. According to Eq. (B.145), we have $\frac{d\pi_\theta^\top r}{d\theta(1)}\geq\frac{d\pi_\theta^\top r}{d\theta(K)}$. Therefore we have $\theta\in\mathcal{R}_1$. □

### B.1.3 Proofs for Softmax Parametrization in MDPs

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

*Proof.* See Agarwal et al. (2019, Lemma E.4). Our proof is for completeness.

Denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. For any $s \in \mathcal{S}$,

$$\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{B.165}$$

$$= \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta}, u \right\rangle \right|. \tag{B.166}$$

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

$$\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \right\rangle \right| \tag{B.167}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \tag{B.168}$$

$$\leq \max_a |u(s,a)| + |\pi_\theta(\cdot|s)^\top u(s,\cdot)| \leq 2 \cdot \|u\|_2. \tag{B.169}$$

Similarly,

$$\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{B.170}$$

$$= \sum_a \left| \left\langle \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{B.171}$$

$$= \sum_a \left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right|. \tag{B.172}$$

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

$$S_{i,j} = \frac{\partial \{\delta_{ia} \pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(i|s)\}}{\partial \theta(s,j)} \tag{B.173}$$

$$= \delta_{ia} \cdot [\delta_{ja} \pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] \tag{B.174}$$

$$- \pi_\theta(a|s) \cdot [\delta_{ij} \pi_\theta(j|s) - \pi_\theta(i|s)\pi_\theta(j|s)] \tag{B.175}$$

$$- \pi_\theta(i|s) \cdot [\delta_{ja} \pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)], \tag{B.176}$$

where the $\delta$ notation is as defined in Eq. (B.22). Then we have,

$$\left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right| = \left| \sum_{i=1}^A \sum_{j=1}^A S_{i,j} u(s,i) u(s,j) \right| \tag{B.177}$$

$$= \pi_\theta(a|s) \cdot \Big| u(s,a)^2 - 2 \cdot u(s,a) \cdot \pi_\theta(\cdot|s)^\top u(s,\cdot) \tag{B.178}$$

$$- \pi_\theta(\cdot|s)^\top (u(s,\cdot) \odot u(s,\cdot)) + 2 \cdot \left(\pi_\theta(\cdot|s)^\top u(s,\cdot)\right)^2 \Big|. \tag{B.179}$$

Therefore we have,

$$\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \le \max_a \left\{ u(s,a)^2 + 2 \cdot \left| u(s,a) \cdot \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \right\} \tag{B.180}$$

$$+ \pi_\theta(\cdot|s)^\top \left( u(s,\cdot) \odot u(s,\cdot) \right) + 2 \cdot \left( \pi_\theta(\cdot|s)^\top u(s,\cdot) \right)^2 \tag{B.181}$$

$$\le \|u(s,\cdot)\|_2^2 + 2 \cdot \|u(s,\cdot)\|_2^2 + \|u(s,\cdot)\|_2^2 + 2 \cdot \|u(s,\cdot)\|_2^2 \tag{B.182}$$

$$\le 6 \cdot \|u\|_2^2. \tag{B.183}$$

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

$$[P(\alpha)]_{(s,s')} = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \mathcal{P}(s'|s,a). \tag{B.184}$$

The derivative w.r.t. $\alpha$ is

$$\left[ \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{B.185}$$

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

$$\left[ \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} x \right]_{(s)} = \sum_{s'} \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s'). \tag{B.186}$$

The $\ell_\infty$ norm is upper bounded as

$$\left\| \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} x \right\|_\infty = \max_s \left| \sum_{s'} \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \tag{B.187}$$

$$\le \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{B.188}$$

$$= \max_s \sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{B.189}$$

$$\le 2 \cdot \|u\|_2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (B.167))} \tag{B.190}$$

Similarly, taking second derivative w.r.t. $\alpha$,

$$\left[ \frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[ \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{B.191}$$

The $\ell_\infty$ norm is upper bounded as

$$\left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0} x \right\|_\infty = \max_s \left| \sum_{s'} \sum_a \left[ \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \tag{B.192}$$

$$\leq \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{B.193}$$

$$= \max_s \sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{B.194}$$

$$\leq 6 \cdot \|u\|_2^2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (B.180))} \tag{B.195}$$

Next, consider the state value function of $\pi_{\theta_\alpha}$,

$$V^{\pi_{\theta_\alpha}}(s) = \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a) \tag{B.196}$$

$$+ \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot V^{\pi_{\theta_\alpha}}(s'), \tag{B.197}$$

which implies,

$$V^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) r_{\theta_\alpha}, \tag{B.198}$$

where

$$M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1}, \tag{B.199}$$

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

$$r_{\theta_\alpha}(s) = \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a). \tag{B.200}$$

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

$$M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1} = \sum_{t=0}^{\infty} \gamma^t \left[P(\alpha)\right]^t, \tag{B.201}$$

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

$$\mathbf{1} = \frac{1}{1-\gamma} \cdot (\mathbf{Id} - \gamma P(\alpha))\, \mathbf{1} \Longrightarrow M(\alpha)\mathbf{1} = \frac{1}{1-\gamma} \cdot \mathbf{1}, \tag{B.202}$$

which implies, $\forall i$,

$$\left\|[M(\alpha)]_{i,:}\right\|_1 = \sum_j [M(\alpha)]_{(i,j)} = \frac{1}{1-\gamma}. \tag{B.203}$$

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

$$\|M(\alpha)x\|_\infty = \max_i \left|[M(\alpha)]_{i,:}^\top x\right| \tag{B.204}$$

$$\leq \max_i \left\|[M(\alpha)]_{i,:}\right\|_1 \cdot \|x\|_\infty \tag{B.205}$$

$$= \frac{1}{1-\gamma} \cdot \|x\|_\infty. \tag{B.206}$$

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

$$\|r_{\theta_\alpha}\|_\infty = \max_s |r_{\theta_\alpha}(s)| = \max_s \left|\sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a)\right| \leq 1. \tag{B.207}$$

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

$$\left|\frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha}\right| = \left|\left(\frac{\partial r_{\theta_\alpha}(s)}{\partial \theta_\alpha}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \tag{B.208}$$

$$= \left|\left(\frac{\partial \{\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\}}{\partial \theta_\alpha(s,\cdot)}\right)^\top u(s,\cdot)\right| \tag{B.209}$$

$$= \left|\left(H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot)\right)^\top u(s,\cdot)\right| \tag{B.210}$$

$$\leq \left\|H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot)\right\|_1 \cdot \|u(s,\cdot)\|_\infty. \tag{B.211}$$

Similarly to Eq. (B.32), the $\ell_1$ norm is upper bounded as

$$\left\|H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot)\right\|_1 = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left|r(s,a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\right| \tag{B.212}$$

$$\leq \max_a \left|r(s,a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\right| \tag{B.213}$$

$$\leq 1. \qquad (\text{since } r(s,a) \in [0,1]) \tag{B.214}$$

Therefore we have,

$$\left\|\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\right\|_\infty = \max_s \left|\frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha}\right| \tag{B.215}$$

$$\leq \max_s \left\|H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot)\right\|_1 \cdot \|u(s,\cdot)\|_\infty \tag{B.216}$$

$$\leq \|u\|_2. \tag{B.217}$$

Similarly,

$$\left\|\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\right\|_\infty = \max_s \left|\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \alpha^2}\right| \tag{B.218}$$

$$= \max_s \left|\left(\frac{\partial}{\partial \theta_\alpha}\left\{\frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha}\right\}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \tag{B.219}$$

$$= \max_s \left|\left(\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \theta_\alpha^2}\frac{\partial \theta_\alpha}{\partial \alpha}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \tag{B.220}$$

$$= \max_s \left|u(s,\cdot)^\top \frac{\partial^2 \{\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\}}{\partial \theta_\alpha(s,\cdot)^2} u(s,\cdot)\right| \tag{B.221}$$

$$\leq 5/2 \cdot \|u(s,\cdot)\|_2^2 \leq 3 \cdot \|u\|_2^2. \qquad \text{(by Eq. (B.35))} \tag{B.222}$$

Taking derivative w.r.t. $\alpha$ in Eq. (B.198),

$$\frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} = \gamma \cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + e_s^\top M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}. \tag{B.223}$$

Taking second derivative w.r.t. $\alpha$,

$$\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} = 2\gamma^2 \cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \tag{B.224}$$

$$+ \gamma \cdot e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \tag{B.225}$$

$$+ 2\gamma \cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}. \tag{B.226}$$

For the last term,

$$\left|e_s^\top M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| \leq \|e_s\|_1 \cdot \left\|M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right\|_\infty \tag{B.227}$$

$$\leq \frac{1}{1-\gamma} \cdot \left\|\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.228}$$

$$\leq \frac{3}{1-\gamma} \cdot \|u\|_2^2. \qquad \text{(by Eq. (B.218))} \tag{B.229}$$

For the second last term,

$$\left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| \le \left\|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \tag{B.230}$$

$$\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.231}$$

$$\le \frac{2\cdot\|u\|_2}{1-\gamma}\cdot\left\|M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.187))} \tag{B.232}$$

$$\le \frac{2\cdot\|u\|_2}{(1-\gamma)^2}\cdot\left\|\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.233}$$

$$\le \frac{2\cdot\|u\|_2}{(1-\gamma)^2}\cdot\|u\|_2 = \frac{2}{(1-\gamma)^2}\cdot\|u\|_2^2. \qquad \text{(by Eq. (B.215))} \tag{B.234}$$

For the second term,

$$\left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right| \le \left\|M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{B.235}$$

$$\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.236}$$

$$\le \frac{6\cdot\|u\|_2^2}{1-\gamma}\cdot\left\|M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.192))} \tag{B.237}$$

$$\le \frac{6\cdot\|u\|_2^2}{(1-\gamma)^2}\cdot\left\|r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.238}$$

$$\le \frac{6}{(1-\gamma)^2}\cdot\|u\|_2^2. \qquad \text{(by Eq. (B.207))} \tag{B.239}$$

For the first term, according to Eq. (B.187), Eqs. (B.204) and (B.207),

$$\left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{B.240}$$

$$\le \left\|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{B.241}$$

$$\le \frac{1}{1-\gamma}\cdot 2\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot 2\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot 1 \tag{B.242}$$

$$= \frac{4}{(1-\gamma)^3}\cdot\|u\|_2^2. \tag{B.243}$$

Combining Eqs. (B.227), (B.230), (B.235) and (B.240) with Eq. (B.224),

$$\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \leq 2\gamma^2 \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{B.244}$$

$$+\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{B.245}$$

$$+2\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| + \left|e_s^\top M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| \tag{B.246}$$

$$\leq \left(2\gamma^2 \cdot \frac{4}{(1-\gamma)^3} + \gamma \cdot \frac{6}{(1-\gamma)^2} + 2\gamma \cdot \frac{2}{(1-\gamma)^2} + \frac{3}{1-\gamma}\right) \cdot \|u\|_2^2 \tag{B.247}$$

$$\leq \frac{8}{(1-\gamma)^3} \cdot \|u\|_2^2, \tag{B.248}$$

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

$$\left|y^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} y\right| = \left|\left(\frac{y}{\|y\|_2}\right)^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2}\left(\frac{y}{\|y\|_2}\right)\right| \cdot \|y\|_2^2 \tag{B.249}$$

$$\leq \max_{\|u\|_2=1}\left|\left\langle \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} u, u\right\rangle\right| \cdot \|y\|_2^2 \tag{B.250}$$

$$= \max_{\|u\|_2=1}\left|\left\langle \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2}\Big|_{\alpha=0}\frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha}\right\rangle\right| \cdot \|y\|_2^2 \tag{B.251}$$

$$= \max_{\|u\|_2=1}\left|\left\langle \frac{\partial}{\partial \theta_\alpha}\left\{\frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha}\right\}\Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha}\right\rangle\right| \cdot \|y\|_2^2 \tag{B.252}$$

$$= \max_{\|u\|_2=1}\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \cdot \|y\|_2^2 \tag{B.253}$$

$$\leq \frac{8}{(1-\gamma)^3} \cdot \|y\|_2^2. \qquad \text{(by Eq. (B.244))} \tag{B.254}$$

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

$$\left|V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta\right\rangle\right| \tag{B.255}$$

$$= \frac{1}{2} \cdot \left|(\theta' - \theta)^\top \frac{\partial^2 V^{\pi_{\theta_\xi}}(s)}{\partial \theta_\xi^2}(\theta' - \theta)\right| \tag{B.256}$$

$$\leq \frac{4}{(1-\gamma)^3} \cdot \|\theta' - \theta\|_2^2. \qquad \text{(by Eq. (B.249))} \tag{B.257}$$

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

**Lemma 8** (Non-uniform Łojasiewicz)**.** Let $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s,\cdot))$, $s \in \mathcal{S}$ and fix an arbitrary optimal policy $\pi^*$. We have,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot [V^*(\rho) - V^{\pi_\theta}(\rho)], \quad \text{(B.258)}$$

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

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{SA}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s)\right] \cdot [V^*(\rho) - V^{\pi_\theta}(\rho)], \quad \text{(B.259)}$$

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

*Proof.* We have,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 = \left[\sum_{s,a}\left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)}\right)^2\right]^{\frac{1}{2}} \quad \text{(B.260)}$$

$$\geq \left[\sum_{s}\left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))}\right)^2\right]^{\frac{1}{2}} \quad \text{(B.261)}$$

$$\geq \frac{1}{\sqrt{S}} \sum_s \left|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))}\right| \qquad (\|x\|_1 = |\langle \mathbf{1},\ |x|\rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2) \quad \text{(B.262)}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \left|d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot A^{\pi_\theta}(s, a^*(s))\right| \quad \text{(B.263)}$$

(by Lemma 1) $\quad \text{(B.264)}$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot |A^{\pi_\theta}(s, a^*(s))|, \quad \text{(B.265)}$$

where the last inequality is because of $d_\mu^{\pi_\theta}(s) \geq 0$ and $\pi_\theta(a^*(s)|s) \geq 0$. Define

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

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \tag{B.266}$$

$$\geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)}\cdot d_\rho^{\pi^*}(s)\cdot\pi_\theta(a^*(s)|s)\cdot|A^{\pi_\theta}(s,a^*(s))| \tag{B.267}$$

$$\geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s) \tag{B.268}$$

$$\cdot\sum_s d_\rho^{\pi^*}(s)\cdot|A^{\pi_\theta}(s,a^*(s))| \tag{B.269}$$

$$\geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s) \tag{B.270}$$

$$\cdot\sum_s d_\rho^{\pi^*}(s)\cdot A^{\pi_\theta}(s,a^*(s)) \tag{B.271}$$

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s) \tag{B.272}$$

$$\cdot\frac{1}{1-\gamma}\sum_s d_\rho^{\pi^*}(s)\sum_a \pi^*(a|s)\cdot A^{\pi_\theta}(s,a) \tag{B.273}$$

$$= \frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\min_s \pi_\theta(a^*(s)|s)\cdot\left[V^*(\rho)-V^{\pi_\theta}(\rho)\right], \tag{B.274}$$

where the one but last equality used that $\pi^*$ is deterministic and in state $s$ chooses $a^*(s)$ with probability one, and the last equality uses the performance difference formula (Lemma 34).

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

$$\bar{\mathcal{A}}^\pi(s) = \left\{\bar{a}(s)\in\mathcal{A} : Q^\pi(s,\bar{a}(s)) = \max_a Q^\pi(s,a)\right\}. \tag{B.275}$$

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

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{SA}} \sum_{s,a} \left|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)}\right| \qquad \text{(by Cauchy-Schwarz)} \tag{B.276}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \cdot |A^{\pi_\theta}(s,a)| \tag{B.277}$$

$$\text{(by Lemma 1)} \tag{B.278}$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \sum_s d_\mu^{\pi_\theta}(s) \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \cdot |A^{\pi_\theta}(s, \bar{a}(s))| \tag{B.279}$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s)\right] \tag{B.280}$$

$$\cdot \sum_s d_\rho^{\pi^*}(s) \cdot \left|\max_a Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\right|, \tag{B.281}$$

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

$$A^{\pi_\theta}(s, \bar{a}(s)) = \max_a Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s), \tag{B.282}$$

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

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s)\right] \tag{B.283}$$

$$\cdot \sum_s d_\rho^{\pi^*}(s) \cdot \left[\max_a Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)\right] \tag{B.284}$$

$$\geq \frac{1}{\sqrt{SA}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s)\right] \tag{B.285}$$

$$\cdot \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \cdot [Q^{\pi_\theta}(s, a^*(s)) - V^{\pi_\theta}(s)] \tag{B.286}$$

$$= \frac{1}{\sqrt{SA}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s)\right] \tag{B.287}$$

$$\cdot \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \sum_a \pi^*(a|s) \cdot A^{\pi_\theta}(s,a) \tag{B.288}$$

$$= \frac{1}{\sqrt{SA}} \cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s)\right] \tag{B.289}$$

$$\cdot [V^*(\rho) - V^{\pi_\theta}(\rho)], \tag{B.290}$$

where the last equation is again according to Lemma 34. □

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

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

$$\mathcal{R}_1(s) = \left\{\theta : \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a^*(s))} \geq \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a)},\ \forall a \neq a^*\right\}, \tag{B.291}$$

$$\mathcal{R}_2(s) = \left\{\theta : Q^{\pi_\theta}(s,a^*(s)) \geq Q^*(s,a^*(s)) - \Delta^*(s)/2\right\}, \tag{B.292}$$

$$\mathcal{R}_3(s) = \big\{\theta_t : V^{\pi_{\theta_t}}(s) \geq Q^{\pi_{\theta_t}}(s,a^*(s)) - \Delta^*(s)/2, \tag{B.293}$$

$$\text{for all } t \geq 1 \text{ large enough}\big\}, \tag{B.294}$$

$$\mathcal{N}_c(s) = \left\{\theta : \pi_\theta(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}\right\}, \tag{B.295}$$

$$\text{where } c(s) = \frac{A}{(1-\gamma)\cdot\Delta^*(s)} - 1. \tag{B.296}$$

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

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

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

**Claim III.** There exists a finite time $t_0(s) \geq 1$, such that $\theta_{t_0(s)} \in \mathcal{N}_c(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, and thus $\theta_{t_0(s)} \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, which implies $\inf_{t\geq 1}\pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0(s)}\pi_{\theta_t}(a^*(s)|s)$.

**Claim IV.** Define $t_0 = \max_s t_0(s)$. Then, we have

$$\inf_{s\in\mathcal{S},t\geq 1}\pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0}\min_s \pi_{\theta_t}(a^*(s)|s). \tag{B.297}$$

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

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

$$Q^{\pi_{\theta_t}}(s, a^*(s)) \geq Q^*(s, a^*(s)) - \Delta^*(s)/2. \tag{B.298}$$

According to smoothness arguments as Eq. (B.357), we have $V^{\pi_{\theta_{t+1}}}(s') \geq V^{\pi_{\theta_t}}(s')$, and

$$\begin{aligned}
Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) &= Q^{\pi_{\theta_t}}(s, a^*(s)) + Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a^*(s)) && \text{(B.299)}\\
&= Q^{\pi_{\theta_t}}(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s)) \cdot [V^{\pi_{\theta_{t+1}}}(s') - V^{\pi_{\theta_t}}(s')] && \text{(B.300)}\\
&\geq Q^{\pi_{\theta_t}}(s, a^*(s)) + 0 && \text{(B.301)}\\
&\geq Q^*(s, a^*(s)) - \Delta^*(s)/2, && \text{(B.302)}
\end{aligned}$$

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

$$\begin{aligned}
&Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) && \text{(B.303)}\\
&= Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^*(s, a^*(s)) + Q^*(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) && \text{(B.304)}\\
&\geq -\Delta^*(s)/2 + Q^*(s, a^*(s)) - Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) && \text{(B.305)}\\
&\geq -\Delta^*(s)/2 + Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) && \text{(B.306)}\\
&= -\Delta^*(s)/2 + \Delta^*(s) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \cdot [V^*(s') - V^{\pi_{\theta_t}}(s')] && \text{(B.307)}\\
&\geq -\Delta^*(s)/2 + \Delta^*(s) + 0 && \text{(B.308)}\\
&= \Delta^*(s)/2. && \text{(B.309)}
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot A^{\pi_{\theta_t}}(s, a) && \text{(B.310)}\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)]\,. && \text{(B.311)}
\end{aligned}$$

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

$$\begin{aligned}
&\pi_{\theta_t}(a^*(s)|s) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] && \text{(B.312)}\\
&\geq \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)]\,. && \text{(B.313)}
\end{aligned}$$

Similarly to the first part in the proof for Lemma 5. There are two cases.
Case (a): If $\pi_{\theta_t}(a^*(s)|s) \geq \pi_{\theta_t}(a|s)$, then $\theta_t(s, a^*(s)) \geq \theta_t(s, a)$. After an update of the parameters,

$$\theta_{t+1}(s, a^*(s)) = \theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \tag{B.314}$$

$$\geq \theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} = \theta_{t+1}(s, a), \tag{B.315}$$

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

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \tag{B.316}$$

$$\geq \pi_{\theta_{t+1}}(a|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)], \tag{B.317}$$

which is equivalent to $\frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s, a)}$, i.e., $\theta_{t+1} \in \mathcal{R}_1(s)$.
Case (b): If $\pi_{\theta_t}(a^*(s)|s) < \pi_{\theta_t}(a|s)$, then by $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}$,

$$\pi_{\theta_t}(a^*(s)|s) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] \tag{B.318}$$

$$\geq \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)] \tag{B.319}$$

$$= \pi_{\theta_t}(a|s) \cdot \big[Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s) \tag{B.320}$$

$$+ Q^{\pi_{\theta_t}}(s, a) - Q^{\pi_{\theta_t}}(s, a^*(s))\big], \tag{B.321}$$

which, after rearranging, is equivalent to

$$Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) \tag{B.322}$$

$$\geq \left(1 - \frac{\pi_{\theta_t}(a^*(s)|s)}{\pi_{\theta_t}(a|s)}\right) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] \tag{B.323}$$

$$= (1 - \exp\{\theta_t(s, a^*(s)) - \theta_t(s, a)\}) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)]. \tag{B.324}$$

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

$$Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \leq \Delta^*(s)/2 \tag{B.325}$$

$$\leq Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a). \tag{B.326}$$

On the other hand,

$$\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a) \tag{B.327}$$
$$= \theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} - \theta_t(s, a) - \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} \tag{B.328}$$
$$\geq \theta_t(s, a^*(s)) - \theta_t(s, a), \tag{B.329}$$

which implies

$$1 - \exp\left\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\right\} \tag{B.330}$$
$$\leq 1 - \exp\left\{\theta_t(s, a^*(s)) - \theta_t(s, a)\right\}. \tag{B.331}$$

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

$$\left(1 - \exp\left\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\right\}\right) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)\right] \tag{B.332}$$
$$\leq Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a), \tag{B.333}$$

which after rearranging is equivalent to

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)\right] \tag{B.334}$$
$$\geq \pi_{\theta_{t+1}}(a|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)\right], \tag{B.335}$$

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

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

$$\pi_{\theta_{t+1}}(a^*(s)|s) = \frac{\exp\left\{\theta_{t+1}(s, a^*(s))\right\}}{\sum_a \exp\left\{\theta_{t+1}(s, a)\right\}} \tag{B.336}$$
$$= \frac{\exp\left\{\theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))}\right\}}{\sum_a \exp\left\{\theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}\right\}} \tag{B.337}$$
$$\geq \frac{\exp\left\{\theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))}\right\}}{\sum_a \exp\left\{\theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))}\right\}} \tag{B.338}$$
$$\left(\text{because } \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}\right) \tag{B.339}$$
$$= \frac{\exp\left\{\theta_t(s, a^*(s))\right\}}{\sum_a \exp\left\{\theta_t(s, a)\right\}} = \pi_{\theta_t}(a^*(s)|s). \tag{B.340}$$

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

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot [Q^{\pi_\theta}(s, a^*(s)) - V^{\pi_\theta}(s)] \tag{B.341}$$

$$> \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot [Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)] \tag{B.342}$$

$$= \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a)}, \tag{B.343}$$

where the inequality is since $Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s, a) \geq \Delta^*(s)/2 > 0$, $\forall a \neq a^*(s)$, similarly to Eq. (B.303).

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

$$\pi_\theta(a^*(s)|s) + \pi_\theta(a|s) > \frac{2 \cdot c(s)}{c(s)+1} = 2 - \frac{2 \cdot (1-\gamma) \cdot \Delta^*(s)}{A} > 1, \tag{B.344}$$

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

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

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

**Claim IV.** Define $t_0 = \max_s t_0(s)$. Then we have

$$\inf_{s\in\mathcal{S},t\geq 1} \pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0} \min_s \pi_{\theta_t}(a^*(s)|s). \qquad \square$$

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

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{16S}{c^2(1-\gamma)^6 t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty . \tag{B.345}$$

*Proof.* Let us first note that for any $\theta$ and $\mu$,

$$\begin{aligned}
d_\mu^{\pi_\theta}(s) &= \mathop{\mathbb{E}}_{s_0\sim\mu} \left[ d_\mu^{\pi_\theta}(s) \right] && \text{(B.346)}\\
&= \mathop{\mathbb{E}}_{s_0\sim\mu} \left[ (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | s_0, \pi_\theta, \mathcal{P}) \right] && \text{(B.347)}\\
&\geq \mathop{\mathbb{E}}_{s_0\sim\mu} \left[ (1-\gamma) \Pr(s_0 = s|s_0) \right] && \text{(B.348)}\\
&= (1-\gamma)\cdot\mu(s)\,. && \text{(B.349)}
\end{aligned}$$

According to the value sub-optimality lemma of Lemma 36,

$$\begin{aligned}
V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) && \text{(B.350)}\\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) && \text{(B.351)}\\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) && \text{(B.352)}\\
&\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) && \text{(B.353)}\\
&\quad \Big( \text{by Eq. (B.346) and } \min_s \mu(s) > 0 \Big) && \text{(B.354)}\\
&= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[ V^*(\mu) - V^{\pi_\theta}(\mu) \right], && \text{(B.355)}
\end{aligned}$$

where the first inequality is because of

$$\sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \geq 0, \tag{B.356}$$

and the last equation is again by Lemma 36. According to Lemma 7, $V^{\pi_\theta}(\mu)$ is $\beta$-smooth with $\beta = 8/(1-\gamma)^3$. Denote $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. And note

$\eta = \frac{(1-\gamma)^3}{8}$. We have,

$$\delta_{t+1} - \delta_t = V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) \tag{B.357}$$

$$\leq -\frac{(1-\gamma)^3}{16} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 \qquad \text{(by Lemma 33)} \tag{B.358}$$

$$\leq -\frac{(1-\gamma)^3}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-2} \cdot \left[ \min_s \pi_{\theta_t}(a^*(s)|s) \right]^2 \tag{B.359}$$

$$\cdot \left[ V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right]^2 \text{(by Lemma 8)} \tag{B.360}$$

$$\leq -\frac{(1-\gamma)^5}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \left[ \min_s \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot \delta_t^2 \tag{B.361}$$

$$\leq -\frac{(1-\gamma)^5}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \left[ \inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot \delta_t^2, \tag{B.362}$$

where the second to last inequality is by $d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (B.346)). According to Lemma 9, $c = \inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$. Using similar induction arguments as in Eq. (B.52),

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \leq \frac{16S}{c^2(1-\gamma)^5 t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2, \tag{B.363}$$

which leads to the final result,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[ V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right] \tag{B.364}$$

$$\leq \frac{16S}{c^2(1-\gamma)^6 t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty, \tag{B.365}$$

thus, finishing the proof. □

## B.2 Proofs for Section 2.4: Entropy Regularized Softmax Policy Gradient

### B.2.1 Preliminaries

**Lemma 10.** Entropy regularized policy gradient w.r.t. $\theta$ is

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a) \tag{B.366}$$

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s)\right] \tag{B.367}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot)\right], \ \forall s \tag{B.368}$$

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

$$\tilde{A}^{\pi_\theta}(s,a) = \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \tag{B.369}$$

$$\tilde{Q}^{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s'). \tag{B.370}$$

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

$$\tilde{V}^{\pi_\theta}(\mu) = \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right]. \tag{B.371}$$

Taking derivative w.r.t. $\theta$,

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} = \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{B.372}$$

$$+ \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} - \tau \cdot \frac{1}{\pi_\theta(a|s)} \cdot \frac{\partial \pi_\theta(a|s)}{\partial \theta}\right] \tag{B.373}$$

$$= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{B.374}$$

$$+ \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} \tag{B.375}$$

$$= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{B.376}$$

$$+ \gamma \cdot \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \frac{\partial \tilde{V}^{\pi_\theta}(s')}{\partial \theta} \tag{B.377}$$

$$= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right], \tag{B.378}$$

where the second equation is because of

$$\sum_a \pi_\theta(a|s) \cdot \left[\frac{1}{\pi_\theta(a|s)} \cdot \frac{\partial \pi_\theta(a|s)}{\partial \theta}\right] = \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \tag{B.379}$$

$$= \frac{\partial}{\partial \theta} \sum_a \pi_\theta(a|s) = \frac{\partial 1}{\partial \theta} = 0. \tag{B.380}$$

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

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{B.381}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d\pi(\cdot|s)}{d\theta(s,\cdot)}\right)^\top \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s)\right] \tag{B.382}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s)\right] \tag{B.383}$$

$$\text{(by Eq. (2.8))} \tag{B.384}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \Big[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(\cdot|s) \tag{B.385}$$

$$+ \tau \log \sum_a \exp\{\theta(s,a)\} \cdot \mathbf{1}\Big] \tag{B.386}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(\cdot|s)\right], \tag{B.387}$$

where the last line is from $H(\pi_\theta(\cdot|s))\mathbf{1} = \mathbf{0}$ in Lemma 37. For each component $a$, we have

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \tag{B.388}$$

$$- \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right]\Big] \tag{B.389}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s)\right] \tag{B.390}$$

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

### B.2.2 Proofs for Bandits and Non-uniform Contraction

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

$$\|\zeta_{t+1}\|_2 \leq \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a)\right) \cdot \|\zeta_t\|_2, \tag{B.391}$$

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

*Proof.* Update 2 can be written as

$$\theta_{t+1} = \theta_t - \eta \cdot H(\pi_{\theta_t})(\tau \log \pi_{\theta_t} - r) \tag{B.392}$$

$$= \theta_t - \eta \cdot H(\pi_{\theta_t}) \left[ \tau\theta_t - r - \left( \log \sum_a \exp\{\theta_t(a)\} \right) \cdot \mathbf{1} \right] \tag{B.393}$$

$$= \theta_t - \eta \cdot H(\pi_{\theta_t})(\tau\theta_t - r) \tag{B.394}$$

$$= \theta_t - \eta \cdot H(\pi_{\theta_t}) \left( \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right), \tag{B.395}$$

where the last two equations are from $H(\pi_{\theta_t})\mathbf{1} = \mathbf{0}$ as shown in Lemma 37. For all $t \geq 1$,

$$\zeta_{t+1} = \tau\theta_{t+1} - r - \frac{(\tau\theta_{t+1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{B.396}$$

$$= \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\theta_{t+1} - \theta_t) \tag{B.397}$$

$$+ \left( \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - \frac{(\tau\theta_{t+1} - r)^\top \mathbf{1}}{K} \right) \cdot \mathbf{1} \tag{B.398}$$

$$= \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\theta_{t+1} - \theta_t) \tag{B.399}$$

$$+ \frac{\tau(\theta_t - \theta_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1}. \tag{B.400}$$

For the last term,

$$\frac{\tau(\theta_t - \theta_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{B.401}$$

$$= \frac{\tau}{K} \cdot \left( \eta \cdot H(\pi_{\theta_t}) \left( \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right)^\top \mathbf{1} \cdot \mathbf{1} = \mathbf{0}, \tag{B.402}$$

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

$$\zeta_{t+1} = \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\theta_{t+1} - \theta_t) \tag{B.403}$$

$$= (\mathbf{Id} - \tau\eta \cdot H(\pi_{\theta_t})) \left( \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \tag{B.404}$$

$$= (\mathbf{Id} - \tau\eta \cdot H(\pi_{\theta_t}))\, \zeta_t. \tag{B.405}$$

According to Lemma 38, with $\tau\eta \leq 1$,

$$\|\zeta_{t+1}\|_2 = \|(\mathbf{Id} - \tau\eta \cdot H(\pi_{\theta_t}))\, \zeta_t\|_2 \tag{B.406}$$

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

**Lemma 12.** Let $\pi_{\theta_t} = \text{softmax}(\theta_t)$. Using Update 2 with $\tau\eta \le 1$, $\forall t \ge 1$,

$$\|\zeta_t\|_2 \le \frac{2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\exp\left\{\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}}. \tag{B.407}$$

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

$$\begin{aligned}
\|\zeta_{t+1}\|_2 &\le \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a)\right) \cdot \|\zeta_t\|_2 && \text{(B.408)}\\
&\le \frac{1}{\exp\{\tau\eta \cdot \min_a \pi_{\theta_t}(a)\}} \cdot \|\zeta_t\|_2 && \text{(B.409)}\\
&\le \frac{1}{\exp\{\tau\eta \cdot \min_a \pi_{\theta_t}(a)\}} \cdot \left(1 - \tau\eta \cdot \min_a \pi_{\theta_{t-1}}(a)\right) \cdot \|\zeta_{t-1}\|_2 && \text{(B.410)}\\
&\le \frac{1}{\exp\left\{\tau\eta \sum_{s=t-1}^{t} [\min_a \pi_{\theta_s}(a)]\right\}} \cdot \|\zeta_{t-1}\|_2 && \text{(B.411)}\\
&\le \frac{1}{\exp\left\{\tau\eta \sum_{s=1}^{t} [\min_a \pi_{\theta_s}(a)]\right\}} \cdot \|\zeta_1\|_2. && \text{(B.412)}
\end{aligned}$$

For the initial logit $\theta_1$,

$$\begin{aligned}
\|\zeta_1\|_2 &= \left\|\tau\theta_1 - r - \frac{(\tau\theta_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 && \text{(B.413)}\\
&\le \|\tau\theta_1 - r\|_2 + \left\|\frac{(\tau\theta_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \quad \text{(by triangle inequality)} && \text{(B.414)}\\
&= \|\tau\theta_1 - r\|_2 + \frac{\left|(\tau\theta_1 - r)^\top \mathbf{1}\right|}{\sqrt{K}} && \text{(B.415)}\\
&\le \|\tau\theta_1 - r\|_2 + \frac{\|\tau\theta_1 - r\|_2 \cdot \|\mathbf{1}\|_2}{\sqrt{K}} \quad \text{(by Cauchy-Schwarz)} && \text{(B.416)}\\
&= 2 \cdot \|\tau\theta_1 - r\|_2 && \text{(B.417)}\\
&\le 2 \cdot (\|\tau\theta_1\|_2 + \|r\|_2) && \text{(B.418)}\\
&\le 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}\,, && \text{(B.419)}
\end{aligned}$$

finishing the proof. □

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

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

$$c = \frac{1}{K} \cdot \frac{1}{\exp\{1/\tau\}} \cdot \frac{1}{\exp\{4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\}}. \tag{B.420}$$

First, according to Eq. (B.413), we have,

$$\|\zeta_1\|_2 \leq 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}. \tag{B.421}$$

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

$$\|\zeta_{t+1}\|_2 \leq \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a)\right) \cdot \|\zeta_t\|_2 \tag{B.422}$$

$$\leq 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}. \tag{B.423}$$

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

$$\|\zeta_t\|_2 \leq 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}. \tag{B.424}$$

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

$$\left|\theta_t(a) - \frac{r(a)}{\tau} - \frac{(\theta_t - r/\tau)^\top \mathbf{1}}{K}\right| = \frac{1}{\tau} \cdot \left|\tau\theta_t(a) - r(a) - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K}\right| \tag{B.425}$$

$$\leq \frac{1}{\tau} \cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \tag{B.426}$$

$$= \frac{1}{\tau} \cdot \|\zeta_t\|_2 \tag{B.427}$$

$$\leq 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}. \tag{B.428}$$

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

$$\theta_t(a_1) \geq \frac{r(a_1)}{\tau} + \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}, \tag{B.429}$$

$$-\theta_t(a_2) \geq -\frac{r(a_2)}{\tau} - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}, \tag{B.430}$$

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

$$\min_a \pi_{\theta_t}(a) = \frac{\exp\{\theta_t(a_1)\}}{\sum_a \exp\{\theta_t(a)\}} \tag{B.431}$$

$$\geq \frac{\exp\{\theta_t(a_1)\}}{\sum_a \exp\{\theta_t(a_2)\}} \tag{B.432}$$

$$= \frac{1}{K} \cdot \exp\left\{\theta_t(a_1) - \theta_t(a_2)\right\}, \qquad (\text{since } \theta_t(a) \leq \theta_t(a_2),\ \forall a) \tag{B.433}$$

which can be further lower bounded using the above results,

$$\min_a \pi_{\theta_t}(a) \geq \frac{1}{K} \cdot \exp\left\{\theta_t(a_1) - \theta_t(a_2)\right\} \tag{B.434}$$

$$\geq \frac{1}{K} \cdot \exp\left\{\frac{r(a_1)}{\tau} + \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right. \tag{B.435}$$

$$\left. - \frac{r(a_2)}{\tau} - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right\} \tag{B.436}$$

$$= \frac{1}{K} \cdot \exp\left\{\frac{r(a_1) - r(a_2)}{\tau} - 4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right\} \tag{B.437}$$

$$\geq \frac{1}{K} \cdot \exp\left\{-\frac{1}{\tau} - 4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right\} \tag{B.438}$$

$$\left(\text{because } r \in [0,1]^K \text{ and } r(a_1) - r(a_2) \geq -1\right) \tag{B.439}$$

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

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

$$\left(\pi_\tau^* - \pi_{\theta_t}\right)^\top r \leq \frac{2\sqrt{K}(\|\theta_1\|_\infty + 1/\tau)}{\exp\left\{\tau\eta \cdot c \cdot (t-1)\right\}}, \tag{B.440}$$

$$\tilde{\delta}_t \leq \frac{2(\tau\|\theta_1\|_\infty + 1)^2 K/\tau}{\exp\left\{2\tau\eta \cdot c \cdot (t-1)\right\}}, \tag{B.441}$$

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

*Proof.* According to Hölder's inequality,

$$\left(\pi_\tau^* - \pi_{\theta_t}\right)^\top r \tag{B.442}$$

$$\leq \|\pi_\tau^* - \pi_{\theta_t}\|_1 \cdot \|r\|_\infty \qquad \text{(by Hölder's inequality)} \tag{B.443}$$

$$\leq \|\pi_\tau^* - \pi_{\theta_t}\|_1 \qquad \left(\text{because } r \in [0,1]^K\right) \tag{B.444}$$

$$\leq \left\|\frac{r}{\tau} - \theta_t + \frac{(\tau\theta_t - r)^\top \mathbf{1}}{\tau K} \cdot \mathbf{1}\right\|_\infty \qquad \text{(by Lemma 39)} \tag{B.445}$$

$$= \frac{1}{\tau} \cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_\infty \tag{B.446}$$

$$\leq \frac{1}{\tau} \cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \tag{B.447}$$

$$\leq \frac{1}{\tau} \cdot \frac{2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\exp\left\{\tau\eta \sum_{s=1}^{t-1} \left[\min_a \pi_{\theta_s}(a)\right]\right\}} \qquad \text{(by Lemma 12)} \tag{B.448}$$

$$\leq \frac{2\sqrt{K}}{\tau} \cdot \frac{\tau\|\theta_1\|_\infty + 1}{\exp\left\{\tau\eta \cdot c \cdot (t-1)\right\}}. \qquad \text{(by Lemma 13)} \tag{B.449}$$

On the other hand, we have,

$$
\begin{aligned}
&{\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_{\theta_t}}^\top (r - \tau \log \pi_{\theta_t}) && \text{(B.450)}\\
&= {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_{\theta_t}}^\top (r - \tau \log \pi_\tau^* + \tau \log \pi_\tau^* - \tau \log \pi_{\theta_t}) && \text{(B.451)}\\
&= (\pi_\tau^* - \pi_{\theta_t})^\top (r - \tau \log \pi_\tau^*) + \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t} \| \pi_\tau^*) && \text{(B.452)}\\
&= (\pi_\tau^* - \pi_{\theta_t})^\top \mathbf{1} \cdot \tau \cdot \log \sum_a \exp\{r(a)/\tau\} + \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t} \| \pi_\tau^*) && \text{(B.453)}\\
&= \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t} \| \pi_\tau^*) && \text{(B.454)}\\
&\le \frac{\tau}{2} \cdot \left\| \theta_t - \frac{r}{\tau} - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{\tau K} \cdot \mathbf{1} \right\|_\infty^2 \qquad \text{(by Lemma 42)} && \text{(B.455)}\\
&= \frac{1}{2\tau} \cdot \left\| \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 && \text{(B.456)}\\
&\le \frac{1}{2\tau} \cdot \left\| \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2^2 && \text{(B.457)}\\
&\le \frac{1}{2\tau} \cdot \frac{4(\tau\|\theta_1\|_\infty + 1)^2 K}{\exp\left\{2\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}} \qquad \text{(by Lemma 12)} && \text{(B.458)}\\
&\le \frac{1}{\tau} \cdot \frac{2(\tau\|\theta_1\|_\infty + 1)^2 K}{\exp\{2\tau\eta \cdot c \cdot (t-1)\}}. \qquad \text{(by Lemma 13)} \qquad \square
\end{aligned}
$$

### B.2.3 Proofs for MDPs and Entropy Regularization

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

*Proof.* Denote $\mathbb{H}^{\pi_\theta}(s) = \mathbb{H}(s, \pi_\theta)$. Also denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. According to Eq. (2.33),

$$
\begin{aligned}
\mathbb{H}^{\pi_{\theta_\alpha}}(s) &= \mathop{\mathbb{E}}_{\substack{s_0 = s, a_t \sim \pi_{\theta_\alpha}(\cdot|s_t),\\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[ \sum_{t=0}^{\infty} -\gamma^t \log \pi_{\theta_\alpha}(a_t|s_t) \right] && \text{(B.459)}\\
&= -\sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) && \text{(B.460)}\\
&\quad + \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \mathbb{H}^{\pi_{\theta_\alpha}}(s'), && \text{(B.461)}
\end{aligned}
$$

which implies,

$$
\mathbb{H}^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) h_{\theta_\alpha}, \qquad \text{(B.462)}
$$

where $M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1}$ is defined in Eq. (B.199), $P(\alpha)$ is defined in Eq. (B.184), and $h_{\theta_\alpha} \in \mathbb{R}^S$ for $s \in \mathcal{S}$ is given by

$$h_{\theta_\alpha}(s) = -\sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s). \tag{B.463}$$

According to Eq. (B.463), $h_{\theta_\alpha}(s) \in [0, \log A]$, $\forall s$. Then we have,

$$\|h_{\theta_\alpha}\|_\infty = \max_s |h_{\theta_\alpha}(s)| \leq \log A. \tag{B.464}$$

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

$$\left|\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right| = \left|\left\langle \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle\right| \tag{B.465}$$

$$= \left|\left\langle \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(\cdot|s)}, u(s,\cdot) \right\rangle\right| \tag{B.466}$$

$$= \left|\left(H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s)\right)^\top u(s,\cdot)\right| \tag{B.467}$$

$$\leq \left\|H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s)\right\|_1 \cdot \|u(s,\cdot)\|_\infty . \tag{B.468}$$

The $\ell_1$ norm is upper bounded as

$$\left\|H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s)\right\|_1 \tag{B.469}$$

$$= \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left|\log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s)\right| \tag{B.470}$$

$$\leq \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left(|\log \pi_{\theta_\alpha}(a|s)| + \left|\pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s)\right|\right) \tag{B.471}$$

$$= -2 \cdot \sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) \leq 2 \cdot \log A. \tag{B.472}$$

Therefore we have,

$$\left\|\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\right\|_\infty = \max_s \left|\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right| \tag{B.473}$$

$$\leq \max_s \left\|H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s)\right\|_1 \cdot \|u(s,\cdot)\|_\infty \tag{B.474}$$

$$\leq 2 \cdot \log A \cdot \|u\|_2. \tag{B.475}$$

The second derivative w.r.t. $\alpha$ is

$$\left|\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2}\right| = \left|\left(\frac{\partial}{\partial \theta_\alpha}\left\{\frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha}\right\}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \tag{B.476}$$

$$= \left|\left(\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2}\frac{\partial \theta_\alpha}{\partial \alpha}\right)^\top \frac{\partial \theta_\alpha}{\partial \alpha}\right| \tag{B.477}$$

$$= \left|u(s,\cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} u(s,\cdot)\right| . \tag{B.478}$$

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

$$T(s,\theta_\alpha) = \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} = \frac{\partial}{\partial \theta_\alpha(s,\cdot)} \left\{ \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(s,\cdot)} \right\} \tag{B.479}$$

$$= \frac{\partial}{\partial \theta_\alpha(s,\cdot)} \left\{ \left( \frac{\partial \pi_{\theta_\alpha}(\cdot|s)}{\partial \theta_\alpha(s,\cdot)} \right)^\top \frac{\partial h_{\theta_\alpha}(s)}{\partial \pi_{\theta_\alpha}(\cdot|s)} \right\} \tag{B.480}$$

$$= \frac{\partial}{\partial \theta_\alpha(s,\cdot)} \left\{ H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s)) \right\}. \tag{B.481}$$

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

$$T_{i,j} = \frac{d\{\pi_{\theta_\alpha}(i|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s))\}}{d\theta_\alpha(s,j)} \tag{B.482}$$

$$= \frac{d\pi_{\theta_\alpha}(i|s)}{d\theta_\alpha(s,j)} \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)) \tag{B.483}$$

$$+ \pi_{\theta_\alpha}(i|s) \cdot \frac{d\{-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)\}}{d\theta_\alpha(s,j)} \tag{B.484}$$

$$= (\delta_{ij}\pi_{\theta_\alpha}(j|s) - \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s)) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)) \tag{B.485}$$

$$+ \pi_{\theta_\alpha}(i|s) \cdot \Bigg( -\frac{1}{\pi_{\theta_\alpha}(i|s)} \cdot (\delta_{ij}\pi_{\theta_\alpha}(j|s) - \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s)) \tag{B.486}$$

$$- \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s)) \Bigg) \tag{B.487}$$

$$= \delta_{ij}\pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) \tag{B.488}$$

$$- \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) \tag{B.489}$$

$$- \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s)). \tag{B.490}$$

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

$$\left|y^\top T(s,\theta_\alpha) y\right| = \left|\sum_{i=1}^{A}\sum_{j=1}^{A} T_{i,j}\, y(i) y(j)\right| \tag{B.491}$$

$$\leq \left|\sum_i \pi_{\theta_\alpha}(i|s)\cdot(-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1)\cdot y(i)^2\right| \tag{B.492}$$

$$+ 2\cdot\left|\sum_i \pi_{\theta_\alpha}(i|s)\cdot y(i)\right. \tag{B.493}$$

$$\left.\cdot \sum_j \pi_{\theta_\alpha}(j|s)\cdot(-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s))\cdot y(j)\right| \tag{B.494}$$

$$+ \left(\pi_{\theta_\alpha}(\cdot|s)^\top y\right)^2 \tag{B.495}$$

$$= \left|\left(H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s)\right)^\top (y\odot y)\right| \tag{B.496}$$

$$+ 2\cdot\left|\left(\pi_{\theta_\alpha}(\cdot|s)^\top y\right)\cdot\left(H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right)^\top y\right| \tag{B.497}$$

$$+ \left(\pi_{\theta_\alpha}(\cdot|s)^\top y\right)^2 \tag{B.498}$$

$$\leq \left\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\|_\infty \cdot \|y\odot y\|_1 \tag{B.499}$$

$$+ \|\pi_{\theta_\alpha}(\cdot|s)\|_\infty \cdot \|y\odot y\|_1 \tag{B.500}$$

$$+ 2\cdot \|\pi_{\theta_\alpha}(\cdot|s)\|_1 \cdot \|y\|_\infty \tag{B.501}$$

$$\cdot \left\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\|_1 \cdot \|y\|_\infty \tag{B.502}$$

$$+ \|\pi_{\theta_\alpha}(\cdot|s)\|_2^2 \cdot \|y\|_2^2, \tag{B.503}$$

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

$$\left\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\|_\infty \tag{B.504}$$

$$= \max_a \left|\pi_{\theta_\alpha}(a|s)\cdot\left(-\log \pi_{\theta_\alpha}(a|s) + \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s)\right)\right| \tag{B.505}$$

$$\leq \max_a -\pi_{\theta_\alpha}(a|s)\cdot\log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \tag{B.506}$$

$$\leq \frac{1}{e} + \log A. \qquad \left(\text{since } -x\cdot\log x \leq \frac{1}{e} \text{ for all } x\in[0,1]\right) \tag{B.507}$$

Therefore we have,

$$\begin{aligned}
\left|y^\top T(s,\theta_\alpha)y\right| &\le \left\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\|_\infty \cdot \|y\|_2^2 && \text{(B.508)}\\
&\quad + \|y\|_2^2 + 2\cdot \left\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\|_1 \cdot \|y\|_2^2 + \|y\|_2^2 && \text{(B.509)}\\
&\le \left(\frac{1}{e} + \log A + 2\right)\cdot \|y\|_2^2 && \text{(B.510)}\\
&\quad + 2\cdot \left\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\right\|_1 \cdot \|y\|_2^2 && \text{(B.511)}\\
&\quad \text{(by Eq. (B.504))} && \text{(B.512)}\\
&\le \left(\frac{1}{e} + \log A + 2 + 2\cdot \log A\right)\cdot \|y\|_2^2 \qquad \text{(by Eq. (B.469))} && \text{(B.513)}\\
&\le 3\cdot(1+\log A)\cdot \|y\|_2^2. && \text{(B.514)}
\end{aligned}$$

According to the above results,

$$\begin{aligned}
\left\|\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\right\|_\infty &= \max_s \left|\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2}\right| && \text{(B.515)}\\
&= \max_s \left|u(s,\cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} u(s,\cdot)\right| && \text{(B.516)}\\
&= \max_s \left|u(s,\cdot)^\top T(s,\theta_\alpha) u(s,\cdot)\right| && \text{(B.517)}\\
&\le 3\cdot(1+\log A)\cdot \max_s \|u(s,\cdot)\|_2^2 && \text{(B.518)}\\
&\le 3\cdot(1+\log A)\cdot \|u\|_2^2. && \text{(B.519)}
\end{aligned}$$

Taking derivative w.r.t. $\alpha$ in Eq. (B.462),

$$\frac{\partial \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} = \gamma \cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} + e_s^\top M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha}. \tag{B.520}$$

Taking second derivative w.r.t. $\alpha$,

$$\begin{aligned}
\frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} &= 2\gamma^2 \cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} && \text{(B.521)}\\
&\quad + \gamma \cdot e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} && \text{(B.522)}\\
&\quad + 2\gamma \cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha} M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha)\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}. && \text{(B.523)}
\end{aligned}$$

For the last term,

$$\begin{aligned}
\left|e_s^\top M(\alpha)\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| &\le \|e_s\|_1 \cdot \left\|M(\alpha)\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right\|_\infty && \text{(B.524)}\\
&\le \frac{1}{1-\gamma}\cdot \left\|\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} && \text{(B.525)}\\
&\le \frac{3\cdot(1+\log A)}{1-\gamma}\cdot \|u\|_2^2. \qquad \text{(by Eq. (B.515))} && \text{(B.526)}
\end{aligned}$$

For the second last term,

$$\left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| \le \left\|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \tag{B.527}$$

$$\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.528}$$

$$\le \frac{2\cdot\|u\|_2}{1-\gamma}\cdot\left\|M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.187))} \tag{B.529}$$

$$\le \frac{2\cdot\|u\|_2}{(1-\gamma)^2}\cdot\left\|\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.530}$$

$$\le \frac{2\cdot\|u\|_2}{(1-\gamma)^2}\cdot 2\cdot\log A\cdot\|u\|_2 \tag{B.531}$$

$$= \frac{4\cdot\log A}{(1-\gamma)^2}\cdot\|u\|_2^2. \qquad \text{(by Eq. (B.473))} \tag{B.532}$$

For the second term,

$$\left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right| \le \left\|M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{B.533}$$

$$\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.534}$$

$$\le \frac{6\cdot\|u\|_2^2}{1-\gamma}\cdot\left\|M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.192))} \tag{B.535}$$

$$\le \frac{6\cdot\|u\|_2^2}{(1-\gamma)^2}\cdot\left\|h_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \qquad \text{(by Eq. (B.204))} \tag{B.536}$$

$$\le \frac{6\cdot\log A}{(1-\gamma)^2}\cdot\|u\|_2^2. \qquad \text{(by Eq. (B.464))} \tag{B.537}$$

For the first term, according to Eqs. (B.187) and (B.204), Eq. (B.464),

$$\left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{B.538}$$

$$\le \left\|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{B.539}$$

$$\le \frac{1}{1-\gamma}\cdot 2\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot 2\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot\log A \tag{B.540}$$

$$= \frac{4\cdot\log A}{(1-\gamma)^3}\cdot\|u\|_2^2. \tag{B.541}$$

Combining Eqs. (B.524), (B.527), (B.533) and (B.538) with Eq. (B.521),

$$\left|\frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \leq 2\gamma^2 \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{B.542}$$

$$+\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)h_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{B.543}$$

$$+2\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial h_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| \tag{B.544}$$

$$+\left|e_s^\top M(\alpha)\frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| \tag{B.545}$$

$$\leq \left(2\gamma^2 \cdot \frac{4\cdot\log A}{(1-\gamma)^3} + \gamma \cdot \frac{6\cdot\log A}{(1-\gamma)^2}\right. \tag{B.546}$$

$$\left.+2\gamma \cdot \frac{4\cdot\log A}{(1-\gamma)^2} + \frac{3\cdot(1+\log A)}{1-\gamma}\right) \cdot \|u\|_2^2 \tag{B.547}$$

$$\leq \left(\frac{8\cdot\log A}{(1-\gamma)^3} + \frac{3}{1-\gamma}\right) \cdot \|u\|_2^2 \tag{B.548}$$

$$\leq \frac{4+8\cdot\log A}{(1-\gamma)^3} \cdot \|u\|_2^2, \tag{B.549}$$

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

$$\left|y^\top \frac{\partial^2 \mathbb{H}^{\pi_\theta}(s)}{\partial \theta^2} y\right| = \left|\left(\frac{y}{\|y\|_2}\right)^\top \frac{\partial^2 \mathbb{H}^{\pi_\theta}(s)}{\partial \theta^2}\left(\frac{y}{\|y\|_2}\right)\right| \cdot \|y\|_2^2 \tag{B.550}$$

$$\leq \max_{\|u\|_2=1}\left|\left\langle \frac{\partial^2 \mathbb{H}^{\pi_\theta}(s)}{\partial \theta^2}u, u\right\rangle\right| \cdot \|y\|_2^2 \tag{B.551}$$

$$= \max_{\|u\|_2=1}\left|\left\langle \frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2}\Big|_{\alpha=0}\frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha}\right\rangle\right| \cdot \|y\|_2^2 \tag{B.552}$$

$$= \max_{\|u\|_2=1}\left|\left\langle \frac{\partial}{\partial \theta_\alpha}\left\{\frac{\partial \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha}\right\}\Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha}\right\rangle\right| \cdot \|y\|_2^2 \tag{B.553}$$

$$= \max_{\|u\|_2=1}\left|\frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \cdot \|y\|_2^2 \tag{B.554}$$

$$\leq \frac{4+8\cdot\log A}{(1-\gamma)^3} \cdot \|y\|_2^2. \qquad \text{(by Eq. (B.542))} \tag{B.555}$$

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

$$\left|\mathbb{H}^{\pi_{\theta'}}(s) - \mathbb{H}^{\pi_\theta}(s) - \left\langle \frac{\partial \mathbb{H}^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta\right\rangle\right| \tag{B.556}$$

$$= \frac{1}{2} \cdot \left|(\theta'-\theta)^\top \frac{\partial^2 \mathbb{H}^{\pi_{\theta_\xi}}(s)}{\partial \theta_\xi^2}(\theta'-\theta)\right| \tag{B.557}$$

$$\leq \frac{2+4\cdot\log A}{(1-\gamma)^3} \cdot \|\theta'-\theta\|_2^2. \qquad \text{(by Eq. (B.550))} \tag{B.558}$$

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

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

$$\left\|\frac{\partial\tilde{V}^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2\geq C(\theta)\cdot\left[\tilde{V}^{\pi_\tau^*}(\rho)-\tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}},\tag{B.559}$$

where

$$C(\theta)=\frac{\sqrt{2\tau}}{\sqrt{S}}\cdot\min_s\sqrt{\mu(s)}\cdot\min_{s,a}\pi_\theta(a|s)\cdot\left\|\frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}}.\tag{B.560}$$

*Proof.* According to the definition of soft value functions,

$$\begin{aligned}
&\tilde{V}^{\pi_\tau^*}(\rho)-\tilde{V}^{\pi_\theta}(\rho) && \text{(B.561)}\\
&=\mathop{\mathbb{E}}_{\substack{s_0\sim\rho,a_t\sim\pi_\tau^*(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}}\left[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t)-\tau\log\pi_\tau^*(a_t|s_t))\right]-\tilde{V}^{\pi_\theta}(\rho) && \text{(B.562)}\\
&=\mathop{\mathbb{E}}_{\substack{s_0\sim\rho,a_t\sim\pi_\tau^*(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}}\Bigg[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t)-\tau\log\pi_\tau^*(a_t|s_t) && \text{(B.563)}\\
&\quad+\tilde{V}^{\pi_\theta}(s_t)-\tilde{V}^{\pi_\theta}(s_t))\Bigg]-\tilde{V}^{\pi_\theta}(\rho) && \text{(B.564)}\\
&=\mathop{\mathbb{E}}_{\substack{s_0\sim\rho,a_t\sim\pi_\tau^*(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}}\Bigg[\sum_{t=0}^{\infty}\gamma^t(r(s_t,a_t)-\tau\log\pi_\tau^*(a_t|s_t) && \text{(B.565)}\\
&\quad+\gamma\tilde{V}^{\pi_\theta}(s_{t+1})-\tilde{V}^{\pi_\theta}(s_t))\Bigg] && \text{(B.566)}\\
&=\frac{1}{1-\gamma}\sum_s d_\rho^{\pi_\tau^*}(s)\cdot\Bigg[\sum_a\pi_\tau^*(a|s)\cdot\Big(r(s,a)-\tau\log\pi_\tau^*(a|s) && \text{(B.567)}\\
&\quad+\gamma\sum_{s'}\mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s')-\tilde{V}^{\pi_\theta}(s)\Big)\Bigg] && \text{(B.568)}\\
&=\frac{1}{1-\gamma}\sum_s d_\rho^{\pi_\tau^*}(s)\cdot\left[\sum_a\pi_\tau^*(a|s)\cdot\left[\tilde{Q}^{\pi_\theta}(s,a)-\tau\log\pi_\tau^*(a|s)\right]-\tilde{V}^{\pi_\theta}(s)\right]. && \text{(B.569)}
\end{aligned}$$

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

$$\bar{\pi}_\theta(a|s)=\frac{\exp\left\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\right\}}{\sum_{a'}\exp\left\{\tilde{Q}^{\pi_\theta}(s,a')/\tau\right\}},\ \forall a.\tag{B.570}$$

We have, $\forall s$,

$$\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s)\right] \tag{B.571}$$

$$\leq \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi(a|s)\right] \tag{B.572}$$

$$= \sum_a \bar{\pi}_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s)\right] \tag{B.573}$$

$$= \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\right\}. \tag{B.574}$$

Also note that,

$$\tilde{V}^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{B.575}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) + \tau \log \bar{\pi}_\theta(a|s) - \tau \log \pi_\theta(a|s)\right] \tag{B.576}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s)\right] - \tau D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \tag{B.577}$$

$$= \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\right\} - \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)). \tag{B.578}$$

Combining the above,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \tag{B.579}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s)\right] - \tilde{V}^{\pi_\theta}(s)\right] \tag{B.580}$$

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\right\} - \tilde{V}^{\pi_\theta}(s)\right] \tag{B.581}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \tag{B.582}$$

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{\tau}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)/\tau - \theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty^2 \tag{B.583}$$

(by Lemma 42) (B.584)

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty^2, \tag{B.585}$$

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

$$\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}} \tag{B.586}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s \frac{d_\rho^{\pi_\tau^*}(s)}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty^2 \right]^{\frac{1}{2}} \tag{B.587}$$

$$= \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s \left( \frac{\sqrt{d_\rho^{\pi_\tau^*}(s)}}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty \right)^2 \right]^{\frac{1}{2}} \tag{B.588}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \sum_s \frac{\sqrt{d_\rho^{\pi_\tau^*}(s)}}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty \tag{B.589}$$

$$\text{(by } \|x\|_2 \leq \|x\|_1\text{)} \tag{B.590}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \tag{B.591}$$

$$\cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty . \tag{B.592}$$

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

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \left[\sum_{s,a} \left( \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \tag{B.593}$$

$$= \left[\sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2^2 \right]^{\frac{1}{2}} \tag{B.594}$$

$$\geq \frac{1}{\sqrt{S}} \sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2, \tag{B.595}$$

$$\text{(by Cauchy-Schwarz, } \|x\|_1 = |\langle \mathbf{1},\ |x| \rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2\text{)} \tag{B.596}$$

which is further lower bounded as

$$\left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left\|H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot)\right]\right\|_2 \tag{B.597}$$

(by Eq. (B.366), Lemma 10) (B.598)

$$= \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \tag{B.599}$$

$$\cdot \left\|H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1}\right]\right\|_2 \tag{B.600}$$

(by Lemma 37) (B.601)

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \tag{B.602}$$

$$\cdot \left\|\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1}\right\|_2 \tag{B.603}$$

(by Lemma 38) (B.604)

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \tag{B.605}$$

$$\cdot \left\|\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1}\right\|_\infty . \tag{B.606}$$

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

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \|\zeta_\theta(s)\|_\infty \tag{B.607}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \tag{B.608}$$

$$\cdot \left[ \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \|\zeta_\theta(s)\|_\infty \right] \tag{B.609}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_{s,a} \pi_\theta(a|s) \tag{B.610}$$

$$\cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} \tag{B.611}$$

$$\geq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}, \tag{B.612}$$

where the last inequality is by $d_\mu^{\pi_\theta}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (B.346)). □

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

*Proof.* The augmented value function $\tilde{V}^{\pi_{\theta_t}}(\rho)$ is monotonically increasing following gradient update due to smoothness, i.e., Lemmas 7 and 14. It follows then that $\tilde{V}^{\pi_{\theta_t}}(\rho)$ is upper bounded. Indeed,

$$\tilde{V}^{\pi_{\theta_t}}(\rho) = \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_{\theta_t}(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[ \sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) - \tau \log \pi_{\theta_t}(a_t|s_t)) \right] \tag{B.613}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_{\theta_t}}(s) \cdot \left[ \sum_a \pi_{\theta_t}(a|s) \cdot (r(s,a) - \tau \log \pi_{\theta_t}(a|s)) \right] \tag{B.614}$$

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_{\theta_t}}(s) \cdot (1 + \tau \log A) \tag{B.615}$$

$$\left( \text{by } r(s,a) \leq 1 \text{ and } -\sum_a \pi_{\theta_t}(a|s) \cdot \log \pi_{\theta_t}(a|s) \leq \log A \right) \tag{B.616}$$

$$\leq \frac{1 + \tau \log A}{1-\gamma}. \tag{B.617}$$

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

$$\mathcal{A}_0(s) = \{a : \pi_{\theta_\infty}(a|s) = 0\}, \tag{B.618}$$

$$\mathcal{A}_+(s) = \{a : \pi_{\theta_\infty}(a|s) > 0\}. \tag{B.619}$$

Note that $\mathcal{A} = \mathcal{A}_0(s) \cup \mathcal{A}_+(s)$ since $\pi_\infty(a|s) \geq 0$, $\forall a \in \mathcal{A}$. We prove that for any state $s \in \mathcal{S}$, $\mathcal{A}_0(s) = \emptyset$ by contradiction. Suppose $\exists s \in \mathcal{S}$, such that $\mathcal{A}_0(s)$ is non-empty. For any $a_0 \in \mathcal{A}_0(s)$, we have $\pi_{\theta_t}(a_0|s) \to \pi_{\theta_\infty}(a_0|s) = 0$, which implies $-\log \pi_{\theta_t}(a_0|s) \to \infty$. There exists $t_0 \geq 1$, such that $\forall t \geq t_0$,

$$-\log \pi_{\theta_t}(a_0|s) \geq \frac{1 + \tau \log A}{\tau(1-\gamma)}. \tag{B.620}$$

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

$$\begin{aligned}
\frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \tilde{A}^{\pi_{\theta_t}}(s, a_0) && \text{(B.621)} \\
&= \frac{d_\mu^{\pi_{\theta_t}}(s)}{1-\gamma} \cdot \pi_{\theta_t}(a_0|s) \cdot \left[\tilde{Q}^{\pi_{\theta_t}}(s, a_0) - \tau \log \pi_{\theta_t}(a_0|s) - \tilde{V}^{\pi_{\theta_t}}(s)\right] && \text{(B.622)} \\
&\geq \frac{d_\mu^{\pi_{\theta_t}}(s)}{1-\gamma} \cdot \pi_{\theta_t}(a_0|s) \cdot \left[0 - \tau \log \pi_{\theta_t}(a_0|s) - \frac{1 + \tau \log A}{1-\gamma}\right] && \text{(B.623)} \\
&\geq \frac{d_\mu^{\pi_{\theta_t}}(s)}{1-\gamma} \cdot \pi_{\theta_t}(a_0|s) \cdot \left[0 + \tau \cdot \frac{1 + \tau \log A}{\tau(1-\gamma)} - \frac{1 + \tau \log A}{1-\gamma}\right] && \text{(B.624)} \\
&= 0, && \text{(B.625)}
\end{aligned}$$

where the first inequality is by

$$\tilde{Q}^{\pi_{\theta_t}}(s, a_0) = r(s, a_0) + \gamma \sum_{s'} \mathcal{P}(s'|s, a_0) \tilde{V}^{\pi_{\theta_t}}(s') \geq 0. \tag{B.626}$$

$$\left(\text{by } r(s, a_0) \geq 0 \text{ and } \tilde{V}^{\pi_{\theta_t}}(s') \geq 0\right) \tag{B.627}$$

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

$$\pi_{\theta_\infty}(a_0|s) = \frac{\exp\{\theta_\infty(a_0|s)\}}{\sum_a \exp\{\theta_\infty(a|s)\}} = 0, \tag{B.628}$$

we have,

$$\sum_a \exp\left\{\theta_\infty(a|s)\right\} = \infty. \tag{B.629}$$

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

$$\pi_{\theta_\infty}(a_+|s) = \frac{\exp\left\{\theta_\infty(a_+|s)\right\}}{\sum_a \exp\left\{\theta_\infty(a|s)\right\}} > 0, \tag{B.630}$$

we have,

$$\exp\left\{\theta_\infty(a_+|s)\right\} = \infty, \quad \forall a_+ \in \mathcal{A}_+(s) \tag{B.631}$$

which implies,

$$\sum_{a_+ \in \mathcal{A}_+(s)} \theta_\infty(a_+|s) = \infty. \tag{B.632}$$

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

$$\sum_a \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} = \sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a_0)} + \sum_{a_+ \in \mathcal{A}_+(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a_+)} \tag{B.633}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \sum_a \pi_{\theta_t}(a|s) \cdot \tilde{A}^{\pi_{\theta_t}}(s,a) \tag{B.634}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \left[\tilde{V}^{\pi_{\theta_t}}(s) - \tilde{V}^{\pi_{\theta_t}}(s)\right] = 0. \tag{B.635}$$

According to Eq. (B.621), $\forall t \geq t_0$,

$$\sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a_0)} \geq 0. \tag{B.636}$$

According to Eq. (B.633), $\forall t \geq t_0$,

$$\sum_{a_+ \in \mathcal{A}_+(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a_+)} = 0 - \sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a_0)} \leq 0. \tag{B.637}$$

which means $\sum_{a_+ \in \mathcal{A}_+(s)} \theta_t(s,a_+)$ will decrease for all large enough $t \geq 1$. This contradicts with Eq. (B.632), i.e., $\sum_{a_+ \in \mathcal{A}_+(s)} \theta_t(s,a_+) \to \infty$.

To this point, we have shown that $\mathcal{A}_0(s) = \emptyset$ for any state $s \in \mathcal{S}$, i.e., $\pi_{\theta_t}(\cdot|s)$ will converge in the interior of probabilistic simplex $\Delta(\mathcal{A})$. Furthermore, at the convergent point $\pi_{\theta_\infty}(\cdot|s)$, the gradient is zero, otherwise by

smoothness the objective can be further improved, which is a contradiction with convergence. According to Lemma 10, $\forall s$,

$$\frac{\partial \tilde{V}^{\pi_{\theta_\infty}}(\mu)}{\partial \theta_\infty(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_\infty}}(s) \cdot H(\pi_{\theta_\infty}(\cdot|s)) \left[\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s)\right] \tag{B.638}$$

$$= \mathbf{0}. \tag{B.639}$$

We have $d_\mu^{\pi_{\theta_\infty}}(s) \geq (1-\gamma) \cdot \mu(s) > 0$ for all states $s$ (cf. Eq. (B.346)). Therefore we have, $\forall s$,

$$H(\pi_{\theta_\infty}(\cdot|s)) \left[\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s)\right] = \mathbf{0}. \tag{B.640}$$

According to Lemma 37, $H(\pi_{\theta_\infty}(\cdot|s))$ has eigenvalue 0 with multiplicity 1, and its corresponding eigenvector is $c \cdot \mathbf{1}$ for some constant $c \in \mathbb{R}$. Therefore, the gradient is zero implies that for all states $s$,

$$\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s) = c \cdot \mathbf{1}, \tag{B.641}$$

which is equivalent to

$$\pi_{\theta_\infty}(\cdot|s) = \text{softmax}(\tilde{Q}^{\pi_{\theta_\infty}}(s,\cdot)/\tau), \tag{B.642}$$

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

$$0 \leq \tilde{Q}^{\pi_{\theta_\infty}}(s,a) \leq \frac{1+\tau \log A}{1-\gamma}, \tag{B.643}$$

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

$$0.9 \cdot \pi_{\theta_\infty}(a|s) \leq \pi_{\theta_t}(a|s) \leq 1.1 \cdot \pi_{\theta_\infty}(a|s), \ \forall (s,a) \tag{B.644}$$

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

$$\inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) = \min \left\{ \min_{1 \leq t \leq t_0} \min_{s,a} \pi_{\theta_t}(a|s), \ \inf_{t \geq t_0} \min_{s,a} \pi_{\theta_t}(a|s) \right\} \tag{B.645}$$

$$= \min\{\Omega(1), \ \Omega(1)\} \in \Omega(1). \qquad \square$$

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

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \left\|\frac{1}{\mu}\right\|_\infty \cdot \frac{1+\tau\log A}{(1-\gamma)^2} \cdot e^{-C(t-1)}. \tag{B.646}$$

*Proof.* According to the soft sub-optimality lemma of Lemma 41,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) = \frac{1}{1-\gamma}\sum_s \left[d_\rho^{\pi_{\theta_t}}(s)\cdot\tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s)\|\pi_\tau^*(\cdot|s))\right] \tag{B.647}$$

$$= \frac{1}{1-\gamma}\sum_s \frac{d_\rho^{\pi_{\theta_t}}(s)}{d_\mu^{\pi_{\theta_t}}(s)}\cdot\left[d_\mu^{\pi_{\theta_t}}(s)\cdot\tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s)\|\pi_\tau^*(\cdot|s))\right] \tag{B.648}$$

$$\leq \frac{1}{(1-\gamma)^2}\sum_s \frac{1}{\mu(s)}\cdot\left[d_\mu^{\pi_{\theta_t}}(s)\cdot\tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s)\|\pi_\tau^*(\cdot|s))\right] \tag{B.649}$$

$$\leq \frac{1}{(1-\gamma)^2}\cdot\left\|\frac{1}{\mu}\right\|_\infty \sum_s \left[d_\mu^{\pi_{\theta_t}}(s)\cdot\tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s)\|\pi_\tau^*(\cdot|s))\right] \tag{B.650}$$

$$= \frac{1}{1-\gamma}\cdot\left\|\frac{1}{\mu}\right\|_\infty \cdot\left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)\right], \tag{B.651}$$

where the last equation is again by Lemma 41, and the first inequality is according to $d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma)\cdot\mu(s)$ (cf. Eq. (B.346)). According to Lemmas 7 and 14, $V^{\pi_\theta}(\mu)$ is $8/(1-\gamma)^3$-smooth, and $\mathbb{H}(\mu,\pi_\theta)$ is $(4+8\log A)/(1-\gamma)^3$-smooth. Therefore, $\tilde{V}^{\pi_\theta}(\mu) = V^{\pi_\theta}(\mu) + \tau\cdot\mathbb{H}(\mu,\pi_\theta)$ is $\beta$-smooth with $\beta = (8+\tau(4+8\log A))/(1-\gamma)^3$. Denote $\tilde{\delta}_t = \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)$. And note $\eta = \frac{(1-\gamma)^3}{8+\tau(4+8\log A)}$. We have,

$$\tilde{\delta}_{t+1} - \tilde{\delta}_t = \tilde{V}^{\pi_{\theta_t}}(\mu) - \tilde{V}^{\pi_{\theta_{t+1}}}(\mu) \tag{B.652}$$

$$\leq -\frac{(1-\gamma)^3}{16+\tau(8+16\log A)}\cdot\left\|\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t}\right\|_2^2 \quad \text{(by Lemma 33)} \tag{B.653}$$

$$\leq -\frac{(1-\gamma)^3}{16+\tau(8+16\log A)}\cdot\frac{2\tau}{S}\cdot\min_s\mu(s)\cdot\min_{s,a}\pi_{\theta_t}(a|s)^2 \tag{B.654}$$

$$\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{d_\mu^{\pi_{\theta_t}}}\right\|_\infty^{-1}\cdot\left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)\right] \quad \text{(by Lemma 15)} \tag{B.655}$$

$$\leq -\frac{(1-\gamma)^4\cdot\min_s\mu(s)}{(8/\tau+4+8\log A)\cdot S}\cdot\min_{s,a}\pi_{\theta_t}(a|s)^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}\cdot\tilde{\delta}_t \tag{B.656}$$

$$\left(\text{by } d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma)\cdot\mu(s)\right) \tag{B.657}$$

$$\leq -\frac{(1-\gamma)^4\cdot\min_s\mu(s)}{(8/\tau+4+8\log A)\cdot S}\cdot\inf_{t\geq1}\min_{s,a}\pi_{\theta_t}(a|s)^2\cdot\left\|\frac{d_\mu^{\pi_\tau^*}}{\mu}\right\|_\infty^{-1}\cdot\tilde{\delta}_t, \tag{B.658}$$

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

$$\tilde{\delta}_t \leq \left[1 - \frac{(1-\gamma)^4 \cdot \min_s \mu(s) \cdot c^2}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1}\right] \cdot \tilde{\delta}_{t-1} \tag{B.659}$$

$$\leq \exp\left\{-\frac{(1-\gamma)^4 \cdot \min_s \mu(s) \cdot c^2}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1}\right\} \cdot \tilde{\delta}_{t-1} \tag{B.660}$$

$$\leq \exp\left\{-\frac{(1-\gamma)^4 \cdot \min_s \mu(s) \cdot c^2}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1)\right\} \cdot \tilde{\delta}_1 \tag{B.661}$$

$$\leq \exp\left\{-\frac{(1-\gamma)^4 \cdot \min_s \mu(s) \cdot c^2}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1)\right\} \cdot \frac{1+\tau\log A}{1-\gamma}, \tag{B.662}$$

where the last inequality is according to Eq. (B.613). Therefore we have the final result,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)\right] \tag{B.663}$$

$$\leq \frac{1}{\exp\{C \cdot (t-1)\}} \cdot \frac{1+\tau\log A}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty, \tag{B.664}$$

where

$$C = \frac{(1-\gamma)^4}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot c^2 \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} > 0, \tag{B.665}$$

is independent with $t$. □

### B.2.4 Proofs for Two-stage and Decaying Entropy Regularization

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(C^2 \cdot t_2), \tag{B.666}$$

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

*Proof.* In particular, using Update 2 with $\eta \leq 1/\tau$ for the following number of iterations,

$$t_1 = \frac{1}{\tau\eta} \cdot K \cdot \exp\left\{4\|\theta_1\|_\infty \sqrt{K}\right\} \tag{B.667}$$

$$\cdot \exp\left\{\frac{1+4\sqrt{K}}{\tau}\right\} \cdot \log\left(\frac{4(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\Delta}\right) + 1 \tag{B.668}$$

$$\in O\left(e^{1/\tau} \cdot \log\left(\frac{\tau+1}{\Delta}\right)\right), \tag{B.669}$$

we have,

$$t_1 - 1 \geq \frac{1}{\tau\eta} \cdot K \cdot \exp\left\{4\|\theta_1\|_\infty \sqrt{K}\right\} \tag{B.670}$$

$$\cdot \exp\left\{\frac{1+4\sqrt{K}}{\tau}\right\} \cdot \log\left(\frac{4(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\Delta}\right) \tag{B.671}$$

$$= \frac{1}{\tau\eta} \cdot K \cdot \exp\{1/\tau\} \cdot \exp\{4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\} \tag{B.672}$$

$$\cdot \log\left(\frac{4(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\Delta}\right) \tag{B.673}$$

$$\geq \frac{1}{\tau\eta} \cdot \frac{1}{c} \cdot \log\left(\frac{4(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\Delta}\right). \quad (c \text{ is from Lemma 13}) \tag{B.674}$$

Therefore we have,

$$\log\left(\frac{4(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\Delta}\right) \leq \tau\eta \cdot c \cdot (t_1 - 1) \tag{B.675}$$

$$\leq \tau\eta \sum_{s=1}^{t_1-1} [\min_a \pi_{\theta_s}(a)] \qquad \text{(by Lemma 13)} \tag{B.676}$$

$$\leq \log\left(\frac{2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\|\zeta_{t_1}\|_2}\right), \qquad \text{(by Lemma 12)} \tag{B.677}$$

which is equivalent to,

$$\|\zeta_{t_1}\|_2 = \left\|\tau\theta_{t_1} - r - \frac{(\tau\theta_{t_1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \leq \frac{\Delta}{2}. \tag{B.678}$$

Then we have, for all $a$,

$$\left|\theta_{t_1}(a) - \frac{r(a)}{\tau} - \frac{(\tau\theta_{t_1} - r)^\top \mathbf{1}}{\tau K}\right| \leq \left\|\theta_{t_1} - \frac{r}{\tau} - \frac{(\tau\theta_{t_1} - r)^\top \mathbf{1}}{\tau K} \cdot \mathbf{1}\right\|_2 \tag{B.679}$$

$$= \frac{1}{\tau} \cdot \left\|\tau\theta_{t_1} - r - \frac{(\tau\theta_{t_1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \leq \frac{\Delta}{2\tau}, \tag{B.680}$$

which implies,

$$\theta_{t_1}(a^*) \geq \frac{r(a^*)}{\tau} - \frac{\Delta}{2\tau} + \frac{(\tau\theta_{t_1} - r)^\top \mathbf{1}}{\tau K}, \quad \text{and} \tag{B.681}$$

$$\theta_{t_1}(a) \leq \frac{r(a)}{\tau} + \frac{\Delta}{2\tau} + \frac{(\tau\theta_{t_1} - r)^\top \mathbf{1}}{\tau K}. \quad \text{for all } a \neq a^* \tag{B.682}$$

Then we have, for all $a \neq a^*$,

$$\theta_{t_1}(a^*) - \theta_{t_1}(a) \geq \frac{r(a^*)}{\tau} - \frac{\Delta}{2\tau} - \left(\frac{r(a)}{\tau} + \frac{\Delta}{2\tau}\right) \tag{B.683}$$

$$= \frac{r(a^*)}{\tau} - \frac{r(a)}{\tau} - \frac{\Delta}{\tau} \geq 0, \tag{B.684}$$

which means $\pi_{\theta_{t_1}}(a^*) \geq \pi_{\theta_{t_1}}(a)$. Now we turn off the regularization and use Update 1 for $t_2 \geq 1$ iterations. According to similar arguments as in Theorem 3, we have,

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(C^2 \cdot t_2), \tag{B.685}$$

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

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{K}{t^{1/\alpha}} + \frac{\log t}{\exp\left\{\sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}} \cdot \frac{2(\tau_1 \|\theta_1\|_\infty + 1)\sqrt{K}}{\alpha \cdot \Delta}. \tag{B.686}$$

*Proof.* Denote $\pi^*_{\tau_t} = \text{softmax}(r/\tau_t)$ as the softmax optimal policy at time $t$. We have,

$$(\pi^* - \pi_{\theta_t})^\top r = \underbrace{(\pi^* - \pi^*_{\tau_t})^\top r}_{\text{“decaying”}} + \underbrace{(\pi^*_{\tau_t} - \pi_{\theta_t})^\top r}_{\text{“tracking”}}. \tag{B.687}$$

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

$$(\pi^* - \pi^*_{\tau_t})^\top r = \sum_a \pi^*_{\tau_t}(a) \cdot r(a^*) - \sum_a \pi^*_{\tau_t}(a) \cdot r(a) \tag{B.688}$$

$$= \sum_{a \neq a^*} \pi^*_{\tau_t}(a) \cdot \Delta(a) \tag{B.689}$$

$$= \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau_t}} \cdot \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau_t}}} \tag{B.690}$$

$$\leq \sum_{a \neq a^*} \frac{e^{\frac{r(a)}{\tau_t}} \cdot \Delta(a)}{e^{\frac{r(a^*)}{\tau_t}} + e^{\frac{r(a)}{\tau_t}}} \tag{B.691}$$

$$= \sum_{a \neq a^*} \frac{\Delta(a)}{e^{\frac{\Delta(a)}{\tau_t}} + 1} \tag{B.692}$$

$$\leq \sum_{a \neq a^*} \frac{1}{e^{\frac{\Delta}{\tau_t}} + 1} = \frac{K-1}{1 + e^{\frac{\Delta}{\tau_t}}} \leq \frac{K}{e^{\frac{\Delta}{\tau_t}}}. \tag{B.693}$$

Using the decaying temperature $\tau_t = \frac{\alpha \cdot \Delta}{\log t}$, for $t \geq 2$, where $\alpha > 0$, we have,

$$(\pi^* - \pi^*_{\tau_t})^\top r \leq \frac{K}{t^{1/\alpha}}. \tag{B.694}$$

**"tracking" part.** Using Update 3, we have,

$$\tau_{t+1}\theta_{t+1} - r - \frac{(\tau_{t+1}\theta_{t+1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} = \tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{B.695}$$

$$+ (\tau_{t+1}\theta_{t+1} - \tau_t\theta_t) + \left( \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} - \frac{(\tau_{t+1}\theta_{t+1} - r)^\top \mathbf{1}}{K} \right) \cdot \mathbf{1} \tag{B.696}$$

$$= \tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau_t\eta_t \cdot H(\pi_{\theta_t})(r - \tau_t \log \pi_{\theta_t}) \tag{B.697}$$

$$+ \frac{(\tau_t\theta_t - \tau_{t+1}\theta_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1} \qquad \text{(by Update 3)} \tag{B.698}$$

$$= (\mathbf{Id} - \tau_t\eta_t \cdot H(\pi_{\theta_t})) \left( \tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \tag{B.699}$$

$$\left( H(\pi_{\theta_t})\mathbf{1} = H(\pi_{\theta_t})^\top \mathbf{1} = \mathbf{0}, \text{ cf. Eq. (B.401)} \right) \tag{B.700}$$

$$= (\mathbf{Id} - H(\pi_{\theta_t})) \left( \tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right). \quad (\eta_t = 1/\tau_t) \tag{B.701}$$

Therefore we have,

$$\left\|\tau_{t+1}\theta_{t+1} - r - \frac{(\tau_{t+1}\theta_{t+1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \tag{B.702}$$

$$= \left\|(\mathbf{Id} - H(\pi_{\theta_t}))\left(\tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right)\right\|_2 \tag{B.703}$$

$$\leq \left(1 - \min_a \pi_{\theta_t}(a)\right) \cdot \left\|\tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \tag{B.704}$$

(by Lemma 38) (B.705)

$$\leq \exp\left\{-\min_a \pi_{\theta_t}(a)\right\} \cdot \left\|\tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2. \tag{B.706}$$

Then we have,

$$(\pi^*_{\tau_t} - \pi_{\theta_t})^\top r \leq \left\|\pi^*_{\tau_t} - \pi_{\theta_t}\right\|_1 \tag{B.707}$$

(by Hölder's inequality, and $\|r\|_\infty \leq 1$) (B.708)

$$\leq \left\|\theta_t - \frac{r}{\tau_t} - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{\tau_t K} \cdot \mathbf{1}\right\|_\infty \qquad \text{(by Lemma 39)} \tag{B.709}$$

$$\leq \frac{1}{\tau_t} \cdot \left\|\tau_t\theta_t - r - \frac{(\tau_t\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \qquad (\|x\|_\infty \leq \|x\|_2) \tag{B.710}$$

$$\leq \frac{1}{\tau_t} \cdot \exp\left\{-\min_a \pi_{\theta_{t-1}}(a)\right\} \tag{B.711}$$

$$\cdot \left\|\tau_{t-1}\theta_{t-1} - r - \frac{(\tau_{t-1}\theta_{t-1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \tag{B.712}$$

(by Eq. (B.702)) (B.713)

$$\leq \frac{1}{\tau_t} \cdot \exp\left\{-\sum_{s=1}^{t-1}[\min_a \pi_{\theta_s}(a)]\right\} \tag{B.714}$$

$$\cdot \left\|\tau_1\theta_1 - r - \frac{(\tau_1\theta_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \tag{B.715}$$

$$\leq \frac{1}{\tau_t} \cdot \exp\left\{-\sum_{s=1}^{t-1}[\min_a \pi_{\theta_s}(a)]\right\} \cdot 2(\tau_1\|\theta_1\|_\infty + 1)\sqrt{K} \tag{B.716}$$

(by Eq. (B.413)) (B.717)

$$= \frac{\log t}{\exp\left\{\sum_{s=1}^{t-1}[\min_a \pi_{\theta_s}(a)]\right\}} \cdot \frac{2(\tau_1\|\theta_1\|_\infty + 1)\sqrt{K}}{\alpha \cdot \Delta}. \qquad \square$$

## B.3 Proofs for Section 2.5: A Theoretical Understanding of Entropy Regularization in Policy Gradient

### B.3.1 Proofs for the Bandit Case

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \leq \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r. \tag{B.718}$$

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

$$\begin{aligned}
(\pi^* - \pi_\theta)^\top r &= \sum_a \pi_\theta(a) \cdot r(a^*) - \sum_a \pi_\theta(a) \cdot r(a) && \text{(B.719)} \\
&= \sum_{a \neq a^*} \pi_\theta(a) \cdot r(a^*) - \sum_{a \neq a^*} \pi_\theta(a) \cdot r(a) && \text{(B.720)} \\
&= \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta(a) && \text{(B.721)} \\
&\geq \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta. && \text{(B.722)}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
0 \leq r(a^*) - \pi_\theta^\top r &= (\pi^* - \pi_\theta)^\top r && \text{(B.723)} \\
&= \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta(a) \leq \sum_{a \neq a^*} \pi_\theta(a) \cdot 1 = \sum_{a \neq a^*} \pi_\theta(a). && \text{(B.724)}
\end{aligned}$$

Therefore the $\ell_2$ norm of gradient can be upper bounded as

$$\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{B.725}$$

$$= \left(\pi_\theta(a^*)^2 \cdot \left[r(a^*) - \pi_\theta^\top r\right]^2 + \sum_{a\neq a^*} \pi_\theta(a)^2 \cdot (r(a) - \pi_\theta^\top r)^2\right)^{\frac{1}{2}} \tag{B.726}$$

$$\leq \left(1^2 \cdot \left[\sum_{a\neq a^*} \pi_\theta(a)\right]^2 + \sum_{a\neq a^*} \pi_\theta(a)^2 \cdot 1^2\right)^{\frac{1}{2}} \tag{B.727}$$

$$\leq \left(\left[\sum_{a\neq a^*} \pi_\theta(a)\right]^2 + \left[\sum_{a\neq a^*} \pi_\theta(a)\right]^2\right)^{\frac{1}{2}} \qquad (\text{by } \|x\|_2 \leq \|x\|_1) \tag{B.728}$$

$$= \sqrt{2} \cdot \sum_{a\neq a^*} \pi_\theta(a). \tag{B.729}$$

Combining the results, we have

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

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

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

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

$$\delta_t - \delta_{t+1} \tag{B.730}$$

$$= (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{B.731}$$

$$\leq \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \qquad (\text{by Lemma 2}) \tag{B.732}$$

$$= \left(\frac{5\eta_t^2}{4} + \eta_t\right) \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 \qquad \left(\text{by } \theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \tag{B.733}$$

$$\leq \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2. \qquad (\text{by } \eta_t \in (0,1] \text{ and by Lemma 17}) \tag{B.734}$$

According to convergence result Theorem 2 we have $\delta_t > 0$, $\delta_t \to 0$ as $t \to \infty$. We prove that for all large enough $t \geq 1$, $\delta_t \leq \frac{10}{9} \cdot \delta_{t+1}$ by contradiction.

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

$$\delta_{t+1} \geq \delta_t - \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2 \tag{B.735}$$

$$> \frac{10}{9} \cdot \delta_{t+1} - \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \left(\frac{10}{9} \cdot \delta_{t+1}\right)^2 \tag{B.736}$$

$$\left(\text{since } f(x) = x - ax^2 \text{ is increasing for all } x < \frac{1}{2a} \text{ and } a > 0\right) \tag{B.737}$$

$$= \frac{10}{9} \cdot \delta_{t+1} - \frac{50}{9} \cdot \frac{1}{\Delta^2} \cdot \delta_{t+1}^2, \tag{B.738}$$

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

$$\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \leq \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \frac{\delta_t}{\delta_{t+1}} \leq \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \frac{10}{9} = \frac{5}{\Delta^2}. \tag{B.739}$$

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

$$\frac{1}{\delta_{T_1+t}} - \frac{1}{\delta_{T_1}} \leq \frac{5}{\Delta^2} \cdot (t-1) \leq \frac{5}{\Delta^2} \cdot t. \tag{B.740}$$

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

$$(\pi^* - \pi_{\theta_{T_1+t}})^\top r = \delta_{T_1+t} \geq \frac{1}{\frac{1}{\delta_{T_1}} + \frac{5}{\Delta^2} \cdot t} \tag{B.741}$$

$$\geq \frac{1}{C + \frac{5}{\Delta^2} \cdot t} \geq \frac{1}{C + \frac{5}{\Delta^2} \cdot (T_1 + t)}. \tag{B.742}$$

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

$$(\pi^* - \pi_{\theta_t})^\top r \geq \frac{1}{C + \frac{5}{\Delta^2} \cdot t} \geq \frac{1}{\frac{t}{\Delta^2} + \frac{5}{\Delta^2} \cdot t} = \frac{\Delta^2}{6 \cdot t}, \tag{B.743}$$

for all large enough $t \geq 1$. □

### B.3.2 Proofs for General MDPs

**Lemma 18** (Reversed Łojasiewicz)**.** Denote

$$\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) > 0, \tag{B.744}$$

as the optimal value gap of state $s$, where $a^*(s)$ is the action that the optimal policy selects under state $s$, and $\Delta^* = \min_{s\in\mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. Then we have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \leq \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \left[ V^*(\mu) - V^{\pi_\theta}(\mu) \right]. \tag{B.745}$$

*Proof.* Denote $\Delta^*(s,a) = Q^*(s,a^*(s)) - Q^*(s,a)$, and $\Delta^*(s) = \min_{a\neq a^*(s)} \Delta^*(s,a)$. We have,

$$V^*(\mu) - V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s,a) \tag{B.746}$$

$$\text{(by Lemma 36)} \tag{B.747}$$

$$= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \Bigg[ \sum_a \pi_\theta(a|s) \cdot Q^*(s,a^*(s)) \tag{B.748}$$

$$- \sum_a \pi_\theta(a|s) \cdot Q^*(s,a) \Bigg] \tag{B.749}$$

$$= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \Bigg[ \sum_{a\neq a^*(s)} \pi_\theta(a|s) \cdot Q^*(s,a^*(s)) \tag{B.750}$$

$$- \sum_{a\neq a^*(s)} \pi_\theta(a|s) \cdot Q^*(s,a) \Bigg] \tag{B.751}$$

$$= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[ \sum_{a\neq a^*(s)} \pi_\theta(a|s) \cdot \Delta^*(s,a) \right] \tag{B.752}$$

$$\geq \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[ \sum_{a\neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^*(s). \tag{B.753}$$

Since $Q^{\pi_\theta}(s,a) \in [0, 1/(1-\gamma)]$, and $V^{\pi_\theta}(s) \in [0, 1/(1-\gamma)]$, we have $|A^{\pi_\theta}(s,a)| \in [0, 1/(1-\gamma)]$. Also,

$$|A^{\pi_\theta}(s,a^*(s))| = \left| Q^{\pi_\theta}(s,a^*(s)) - \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right| \tag{B.754}$$

$$= \left| \sum_{a\neq a^*(s)} \pi_\theta(a|s) \cdot \left[ Q^{\pi_\theta}(s,a^*(s)) - Q^{\pi_\theta}(s,a) \right] \right| \tag{B.755}$$

$$\leq \sum_{a\neq a^*(s)} \pi_\theta(a|s) \cdot \left| Q^{\pi_\theta}(s,a^*(s)) - Q^{\pi_\theta}(s,a) \right| \tag{B.756}$$

$$\text{(by the triangle inequality)} \tag{B.757}$$

$$\leq \frac{1}{1-\gamma} \sum_{a\neq a^*(s)} \pi_\theta(a|s). \qquad \text{(because } Q^{\pi_\theta}(s,a) \in [0, 1/(1-\gamma)]\text{)} \tag{B.758}$$

Therefore the $\ell_2$ norm of gradient can be upper bounded as

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 = \frac{1}{1-\gamma}\cdot\left[\sum_s d_\mu^{\pi_\theta}(s)^2 \sum_a \pi_\theta(a|s)^2\cdot A^{\pi_\theta}(s,a)^2\right]^{\frac{1}{2}} \tag{B.759}$$

$$= \frac{1}{1-\gamma}\cdot\Bigg[\sum_s d_\mu^{\pi_\theta}(s)^2\cdot\Big(\pi_\theta(a^*(s)|s)^2\cdot A^{\pi_\theta}(s,a^*(s))^2 \tag{B.760}$$

$$+\sum_{a\neq a^*(s)}\pi_\theta(a|s)^2\cdot A^{\pi_\theta}(s,a)^2\Big)\Bigg]^{\frac{1}{2}} \tag{B.761}$$

$$\leq \frac{1}{1-\gamma}\cdot\Bigg[\sum_s d_\mu^{\pi_\theta}(s)^2\cdot\Big(1\cdot\frac{1}{(1-\gamma)^2}\cdot\Big[\sum_{a\neq a^*(s)}\pi_\theta(a|s)\Big]^2 \tag{B.762}$$

$$+\sum_{a\neq a^*(s)}\pi_\theta(a|s)^2\cdot\frac{1}{(1-\gamma)^2}\Big)\Bigg]^{\frac{1}{2}} \tag{B.763}$$

$$\leq \frac{1}{(1-\gamma)^2}\cdot\left[\sum_s d_\mu^{\pi_\theta}(s)^2\cdot 2\cdot\Big[\sum_{a\neq a^*(s)}\pi_\theta(a|s)\Big]^2\right]^{\frac{1}{2}} \tag{B.764}$$

$$\text{(by } \|x\|_2\leq\|x\|_1\text{)} \tag{B.765}$$

$$\leq \frac{1}{(1-\gamma)^2}\cdot\sqrt{2}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\Big[\sum_{a\neq a^*(s)}\pi_\theta(a|s)\Big]. \tag{B.766}$$

$$\text{(by } \|x\|_2\leq\|x\|_1\text{)} \tag{B.767}$$

Combining the results, we have

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \leq \frac{1}{1-\gamma}\cdot\sqrt{2}\cdot\frac{1}{1-\gamma}\sum_s d_\mu^{\pi_\theta}(s)\cdot\left[\sum_{a\neq a^*(s)}\pi_\theta(a|s)\right] \tag{B.768}$$

$$= \frac{1}{1-\gamma}\cdot\frac{\sqrt{2}}{\Delta^*}\cdot\frac{1}{1-\gamma}\sum_s d_\mu^{\pi_\theta}(s)\cdot\left[\sum_{a\neq a^*(s)}\pi_\theta(a|s)\right]\cdot\Delta^* \tag{B.769}$$

$$\leq \frac{1}{1-\gamma}\cdot\frac{\sqrt{2}}{\Delta^*}\cdot\frac{1}{1-\gamma}\sum_s d_\mu^{\pi_\theta}(s)\cdot\left[\sum_{a\neq a^*(s)}\pi_\theta(a|s)\right]\cdot\Delta^*(s) \tag{B.770}$$

$$\text{(by } \Delta^*\leq\Delta^*(s) \text{ holds for all } s\text{)} \tag{B.771}$$

$$\leq \frac{1}{1-\gamma}\cdot\frac{\sqrt{2}}{\Delta^*}\cdot\left[V^*(\mu)-V^{\pi_\theta}(\mu)\right]. \qquad \square$$

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

$$V^*(\mu)-V^{\pi_{\theta_t}}(\mu)\geq\frac{(1-\gamma)^5\cdot(\Delta^*)^2}{12\cdot t}, \tag{B.772}$$

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

*Proof.* Suppose Algorithm 1 can converge faster than $O(1/t)$ for general MDPs, then it can converge faster than $O(1/t)$ for any one-state MDPs, which are special cases of general MDPs. This is a contradiction with Theorem 9.

The above one-sentence argument implies a $\Omega(1/t)$ rate lower bound. To calculate the constant in the lower bound, we need results similar to Lemma 17. According to the reversed Łojasiewicz inequality of Lemma 18,

$$\left\|\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}\right\|_2 \leq \frac{1}{1-\gamma}\cdot\frac{\sqrt{2}}{\Delta^*}\cdot\delta_t, \tag{B.773}$$

where $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu) > 0$. Let $\theta_{t+1} = \theta_t + \eta_t \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}$, and $\pi_{\theta_{t+1}}(\cdot|s) = \text{softmax}(\theta_{t+1}(s,\cdot))$, $\forall s\in\mathcal{S}$ be the next policy after one step gradient update. Using similar calculations as in Eq. (B.730),

$$\begin{aligned}
\delta_t - \delta_{t+1} &= V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t}, \theta_{t+1}-\theta_t\right\rangle && \text{(B.774)}\\
&\quad + \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t}, \theta_{t+1}-\theta_t\right\rangle && \text{(B.775)}\\
&\leq \frac{4}{(1-\gamma)^3}\cdot\|\theta_{t+1}-\theta_t\|_2^2 + \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t}, \theta_{t+1}-\theta_t\right\rangle && \text{(B.776)}\\
&\quad \text{(by Lemma 7)} && \text{(B.777)}\\
&= \left(\frac{4\eta_t^2}{(1-\gamma)^3}+\eta_t\right)\cdot\left\|\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t}\right\|_2^2 && \text{(B.778)}\\
&\quad \left(\text{by } \theta_{t+1} = \theta_t + \eta_t\cdot\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t}\right) && \text{(B.779)}\\
&\leq \frac{10}{(1-\gamma)^5}\cdot\frac{1}{(\Delta^*)^2}\cdot\delta_t^2. \quad \text{(by } \eta_t\in(0,1] \text{ and by Lemma 18)} && \text{(B.780)}
\end{aligned}$$

According to Theorem 4, we have $\delta_t > 0$, $\delta_t \to 0$ as $t\to\infty$. Using similar arguments as in Eq. (B.735), we can show that for all large enough $t \geq 1$, $\delta_t \leq \frac{11}{10}\cdot\delta_{t+1}$. Divide both sides of $\delta_t - \delta_{t+1} \leq \frac{10}{(1-\gamma)^5}\cdot\frac{1}{(\Delta^*)^2}\cdot\delta_t^2$ by $\delta_t\cdot\delta_{t+1}$,

$$\begin{aligned}
\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} &\leq \frac{10}{(1-\gamma)^5}\cdot\frac{1}{(\Delta^*)^2}\cdot\frac{\delta_t}{\delta_{t+1}} && \text{(B.781)}\\
&\leq \frac{10}{(1-\gamma)^5}\cdot\frac{1}{(\Delta^*)^2}\cdot\frac{11}{10} = \frac{11}{(1-\gamma)^5\cdot(\Delta^*)^2}. && \text{(B.782)}
\end{aligned}$$

Using similar calculations as in the proof of Theorem 9, we have,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) = \delta_t \geq \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12 \cdot t}, \tag{B.783}$$

for all large enough $t \geq 1$. □

### B.3.3 Proofs for the Non-uniform Łojasiewicz Degree

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

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq C(\theta) \cdot \left[ (\pi^* - \pi_\theta)^\top r \right]^{1-\xi}. \tag{B.784}$$

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

$$(\pi^* - \pi_\theta)^\top r = r(a^*) - \pi_\theta^\top r = 0.6 - (0.6 - 0.8\epsilon) = 0.8 \cdot \epsilon. \tag{B.785}$$

According to the reversed Łojasiewicz inequality of Lemma 17,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \leq \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r = \frac{\sqrt{2}}{2} \cdot (\pi^* - \pi_\theta)^\top r \tag{B.786}$$

$$\leq \frac{1.5}{2} \cdot (\pi^* - \pi_\theta)^\top r = 0.6 \cdot \epsilon. \tag{B.787}$$

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \leq 0.6 \cdot \epsilon = \frac{1}{4} \cdot 3 \cdot 0.8 \cdot \epsilon \tag{B.788}$$

$$< \pi_\theta(a^*) \cdot 3 \cdot 0.8 \cdot \epsilon = C(\theta) \cdot 3 \cdot 0.8 \cdot \epsilon. \tag{B.789}$$

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 < C(\theta) \cdot 3 \cdot 0.8 \cdot \epsilon = C(\theta) \cdot 3 \cdot (0.8 \cdot \epsilon)^\xi \cdot (0.8 \cdot \epsilon)^{1-\xi} \tag{B.790}$$

$$< C(\theta) \cdot (0.8 \cdot \epsilon)^{1-\xi} = C(\theta) \cdot \left[ (\pi^* - \pi_\theta)^\top r \right]^{1-\xi}, \tag{B.791}$$

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

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

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

$$\begin{aligned}
\delta_\theta &= \mathop{\mathbb{E}}_{a\sim\pi_\tau^*}[r(a) - \tau \log \pi_\tau^*(a)] - \mathop{\mathbb{E}}_{a\sim\pi_\theta}[r(a) - \tau \log \pi_\tau^*(a)] && \text{(B.792)}\\
&\quad - \mathop{\mathbb{E}}_{a\sim\pi_\theta}[\tau \log \pi_\tau^*(a) - \tau \log \pi_\theta(a)] && \text{(B.793)}\\
&= \tau \log \sum_a \exp\{r(a)/\tau\} - \tau \log \sum_a \exp\{r(a)/\tau\} && \text{(B.794)}\\
&\quad + \tau \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_\tau^*) \qquad (\text{since } \pi_\tau^* = \mathrm{softmax}(r/\tau)) && \text{(B.795)}\\
&= \tau \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_\tau^*) && \text{(B.796)}\\
&\le \frac{\tau}{2} \cdot \left\| \frac{r}{\tau} - \theta - \frac{(r/\tau - \theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 \qquad (\text{by Lemma 42}) && \text{(B.797)}\\
&= \frac{1}{2\tau} \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2. && \text{(B.798)}
\end{aligned}$$

Next, the entropy regularized policy gradient w.r.t. $\theta$ is

$$\begin{aligned}
\frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} &= H(\pi_\theta)(r - \tau \log \pi_\theta) && \text{(B.799)}\\
&= H(\pi_\theta) \left( r - \tau\theta + \tau \log \sum_a \exp\{\theta(a)\} \cdot \mathbf{1} \right) && \text{(B.800)}\\
&= H(\pi_\theta)\,(r - \tau\theta) && \text{(B.801)}\\
&= H(\pi_\theta) \left( r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right), && \text{(B.802)}
\end{aligned}$$

where the last two equations are by $H(\pi_\theta)\mathbf{1} = \mathbf{0}$ as shown in Lemma 37. Then we have,

$$\begin{aligned}
\left\| \frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} \right\|_2 &= \left\| H(\pi_\theta) \left( r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 && \text{(B.803)}\\
&\ge \min_a \pi_\theta(a) \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 \qquad (\text{by Lemma 38}) && \text{(B.804)}\\
&\ge \min_a \pi_\theta(a) \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty && \text{(B.805)}\\
&\ge \min_a \pi_\theta(a) \cdot \sqrt{2\tau} \cdot \sqrt{\delta_\theta} \qquad (\text{by Eq. (B.792)}) && \text{(B.806)}\\
&= \sqrt{2\tau} \cdot \min_a \pi_\theta(a) \cdot \left( \mathop{\mathbb{E}}_{a\sim\pi_\tau^*}[r(a) - \tau \log \pi_\tau^*(a)] - \mathop{\mathbb{E}}_{a\sim\pi_\theta}[r(a) - \tau \log \pi_\theta(a)] \right)^{\frac{1}{2}}, && \text{(B.807)}
\end{aligned}$$

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

## B.4 Miscellaneous Extra Supporting Results

**Lemma 33** (Ascent lemma for smooth function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\beta$-smooth function, $\theta \in \mathbb{R}^d$ and $\theta' = \theta + \frac{1}{\beta} \cdot \frac{\partial f(\theta)}{\partial \theta}$. We have,*

$$f(\theta) - f(\theta') \leq -\frac{1}{2\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \tag{B.808}$$

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

$$\left| f(\theta') - f(\theta) - \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2, \tag{B.809}$$

which implies,

$$\begin{aligned} f(\theta) - f(\theta') &\leq -\left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle + \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2 && \text{(B.810)} \\ &= -\frac{1}{\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 + \frac{\beta}{2} \cdot \frac{1}{\beta^2} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 && \text{(B.811)} \\ &\qquad \left( \theta' = \theta + \frac{1}{\beta} \cdot \frac{\partial f(\theta)}{\partial \theta} \right) && \text{(B.812)} \\ &= -\frac{1}{2\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \end{aligned}$$ □

**Lemma 34** (First performance difference lemma (Kakade and Langford, 2002))**.** *For any policies $\pi$ and $\pi'$,*

$$\begin{aligned} V^{\pi'}(\rho) - V^{\pi}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi'}(s) \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^\pi(s,a) && \text{(B.813)} \\ &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi'}(s) \sum_a \pi'(a|s) \cdot A^\pi(s,a). && \text{(B.814)} \end{aligned}$$

*Proof.* According to the definition of value function,

$$V^{\pi'}(s) - V^{\pi}(s) = \sum_a \pi'(a|s) \cdot Q^{\pi'}(s,a) - \sum_a \pi(a|s) \cdot Q^{\pi}(s,a) \tag{B.815}$$

$$= \sum_a \pi'(a|s) \cdot \Big( Q^{\pi'}(s,a) - Q^{\pi}(s,a) \Big) \tag{B.816}$$

$$+ \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi}(s,a) \tag{B.817}$$

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi}(s,a) \tag{B.818}$$

$$+ \gamma \sum_a \pi'(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \Big[ V^{\pi'}(s') - V^{\pi}(s') \Big] \tag{B.819}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} (\pi'(a'|s') - \pi(a'|s')) \cdot Q^{\pi}(s',a') \tag{B.820}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \pi'(a'|s') \cdot (Q^{\pi}(s',a') - V^{\pi}(s')) \tag{B.821}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \pi'(a'|s') \cdot A^{\pi}(s',a'). \qquad \square$$

**Lemma 35** (Second performance difference lemma)**.** *For any policies $\pi$ and $\pi'$,*

$$V^{\pi'}(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi}(s) \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi'}(s,a). \tag{B.822}$$

*Proof.* According to the definition of value function,

$$V^{\pi'}(s) - V^{\pi}(s) = \sum_a \pi'(a|s) \cdot Q^{\pi'}(s,a) - \sum_a \pi(a|s) \cdot Q^{\pi}(s,a) \tag{B.823}$$

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi'}(s,a) \tag{B.824}$$

$$+ \sum_a \pi(a|s) \cdot \Big( Q^{\pi'}(s,a) - Q^{\pi}(s,a) \Big) \tag{B.825}$$

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot Q^{\pi'}(s,a) \tag{B.826}$$

$$+ \gamma \sum_a \pi(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \Big[ V^{\pi'}(s') - V^{\pi}(s') \Big] \tag{B.827}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi}(s') \sum_{a'} (\pi'(a'|s') - \pi(a'|s')) \cdot Q^{\pi'}(s',a'). \qquad \square$$

**Lemma 36** (Value sub-optimality lemma)**.** *For any policy $\pi$,*

$$V^{*}(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi}(s) \sum_a (\pi^*(a|s) - \pi(a|s)) \cdot Q^{*}(s,a). \tag{B.828}$$

*Proof.* According to the second performance difference lemma of Lemma 35, the result immediately holds. □

**Lemma 37** (Spectrum of H matrix)**.** *Let* $\pi \in \Delta(\mathcal{A})$*. Denote* $H(\pi) = diag(\pi) - \pi\pi^\top$*. Let*

$$\pi(1) \leq \pi(2) \leq \cdots \leq \pi(K). \tag{B.829}$$

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

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_K. \tag{B.830}$$

*Then we have,*

$$\lambda_1 = 0, \tag{B.831}$$

$$\pi(i-1) \leq \lambda_i \leq \pi(i),\ i = 2, 3, \ldots, K. \tag{B.832}$$

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

$$\pi(1) - \pi^\top\pi \leq \lambda_1 \leq \pi(1), \tag{B.833}$$

$$\pi(i-1) \leq \lambda_i \leq \pi(i),\ i = 2, 3, \ldots, K. \tag{B.834}$$

We show $\lambda_1 = 0$. Note

$$H(\pi)\mathbf{1} = (\mathrm{diag}(\pi) - \pi\pi^\top)\mathbf{1} = \pi - \pi = 0 \cdot \mathbf{1}. \tag{B.835}$$

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

$$x^\top H(\pi)x = \mathop{\mathbb{E}}_{a\sim\pi}[x(a)^2] - \left(\mathop{\mathbb{E}}_{a\sim\pi}[x(a)]\right)^2 = \mathrm{Var}_{a\sim\pi}[x(a)] \geq 0, \tag{B.836}$$

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

**Lemma 38.** *Let* $\pi \in \Delta(\mathcal{A})$*. Denote* $H(\pi) = diag(\pi) - \pi\pi^\top$*. For any vector* $x \in \mathbb{R}^K$*,*

$$\left\|(\mathbf{Id} - H(\pi))\left(x - \frac{x^\top\mathbf{1}}{K}\cdot\mathbf{1}\right)\right\|_2 \leq \left(1 - \min_a \pi(a)\right)\cdot\left\|x - \frac{x^\top\mathbf{1}}{K}\cdot\mathbf{1}\right\|_2, \tag{B.837}$$

$$\left\|H(\pi)\left(x - \frac{x^\top\mathbf{1}}{K}\cdot\mathbf{1}\right)\right\|_2 \geq \min_a \pi(a)\cdot\left\|x - \frac{x^\top\mathbf{1}}{K}\cdot\mathbf{1}\right\|_2. \tag{B.838}$$

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

$$x = a_1 \cdot \frac{\mathbf{1}}{\sqrt{K}} + a_2 v_2 + \cdots + a_K v_K \tag{B.839}$$

$$= \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} + a_2 v_2 + \cdots + a_K v_K. \tag{B.840}$$

Since $H(\pi)$ 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

$$a_1 = x^\top \frac{\mathbf{1}}{\sqrt{K}} = \frac{x^\top \mathbf{1}}{\sqrt{K}}. \tag{B.841}$$

Denote

$$x' = x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} = a_2 v_2 + \cdots + a_K v_K. \tag{B.842}$$

We have

$$\|x'\|_2^2 = a_2^2 + \cdots + a_K^2. \tag{B.843}$$

On the other hand,

$$(\mathbf{Id} - H(\pi))x' = a_2(1-\lambda_2)v_2 + \cdots + a_K(1-\lambda_K)v_K. \tag{B.844}$$

Therefore

$$\|(\mathbf{Id} - H(\pi))x'\|_2 = \left(a_2^2(1-\lambda_2)^2 + \cdots + a_K^2(1-\lambda_K)^2\right)^{\frac{1}{2}} \tag{B.845}$$

$$\leq \left((a_2^2 + \cdots + a_K^2) \cdot (1-\lambda_2)^2\right)^{\frac{1}{2}} \tag{B.846}$$

$$= (1-\lambda_2) \cdot \|x'\|_2 \tag{B.847}$$

$$\leq \left(1 - \min_a \pi(a)\right) \cdot \|x'\|_2, \tag{B.848}$$

where the first inequality is by $0 \leq \pi(1) \leq \lambda_2 \leq \cdots \leq \lambda_K \leq \pi(K) \leq 1$, and the last inequality is according to $\lambda_2 \geq \pi(1) = \min_a \pi(a)$, and both are shown in Lemma 37. Similarly,

$$\|H(\pi)x'\|_2 = \left(a_2^2\lambda_2^2 + \cdots + a_K^2\lambda_K^2\right)^{\frac{1}{2}} \tag{B.849}$$

$$\geq \left((a_2^2 + \cdots + a_K^2) \cdot \lambda_2^2\right)^{\frac{1}{2}} \tag{B.850}$$

$$= \lambda_2 \cdot \|x'\|_2 \tag{B.851}$$

$$\geq \min_a \pi(a) \cdot \|x'\|_2. \qquad \square$$

**Lemma 39.** *Let* $\pi_\theta = \mathrm{softmax}(\theta)$ *and* $\pi_{\theta'} = \mathrm{softmax}(\theta')$*. Then for any constant* $c \in \mathbb{R}$*,*

$$\|\pi_\theta - \pi'_\theta\|_1 \leq \|\theta' - \theta - c \cdot \mathbf{1}\|_\infty . \tag{B.852}$$

*Proof.* This results improves the results of $\|\pi_\theta - \pi_{\theta'}\|_\infty \leq 2 \cdot \|\theta - \theta'\|_\infty$ in Xiao et al. (2019, Lemma 5). According to the $\ell_1$ norm strong convexity of negative entropy over probabilistic simplex, i.e., for any policies $\pi$, $\pi'$,

$$\pi^\top \log \pi \geq {\pi'}^\top \log \pi' + (\pi - \pi')^\top \log \pi' + \frac{1}{2} \cdot \|\pi' - \pi\|_1^2 , \tag{B.853}$$

we have (letting $\pi = \pi_\theta$, and $\pi' = \pi_{\theta'}$),

$$\begin{aligned} D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) &= \pi_\theta^\top \log \pi_\theta - {\pi_{\theta'}}^\top \log \pi_{\theta'} - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} && \text{(B.854)} \\ &\geq \frac{1}{2} \cdot \|\pi_\theta - \pi'_\theta\|_1^2 , && \text{(B.855)} \end{aligned}$$

which is the Pinsker's inequality. Then we have,

$$\begin{aligned} \|\pi_\theta - \pi'_\theta\|_1 &\leq \sqrt{2 \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'})} && \text{(B.856)} \\ &\leq \sqrt{2 \cdot \frac{1}{2} \cdot \|\theta' - \theta - c \cdot \mathbf{1}\|_\infty^2} \qquad \text{(by Lemma 42)} && \text{(B.857)} \\ &= \|\theta' - \theta - c \cdot \mathbf{1}\|_\infty . && \square \end{aligned}$$

**Lemma 40** (Soft performance difference lemma)**.** *For any policies* $\pi$ *and* $\pi'$*,*

$$\begin{aligned} \tilde{V}^{\pi'}(\rho) - \tilde{V}^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^\pi(s) && \text{(B.858)} \\ \cdot \Bigg[ \sum_a (\pi'(a|s) - \pi(a|s)) \cdot \Big[ \tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s) \Big] && \text{(B.859)} \\ + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi'(\cdot|s)) \Bigg] . && \text{(B.860)} \end{aligned}$$

*Proof.* According to the definition of soft value function,

$$\tilde{V}^{\pi'}(s) - \tilde{V}^{\pi}(s) = \sum_a \pi'(a|s) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s)\right] \tag{B.861}$$

$$- \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi}(s,a) - \tau \log \pi(a|s)\right] \tag{B.862}$$

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s)\right] \tag{B.863}$$

$$+ \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s)\right. \tag{B.864}$$

$$\left. - \tilde{Q}^{\pi}(s,a) + \tau \log \pi(a|s)\right] \tag{B.865}$$

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s)\right] \tag{B.866}$$

$$+ \tau D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi'(\cdot|s)) \tag{B.867}$$

$$+ \gamma \sum_a \pi(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left[\tilde{V}^{\pi'}(s') - \tilde{V}^{\pi}(s')\right] \tag{B.868}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi}(s') \tag{B.869}$$

$$\cdot \left[\sum_{a'} (\pi'(a'|s') - \pi(a'|s')) \cdot \left[\tilde{Q}^{\pi'}(s',a') - \tau \log \pi'(a'|s')\right]\right. \tag{B.870}$$

$$\left. + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi'(\cdot|s'))\right], \tag{B.871}$$

finishing the proof. □

**Lemma 41** (Soft sub-optimality lemma)**.** *For any policy* $\pi$*,*

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s \left[d_\rho^{\pi}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_\tau^*(\cdot|s))\right]. \tag{B.872}$$

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

$$\tau \log \pi_\tau^*(a|s) = \tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s). \tag{B.873}$$

According to the soft performance difference lemma of Lemma 40,

$$\tilde{V}^{\pi_\tau^*}(s) - \tilde{V}^{\pi}(s) = \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi}(s') \tag{B.874}$$

$$\cdot \left[ \sum_{a'} (\pi_\tau^*(a'|s') - \pi(a'|s')) \cdot \left[ \tilde{Q}^{\pi_\tau^*}(s', a') - \tau \log \pi_\tau^*(a'|s') \right] \right. \tag{B.875}$$

$$\left. + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right] \tag{B.876}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi}(s') \cdot \left[ \sum_{a'} (\pi_\tau^*(a'|s') - \pi(a'|s')) \cdot \tilde{V}^{\pi_\tau^*}(s') \right. \tag{B.877}$$

$$\left. + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right] \qquad \text{(by Eq. (B.873))} \tag{B.878}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi}(s') \cdot \left[ (1-1) \cdot \tilde{V}^{\pi_\tau^*}(s') \right. \tag{B.879}$$

$$\left. + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right] \tag{B.880}$$

$$= \frac{1}{1-\gamma} \sum_{s'} [d_s^{\pi}(s') \cdot \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s'))]. \qquad \square$$

**Lemma 42** (KL-Logit inequality)**.** *Let* $\pi_\theta = \mathrm{softmax}(\theta)$ *and* $\pi_{\theta'} = \mathrm{softmax}(\theta')$. *Then for any constant* $c \in \mathbb{R}$,

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \leq \frac{1}{2} \cdot \|\theta' - \theta - c \cdot \mathbf{1}\|_\infty^2 . \tag{B.881}$$

*In particular, let* $c = \frac{(\theta'-\theta)^\top \mathbf{1}}{K}$, *we have*

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \leq \frac{1}{2} \cdot \left\| \theta' - \theta - \frac{(\theta'-\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 . \tag{B.882}$$

*Proof.* According to the $\ell_1$ norm strong convexity of negative entropy over probabilistic simplex, i.e., for any policies $\pi$, $\pi'$,

$$\pi'^\top \log \pi' \geq \pi^\top \log \pi + (\pi' - \pi)^\top \log \pi + \frac{1}{2} \cdot \|\pi - \pi'\|_1^2 , \tag{B.883}$$

we have (letting $\pi = \pi_\theta$, and $\pi' = \pi_{\theta'}$),

$$
\begin{aligned}
D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) &= \pi_\theta^\top \log \pi_\theta - {\pi_{\theta'}}^\top \log \pi_{\theta'} - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} && \text{(B.884)} \\
&\leq (\pi_\theta - \pi_{\theta'})^\top \log \pi_\theta - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} && \text{(B.885)} \\
&= (\pi_\theta - \pi_{\theta'})^\top (\log \pi_\theta - \log \pi_{\theta'}) - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 && \text{(B.886)} \\
&= (\pi_\theta - \pi_{\theta'})^\top \Big[ \theta - \theta' && \text{(B.887)} \\
&\quad - \Big( \log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \Big) \cdot \mathbf{1} \Big] && \text{(B.888)} \\
&\quad - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 && \text{(B.889)} \\
&= (\pi_\theta - \pi_{\theta'})^\top (\theta - \theta') - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 && \text{(B.890)} \\
&= (\pi_\theta - \pi_{\theta'})^\top (\theta - \theta' - c \cdot \mathbf{1}) - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 && \text{(B.891)} \\
&\quad \big(\text{since } (\pi_\theta - \pi_{\theta'})^\top c \cdot \mathbf{1} = 0 \text{ holds } \forall c \in \mathbb{R}\big) && \text{(B.892)} \\
&\leq \|\theta - \theta' - c \cdot \mathbf{1}\|_\infty \cdot \|\pi_\theta - \pi_{\theta'}\|_1 - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 && \text{(B.893)} \\
&\quad \text{(by Hölder's inequality)} && \text{(B.894)} \\
&\leq \frac{1}{2} \cdot \|\theta - \theta' - c \cdot \mathbf{1}\|_\infty^2 , && \text{(B.895)}
\end{aligned}
$$

where the last inequality is according to $ax - bx^2 \leq \frac{a^2}{4b}$, $\forall a, b > 0$. □

## B.5 Sub-optimality Guarantees for Entropy-Based RL Methods

Some interesting insight worth mentioning in the proof of Lemma 15 is that the intermediate results provide sub-optimality guarantees for existing entropy regularized RL methods. In particular, Eqs. (B.570) and (B.582) provides policy improvement guarantee for Soft Actor-Critic (Haarnoja et al., 2018, SAC), and Eqs. (B.583) and (B.591) provide sub-optimality guarantees for Patch Consistency Learning (Nachum et al., 2017, PCL).

**Remark 20** (Soft policy improvement inequality)**.** *In Haarnoja et al. (2018, Eq. (4) and Lemma 2), the policy is updated by*

$$
\pi_{\theta_{t+1}} = \arg\min_{\pi_\theta} D_{\mathrm{KL}} \left( \pi_\theta(\cdot|s) \middle\| \frac{\exp\{Q^{\pi_{\theta_t}}(s, \cdot)\}}{\sum_a \exp\{Q^{\pi_{\theta_t}}(s, a)\}} \right), \tag{B.896}
$$

*which is exactly the KL divergence in Eq. (B.582), with $\bar{\pi}_\theta(\cdot|s)$ defined in Eq. (B.570). The soft policy improvement inequality of Eq. (B.582) guarantees that if the soft policy improvement is small, then the sub-optimality is small.*

**Remark 21** (Path inconsistency inequality)**.** *In Nachum et al. (2017, Theorems 1 and 3), it is shown that*

- *(i) soft optimal policy $\pi_\tau^*$ satisfies the consistency conditions Eqs. (2.44) and (2.45);*
- *(ii) for any policy $\pi$ that satisfies the consistency conditions, i.e., if $\forall s, a$,*

$$\pi(a|s) = \exp\left\{(\tilde{Q}^\pi(s,a) - \tilde{V}^\pi(s))/\tau\right\}, \text{ and} \tag{B.897}$$

$$\tilde{V}^\pi(s) = \tau \log \sum_a \exp\left\{\tilde{Q}^\pi(s,a)/\tau\right\}, \tag{B.898}$$

*then $\pi = \pi_\tau^*$, and $\tilde{V}^\pi = \tilde{V}^{\pi_\tau^*}$.*

*However, Nachum et al. (2017) does not show if the consistency is violated during learning, how the violation is related to the sub-optimality. To see why Lemma 15 provides insight, define the following "path inconsistency",*

$$\begin{aligned} & r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^\pi(s') - \tilde{V}^\pi(s) - \tau \log \pi(a|s) \\ &= \tilde{Q}^\pi(s,a) - \tilde{V}^\pi(s) - \tau \log \pi(a|s), \end{aligned} \tag{B.899}$$

*which captures the violation of consistency conditions during learning. Note that for softmax policy $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s,\cdot))$, the r.h.s. of Eq. (B.899) can be written in vector form as*

$$\tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) \tag{B.900}$$

$$= \tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau\theta(s,\cdot) + \tau \log \sum_a \exp\{\theta(s,a)\} \cdot \mathbf{1}. \tag{B.901}$$

*Denote $c_\theta(s) = \frac{\tilde{V}^{\pi_\theta}(s)}{\tau} - \log \sum_a \exp\{\theta(s,a)\}$, and using Lemma 42 in the proof of Lemma 15, in particular, Eq. (B.583),*

$$\begin{aligned} D_{\text{KL}}(\pi_\theta(\cdot|s)\|\bar{\pi}_\theta(\cdot|s)) &\le \frac{1}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2 \\ &= \frac{1}{2\tau^2} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) \right\|_\infty^2. \end{aligned}$$

*Using the above results in Eq.* (B.591),

$$\left[\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}} \leq \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi^*_\tau}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \tag{B.902}$$

$$\cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) \right\|_\infty \tag{B.903}$$

$$= \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi^*_\tau}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \tag{B.904}$$

$$\cdot \max_a \left| r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s') - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \right|, \tag{B.905}$$

*where (square of)* $\left| r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s') - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \right|$ *is exactly the (one-step) path inconsistency objective used in PCL (Nachum et al., 2017, Eq. (14)). Therefore, minimizing path inconsistency guarantees small sub-optimality. The path inconsistency inequality of Eq.* (B.902) *implies path consistency of Nachum et al. (2017).*

# Appendix C

# Proofs for Chapter 3: Escaping the Gravitational Pull of Softmax

## C.1 Proofs for Section 3.2: Softmax Gravity Well

**Theorem 11** (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 $\pi_{\theta_1}$ 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

$$(\pi^* - \pi_{\theta_t})^\top r \geq 0.9 \cdot \Delta, \tag{C.1}$$

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

*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,

$$\pi_\theta^\top r = \pi_\theta(1) \cdot (b + \Delta) + (1 - \pi_\theta(1)) \cdot b. \tag{C.2}$$

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

$$r(a^*) - \pi_\theta^\top r = b + \Delta - \pi_\theta^\top r \tag{C.3}$$

$$= (1 - \pi_\theta(1)) \cdot \Delta. \tag{C.4}$$

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

$$r(a) - \pi_\theta^\top r = b - \pi_\theta^\top r \tag{C.5}$$

$$= -\pi_\theta(1) \cdot \Delta. \tag{C.6}$$

Therefore, the $\ell_2$ norm of softmax policy gradient can be upper bounded as

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left[ \pi_\theta(a^*)^2 \cdot \left( r(a^*) - \pi_\theta^\top r \right)^2 + \sum_{a=2}^{K} \pi_\theta(a)^2 \cdot \left( r(a) - \pi_\theta^\top r \right)^2 \right]^{\frac{1}{2}} \tag{C.7}$$

$$= \left[ \pi_\theta(1)^2 \cdot (1 - \pi_\theta(1))^2 \cdot \Delta^2 + \pi_\theta(1)^2 \cdot \Delta^2 \cdot \sum_{a=2}^{K} \pi_\theta(a)^2 \right]^{\frac{1}{2}} \tag{C.8}$$

(by Eqs. (C.3) and (C.5)) (C.9)

$$= \pi_\theta(1) \cdot \Delta \cdot \left[ (1 - \pi_\theta(1))^2 + \sum_{a=2}^{K} \pi_\theta(a)^2 \right]^{\frac{1}{2}} \tag{C.10}$$

$$\leq \pi_\theta(1) \cdot \Delta \cdot \left[ (1 - \pi_\theta(1))^2 + \left( \sum_{a=2}^{K} \pi_\theta(a) \right)^2 \right]^{\frac{1}{2}} \quad (\|x\|_2 \leq \|x\|_1) \tag{C.11}$$

$$= \sqrt{2} \cdot \pi_\theta(1) \cdot (1 - \pi_\theta(1)) \cdot \Delta. \tag{C.12}$$

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

$$t_{\text{good}} \coloneqq \left\{ t \geq 1 : \pi_{\theta_{t+1}}(1) > \pi_{\theta_t}(1) \right\}, \tag{C.13}$$

$$t_{\text{bad}} \coloneqq \left\{ t \geq 1 : \pi_{\theta_{t+1}}(1) \leq \pi_{\theta_t}(1) \right\}. \tag{C.14}$$

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

$$\frac{1}{\pi_{\theta_t}(1)} - \frac{1}{\pi_{\theta_{t+1}}(1)} = \frac{1}{\pi_{\theta_{t+1}}(1) \cdot \pi_{\theta_t}(1)} \cdot \left( \pi_{\theta_{t+1}}(1) - \pi_{\theta_t}(1) \right) \leq 0. \tag{C.15}$$

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

$$\pi_{\theta_{t+1}}(1) - \pi_{\theta_t}(1) = \left[1 - \frac{1}{\Delta} \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right)\right] \tag{C.16}$$

$$- \left[1 - \frac{1}{\Delta} \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right)\right] \qquad \text{(by Eq. (C.3))} \tag{C.17}$$

$$= \frac{1}{\Delta} \cdot \left[(\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right. \tag{C.18}$$

$$\left. + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right] \tag{C.19}$$

$$\leq \frac{1}{\Delta} \cdot \left[\frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right] \tag{C.20}$$

$$\text{(by Lemma 2)} \tag{C.21}$$

$$= \frac{1}{\Delta} \cdot \left(\frac{5\eta_t^2}{4} + \eta_t\right) \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \tag{C.22}$$

$$\leq \frac{1}{\Delta} \cdot \left(\frac{5\eta_t^2}{4} + \eta_t\right) \cdot 2 \cdot \pi_{\theta_t}(1)^2 \cdot (1 - \pi_{\theta_t}(1))^2 \cdot \Delta^2 \tag{C.23}$$

$$\text{(by Eq. (C.7))} \tag{C.24}$$

$$\leq \frac{9}{2} \cdot \pi_{\theta_t}(1)^2 \cdot (1 - \pi_{\theta_t}(1))^2 \cdot \Delta \qquad (\eta_t \in (0, 1]) \tag{C.25}$$

$$\leq \frac{9}{2} \cdot \pi_{\theta_t}(1)^2 \cdot \Delta. \qquad (\pi_{\theta_t}(1) \in [0, 1]) \tag{C.26}$$

Dividing both sides of Eq. (C.16) with $\pi_{\theta_{t+1}}(1) \cdot \pi_{\theta_t}(1)$, we have,

$$\frac{1}{\pi_{\theta_t}(1)} - \frac{1}{\pi_{\theta_{t+1}}(1)} \leq \frac{9}{2} \cdot \frac{\pi_{\theta_t}(1)}{\pi_{\theta_{t+1}}(1)} \cdot \Delta \tag{C.27}$$

$$\leq \frac{9}{2} \cdot \Delta. \qquad \left(\pi_{\theta_{t+1}}(1) \geq \pi_{\theta_t}(1) > 0\right) \tag{C.28}$$

Therefore, we have,

$$\frac{1}{\pi_{\theta_1}(1)} - \frac{1}{\pi_{\theta_t}(1)} = \sum_{s=1}^{t-1} \left[ \frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)} \right] \tag{C.29}$$

$$= \sum_{s=1,\ s \in t_{\text{good}}}^{t-1} \left[ \frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)} \right] \tag{C.30}$$

$$+ \sum_{s=1,\ s \in t_{\text{bad}}}^{t-1} \left[ \frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)} \right] \tag{C.31}$$

$$\leq \sum_{s=1,\ s \in t_{\text{good}}}^{t-1} \left[ \frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)} \right] \qquad \text{(by Eq. (C.15))} \tag{C.32}$$

$$\leq \sum_{s=1,\ s \in t_{\text{good}}}^{t-1} \left[ \frac{9}{2} \cdot \Delta \right] \qquad \text{(by Eq. (C.27))} \tag{C.33}$$

$$\leq \frac{9}{2} \cdot \Delta \cdot t. \tag{C.34}$$

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

$$\frac{1}{\pi_{\theta_t}(1)} \geq \frac{1}{\pi_{\theta_1}(1)} - \frac{9}{2} \cdot \Delta \cdot t \qquad \text{(by Eq. (C.29))} \tag{C.35}$$

$$\geq \frac{1}{\pi_{\theta_1}(1)} \cdot \left( 1 - \frac{1}{c} \right) \tag{C.36}$$

$$\geq c \cdot \left( 1 - \frac{1}{c} \right) = c - 1 \geq 10, \tag{C.37}$$

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

$$(\pi^* - \pi_{\theta_t})^\top r = (1 - \pi_{\theta_t}(1)) \cdot \Delta \qquad \text{(by Eq. (C.3))} \tag{C.38}$$

$$\geq 0.9 \cdot \Delta. \qquad \square$$

## C.2 Proofs for Section 3.3: Escort Policy Gradient

### C.2.1 Escort Policy Gradient Closed Form in Bandits

For completeness, we show the detailed calculations for the escort policy gradient in bandits, i.e., Eqs. (3.13) and (3.14), which are duplicated here for

convenience,

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = p \cdot \mathrm{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r\right] \tag{C.39}$$

$$= \frac{p}{\|\theta\|_p} \cdot \mathrm{sgn}\{\theta(a)\} \cdot \pi_\theta(a)^{1-1/p} \cdot \left[r(a) - \pi_\theta^\top r\right]. \tag{C.40}$$

According to the chain rule, we have,

$$\frac{d\pi_\theta^\top r}{d\theta} = \left(\frac{d\pi_\theta}{d\theta}\right)^\top \left(\frac{d\pi_\theta^\top r}{d\pi_\theta}\right) = \left(\frac{d\pi_\theta}{d\theta}\right)^\top r. \tag{C.41}$$

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

$$\frac{d\pi_\theta(i)}{d\theta(j)} = \frac{d}{d\theta(j)} \left\{ \frac{|\theta(i)|^p}{\sum_{a'} |\theta(a')|^p} \right\} \tag{C.42}$$

$$= \frac{\delta_{ij} \cdot p \cdot |\theta(i)|^{p-1} \cdot \mathrm{sgn}\{\theta(i)\} \cdot (\sum_{a'} |\theta(a')|^p)}{(\sum_{a'} |\theta(a')|^p)^2} \tag{C.43}$$

$$- \frac{|\theta(i)|^p \cdot p \cdot |\theta(j)|^{p-1} \cdot \mathrm{sgn}\{\theta(j)\}}{(\sum_{a'} |\theta(a')|^p)^2} \tag{C.44}$$

$$= \delta_{ij} \cdot p \cdot \mathrm{sgn}\{\theta(i)\} \cdot \frac{|\theta(i)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \tag{C.45}$$

$$- p \cdot \mathrm{sgn}\{\theta(j)\} \cdot \frac{|\theta(j)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \pi_\theta(i), \tag{C.46}$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{C.47}$$

Then we have the Jacobian as,

$$\left(\frac{d\pi_\theta}{d\theta}\right)^\top = p \cdot \frac{\mathrm{diag}(\mathrm{sgn}\{\theta\})\mathrm{diag}(|\theta|^{p-1})}{\sum_{a'} |\theta(a')|^p} \left[\mathbf{Id} - \mathbf{1}\pi_\theta^\top\right]. \tag{C.48}$$

Combing Eqs. (C.41) and (C.48), we have,

$$\frac{d\pi_\theta^\top r}{d\theta} = p \cdot \frac{\mathrm{diag}(\mathrm{sgn}\{\theta\})\mathrm{diag}(|\theta|^{p-1})}{\sum_{a'} |\theta(a')|^p} \left[r - \mathbf{1} \cdot \left(\pi_\theta^\top r\right)\right], \tag{C.49}$$

which implies Eq. (3.13). Using $\pi_\theta(a) = \frac{|\theta(a)|^p}{\sum_{a'}|\theta(a')|^p}$, we have, if $\theta(a) \neq 0$,

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = p \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^{p-1}}{\sum_{a'}|\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r\right] \tag{C.50}$$

$$= \frac{p}{|\theta(a)|} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^p}{\sum_{a'}|\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r\right] \tag{C.51}$$

$$= \frac{p}{\|\theta\|_p \cdot \pi_\theta(a)^{1/p}} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a) \cdot \left[r(a) - \pi_\theta^\top r\right] \tag{C.52}$$

$$= \frac{p}{\|\theta\|_p} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a)^{1-1/p} \cdot \left[r(a) - \pi_\theta^\top r\right], \tag{C.53}$$

which is Eq. (3.14). On the other hand, if $\theta(a) = 0$, then $\operatorname{sgn}\{\theta(a)\} = \operatorname{sgn}\{0\} = 0$ makes Eq. (3.13) trivially equal to Eq. (3.14).

### C.2.2 One-state MDPs

**Proposition 6.** $\theta \mapsto \pi_\theta^\top r$ is a non-concave function over $\mathbb{R}^K$ using the map $\pi_\theta \coloneqq f_p(\theta)$, $p \geq 1$.

*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,

$$\pi_{\theta_1} = (1/3, 1/3, 1/3)^\top, \tag{C.54}$$

$$\pi_{\theta_2} = \frac{1}{2 + 3^p} \cdot (1, 1, 3^p)^\top. \tag{C.55}$$

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

$$\frac{1}{2} \cdot \left(\pi_{\theta_1}^\top r + \pi_{\theta_2}^\top r\right) = \frac{1}{2} \cdot \left(\frac{2}{3} + \frac{19 + 3^p}{10 \cdot (2 + 3^p)}\right) \tag{C.56}$$

$$= \frac{97 + 23 \cdot 3^p}{60 \cdot (2 + 3^p)} \tag{C.57}$$

$$= \frac{194 + 97 \cdot 2^p + 46 \cdot 3^p + 23 \cdot 2^p \cdot 3^p}{60 \cdot (2 + 3^p) \cdot (2 + 2^p)} \tag{C.58}$$

$$= \frac{19 + 2^p}{10 \cdot (2 + 2^p)} + \frac{17 \cdot (-2 + 5 \cdot 2^p - 4 \cdot 3^p + 2^p \cdot 3^p)}{60 \cdot (2 + 3^p) \cdot (2 + 2^p)} \tag{C.59}$$

$$\geq \frac{19 + 2^p}{10 \cdot (2 + 2^p)} \qquad \text{(see below)} \tag{C.60}$$

$$= \pi_{\bar{\theta}}^\top r, \tag{C.61}$$

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 $x \geq 1$. In fact, for any $x \geq 2$, we have,

$$g(x) \coloneqq -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2^x \cdot 3^x \tag{C.62}$$

$$\geq -2 + 5 \cdot 2^x - 4 \cdot 3^x + 4 \cdot 3^x \qquad (x \geq 2) \tag{C.63}$$

$$= -2 + 5 \cdot 2^x \geq 0. \tag{C.64}$$

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

$$g(x) \coloneqq -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2^x \cdot 3^x \tag{C.65}$$

$$\geq -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2 \cdot 3^x \qquad (x \geq 1) \tag{C.66}$$

$$= -2 + 5 \cdot 2^x - 2 \cdot 3^x \geq 0, \tag{C.67}$$

which is easy to verify. According to Eq. (C.56), $\mathbb{E}_{a \sim \pi_\theta(\cdot)}[r(a)]$ is a non-concave function of $\theta$. □

**Lemma 19** (Non-uniform Smoothness)**.** Suppose $r \in [0, 1]^K$. Let $\pi_\theta \coloneqq f_p(\theta)$, and $\pi_{\theta'} \coloneqq f_p(\theta')$. Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. Then, we have,

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_\zeta\|_p^2} \cdot \|\theta' - \theta\|_2^2. \tag{C.68}$$

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{K}{\|\theta_\zeta\|_1^2} \cdot \|\theta' - \theta\|_2^2. \tag{C.69}$$

*Proof.* Denote the second derivative w.r.t. $\theta$ (i.e., Hessian) as

$$S(r, \theta) = \frac{d}{d\theta} \left\{ \frac{d\pi_\theta^\top r}{d\theta} \right\} \tag{C.70}$$

$$= p \cdot \frac{d}{d\theta} \left\{ \operatorname{diag}\left( \frac{\pi_\theta}{\theta} \right) (r - \pi_\theta^\top r \cdot \mathbf{1}) \right\}. \tag{C.71}$$

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

$$S_{i,j} = p \cdot \frac{d\{\frac{\pi_\theta(i)}{\theta(i)} \cdot (r(i) - \pi_\theta^\top r)\}}{d\theta(j)} \tag{C.72}$$

$$= p \cdot \frac{d\{\frac{\pi_\theta(i)}{\theta(i)}\}}{d\theta(j)} \cdot (r(i) - \pi_\theta^\top r) + p \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{d\{r(i) - \pi_\theta^\top r\}}{d\theta(j)} \tag{C.73}$$

$$= p \cdot \frac{\frac{p}{\theta(j)} \cdot [\delta_{ij} \cdot \pi_\theta(i) - \pi_\theta(i) \cdot \pi_\theta(j)] \cdot \theta(i) - \pi_\theta(i) \cdot \delta_{ij}}{\theta(i)^2} \tag{C.74}$$

$$\cdot (r(i) - \pi_\theta^\top r) \tag{C.75}$$

$$- \frac{\pi_\theta(i)}{\theta(i)} \cdot p^2 \cdot \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \tag{C.76}$$

$$= p \cdot (p-1) \cdot \delta_{ij} \cdot \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \tag{C.77}$$

$$- p^2 \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(i) - \pi_\theta^\top r) \tag{C.78}$$

$$- p^2 \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r), \tag{C.79}$$

where $\delta_{ij}$ is defined in Eq. (C.47). We calculate the spectral radius of $S(r,\theta)$. For any nonzero $y \in \mathbb{R}^K$,

$$\left|y^\top S(r,\theta) y\right| = \left|\sum_{i=1}^K \sum_{j=1}^K S_{i,j} y(i) y(j)\right| \tag{C.80}$$

$$= \Bigg| p \cdot (p-1) \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \cdot y(i)^2 \tag{C.81}$$

$$-2 \cdot p^2 \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot y(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \cdot y(j) \Bigg| \tag{C.82}$$

$$\leq p \cdot (p-1) \cdot \left| \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \cdot y(i)^2 \right| \tag{C.83}$$

$$+ 2 \cdot p^2 \cdot \left| \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot y(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \cdot y(j) \right|, \tag{C.84}$$

where the last inequality is by triangle inequality.

**First part.** For $p \geq 2$, the first term in Eq. (C.80) is upper bounded as,

$$\left| \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \cdot y(i)^2 \right| \tag{C.85}$$

$$\leq \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot \left| r(i) - \pi_\theta^\top r \right| \cdot y(i)^2 \qquad \text{(by triangle inequality)} \tag{C.86}$$

$$\leq \left( \max_a \frac{\pi_\theta(a)}{\theta(a)^2} \cdot |r(a) - \pi_\theta^\top r| \right) \cdot \|y\|_2^2 \tag{C.87}$$

$$\text{(by Hölder's inequality)} \tag{C.88}$$

$$\leq \left( \max_a \frac{\pi_\theta(a)}{\theta(a)^2} \right) \cdot \|y\|_2^2 \qquad (r(a) \in [0,1],\ \forall a) \tag{C.89}$$

$$= \frac{1}{\|\theta\|_p^2} \cdot \left( \max_a \pi_\theta(a)^{1-2/p} \right) \cdot \|y\|_2^2 \tag{C.90}$$

$$\leq \frac{1}{\|\theta\|_p^2} \cdot \|y\|_2^2 . \qquad (p \geq 2) \tag{C.91}$$

The last term in Eq. (C.80) is upper bounded as,

$$\left| \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot y(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \cdot y(j) \right| \tag{C.92}$$

$$\leq \left\| \frac{\pi_\theta}{\theta} \right\|_2 \cdot \|y\|_2 \cdot \left\| \operatorname{diag}\left( \frac{\pi_\theta}{\theta} \right) (r - \pi_\theta^\top r \cdot \mathbf{1}) \right\|_2 \cdot \|y\|_2 \tag{C.93}$$

$$\text{(by Cauchy-Schwarz)} \tag{C.94}$$

$$= \left\| \frac{\pi_\theta}{\theta} \right\|_2 \cdot \left[ \sum_a \left( \frac{\pi_\theta(a)}{\theta(a)} \right)^2 \cdot (r(a) - \pi_\theta^\top r)^2 \right]^{\frac{1}{2}} \cdot \|y\|_2^2 \tag{C.95}$$

$$\leq \left\| \frac{\pi_\theta}{\theta} \right\|_2 \cdot \left[ \sum_a \left( \frac{\pi_\theta(a)}{\theta(a)} \right)^2 \right]^{\frac{1}{2}} \cdot \|y\|_2^2 \qquad (r(a) \in [0,1],\ \forall a) \tag{C.96}$$

$$= \sum_a \left( \frac{\pi_\theta(a)}{\theta(a)} \right)^2 \cdot \|y\|_2^2 \tag{C.97}$$

$$= \frac{1}{\|\theta\|_p^2} \sum_a \left( \pi_\theta(a)^{1-1/p} \right)^2 \cdot \|y\|_2^2 \tag{C.98}$$

$$\leq \frac{1}{\|\theta\|_p^2} \cdot \left( \sum_a \pi_\theta(a)^{1-1/p} \right) \cdot \|y\|_2^2 . \qquad \left( \pi_\theta(a)^{1-1/p} \in [0,1] \right) \tag{C.99}$$

The intermediate term is then upper bounded as,

$$\sum_a \pi_\theta(a)^{1-1/p} = K^{1/p} \cdot \frac{1}{K} \sum_a (K \cdot \pi_\theta(a))^{1-1/p} \tag{C.100}$$

$$\leq K^{1/p} \cdot \left( \sum_a \frac{K \cdot \pi_\theta(a)}{K} \right)^{1-1/p} \qquad \text{(by Jensen's inequality)} \tag{C.101}$$

$$= K^{1/p}. \tag{C.102}$$

Combining Eqs. (C.80), (C.85), (C.92) and (C.100), we have

$$\left| y^\top S(r,\theta) y \right| \leq p \cdot (p-1) \cdot \frac{1}{\|\theta\|_p^2} \cdot \|y\|_2^2 + 2 \cdot p^2 \cdot \frac{1}{\|\theta\|_p^2} \cdot K^{1/p} \cdot \|y\|_2^2 \tag{C.103}$$

$$\leq \frac{3 \cdot p^2 \cdot K^{1/p}}{\|\theta\|_p^2} \cdot \|y\|_2^2. \qquad \left(K^{1/p} \geq 1\right) \tag{C.104}$$

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| = \frac{1}{2} \cdot \left| (\theta' - \theta)^\top S(r, \theta_\zeta) (\theta' - \theta) \right| \tag{C.105}$$

$$\leq \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_\zeta\|_p^2} \cdot \|\theta' - \theta\|_2^2. \qquad \text{(by Eq. (C.103))} \tag{C.106}$$

**Second part.** For $p = 1$, according to Eq. (C.80), Eqs. (C.92) and (C.100), we have,

$$\left| y^\top S(r,\theta) y \right| \leq 2 \cdot p^2 \cdot \frac{1}{\|\theta\|_p^2} \cdot K^{1/p} \cdot \|y\|_2^2 = \frac{2 \cdot K}{\|\theta\|_1^2} \cdot \|y\|_2^2. \tag{C.107}$$

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

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{K}{\|\theta_\zeta\|_1^2} \cdot \|\theta' - \theta\|_2^2. \tag{C.108}$$

□

**Lemma 20** (Non-uniform Łojasiewicz)**.** Let $\pi_\theta = f_p(\theta)$. For any $p > 0$, we have,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \frac{p}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r, \tag{C.109}$$

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

*Proof.* The result follows from calculating the gradient norm,

$$\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 = \left[\sum_{a=1}^{K}\left(p\cdot\frac{\pi_\theta(a)}{\theta(a)}\cdot(r(a)-\pi_\theta^\top r)\right)^2\right]^{\frac{1}{2}} \tag{C.110}$$

$$\geq \left|p\cdot\frac{\pi_\theta(a^*)}{\theta(a^*)}\cdot(r(a^*)-\pi_\theta^\top r)\right| \tag{C.111}$$

$$= p\cdot\frac{\pi_\theta(a^*)}{|\theta(a^*)|}\cdot(\pi^*-\pi_\theta)^\top r \tag{C.112}$$

$$= \frac{p}{\|\theta\|_p}\cdot\pi_\theta(a^*)^{1-1/p}\cdot(\pi^*-\pi_\theta)^\top r. \qquad \left(\pi_\theta(a)=\frac{|\theta(a)|^p}{\sum_{a'}|\theta(a')|^p}\right) \qquad \square$$

**Theorem 12.** 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 gap for all $t\geq 1$:

**(gradient flow)** for $p\geq 1$, with $\eta_t=\|\theta_t\|_p^2$,

$$(\pi^*-\pi_{\theta_t})^\top r \leq \frac{1}{c^{2-2/p}\cdot t+1}, \tag{C.113}$$

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

$$(\pi^*-\pi_{\theta_t})^\top r \leq \frac{9\cdot K^{1/p}}{c^{2-2/p}}\cdot\frac{1}{t}, \tag{C.114}$$

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

$$(\pi^*-\pi_{\theta_t})^\top r \leq \frac{9K}{t}, \tag{C.115}$$

where $c\coloneqq\inf_t\pi_{\theta_t}(a^*)>0$ is a problem- and initialization-dependent, but time-independent constant.

*Proof.* **First part.** For $p\geq 2$, according to Lemma 19,

$$\left|(\pi_{\theta_{t+1}}-\pi_{\theta_t})^\top r-\left\langle\frac{d\pi_{\theta_t}^\top r}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle\right| \tag{C.116}$$

$$\leq \frac{3\cdot p^2\cdot K^{1/p}}{2\cdot\|\theta_{\zeta_t}\|_p^2}\cdot\|\theta_{t+1}-\theta_t\|_2^2, \tag{C.117}$$

where

$$\theta_{\zeta_t}\coloneqq\theta_t+\zeta_t\cdot(\theta_{t+1}-\theta_t)=\theta_t+\zeta_t\cdot\eta_t\cdot\frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \tag{C.118}$$

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_p = \left[ \sum_{a=1}^{K} \left| p \cdot \frac{\pi_\theta(a)}{\theta(a)} \cdot (r(a) - \pi_\theta^\top r) \right|^p \right]^{\frac{1}{p}} \tag{C.119}$$

$$= p \cdot \left[ \sum_{a=1}^{K} \left( \frac{\pi_\theta(a)}{|\theta(a)|} \cdot |r(a) - \pi_\theta^\top r| \right)^p \right]^{\frac{1}{p}} \tag{C.120}$$

$$= \frac{p}{\|\theta\|_p} \cdot \left[ \sum_{a=1}^{K} \left( \pi_\theta(a)^{1-1/p} \cdot |r(a) - \pi_\theta^\top r| \right)^p \right]^{\frac{1}{p}} \tag{C.121}$$

$$\leq \frac{p}{\|\theta\|_p} \cdot \left[ \sum_{a=1}^{K} (1 \cdot 1)^p \right]^{\frac{1}{p}} = \frac{p \cdot K^{1/p}}{\|\theta\|_p}. \tag{C.122}$$

According to the triangle inequality, we have,

$$\|\theta_{\zeta_t}\|_p \geq \|\theta_t\|_p - \zeta_t \cdot \eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_p \tag{C.123}$$

$$\geq \|\theta_t\|_p - \zeta_t \cdot \eta_t \cdot \frac{p \cdot K^{1/p}}{\|\theta_t\|_p}. \qquad \text{(by Eq. (C.119))} \tag{C.124}$$

$$= \|\theta_t\|_p \cdot \left( 1 - \zeta_t \cdot \frac{2}{9 \cdot p} \right) \qquad \left( \eta_t = \frac{2 \cdot \|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}} \right) \tag{C.125}$$

$$\geq \|\theta_t\|_p \cdot \left( 1 - \frac{2}{9 \cdot p} \right) \qquad (\zeta_t \in [0, 1]) \tag{C.126}$$

$$= \|\theta_t\|_p \cdot \left[ \left( 1 - \frac{2}{\sqrt{6}} \right) \cdot \left( 1 - \frac{2 \cdot (3 + \sqrt{6})}{9 \cdot p} \right) + \frac{2}{\sqrt{6}} \right] \tag{C.127}$$

$$\geq \frac{2}{\sqrt{6}} \cdot \|\theta_t\|_p. \qquad (p \geq 2) \tag{C.128}$$

Combining Eqs. (C.116) and (C.123), we have,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \tag{C.129}$$

$$\leq \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_{\zeta_t}\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{C.130}$$

$$\leq \frac{9 \cdot p^2 \cdot K^{1/p}}{4 \cdot \|\theta_t\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{C.131}$$

which implies,

$$\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r \le -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{9 \cdot p^2 \cdot K^{1/p}}{4 \cdot \|\theta_t\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{C.132}$$

$$= -\eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{9 \cdot p^2 \cdot K^{1/p}}{4 \cdot \|\theta_t\|_p^2} \cdot \eta_t^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \tag{C.133}$$

$$\left( \theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right) \tag{C.134}$$

$$= -\frac{\|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left( \eta_t = \frac{2 \cdot \|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}} \right) \tag{C.135}$$

$$\le -\frac{\cancel{\|\theta_t\|_p^2}}{9 \cdot \cancel{p^2} \cdot K^{1/p}} \cdot \left[ \frac{\cancel{p}}{\cancel{\|\theta_t\|_p}} \cdot \pi_{\theta_t}(a^*)^{1-1/p} \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \tag{C.136}$$

(by Lemma 20) (C.137)

$$= -\frac{\pi_{\theta_t}(a^*)^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left[ (\pi^* - \pi_{\theta_t})^\top r \right]^2 \tag{C.138}$$

$$\le -\frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left[ (\pi^* - \pi_{\theta_t})^\top r \right]^2, \tag{C.139}$$

where $c_t \coloneqq \min_{1 \le s \le t} \pi_{\theta_s}(a^*) > 0$. Eq. (C.132) is equivalent to,

$$(\pi^* - \pi_{\theta_{t+1}})^\top r - (\pi^* - \pi_{\theta_t})^\top r \le -\frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left[ (\pi^* - \pi_{\theta_t})^\top r \right]^2. \tag{C.140}$$

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

$$\delta_2 \le 1 \le \frac{9 \cdot K^{1/p}}{c_2^{2-2/p}} \cdot \frac{1}{2}. \tag{C.141}$$

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,

$$\delta_{t+1} \leq \delta_t - \frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \delta_t^2 \qquad \text{(by Eq. (C.140))} \tag{C.142}$$

$$\leq \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} - \frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left(\frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t}\right)^2 \tag{C.143}$$

$$\left(\delta_t \leq \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} \leq \frac{9 \cdot K^{1/p}}{2 \cdot c_t^{2-2/p}},\ t \geq 2\right) \tag{C.144}$$

$$= \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \left(\frac{1}{t} - \frac{1}{t^2}\right) \tag{C.145}$$

$$\leq \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t+1}, \tag{C.146}$$

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

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} \tag{C.147}$$

$$\leq \frac{9 \cdot K^{1/p}}{(\inf_{t \geq 1} \pi_{\theta_t}(a^*))^{2-2/p}} \cdot \frac{1}{t}. \tag{C.148}$$

**Second part.** For $p = 1$, according to Lemma 19,

$$\left|(\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle\right| \leq \frac{K}{\|\theta_{\zeta_t}\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{C.149}$$

where

$$\theta_{\zeta_t} \coloneqq \theta_t + \zeta_t \cdot (\theta_{t+1} - \theta_t) = \theta_t + \zeta_t \cdot \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \tag{C.150}$$

for some $\zeta_t \in [0, 1]$. The $\ell_1$ norm can be upper bounded as

$$\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_1 = \sum_{a=1}^{K} \left|\frac{\pi_{\theta_t}(a)}{\theta_t(a)} \cdot \left(r(a) - \pi_\theta^\top r\right)\right| \tag{C.151}$$

$$= \frac{1}{\|\theta_t\|_1} \sum_{a=1}^{K} |r(a) - \pi_\theta^\top r| \tag{C.152}$$

$$\leq \frac{K}{\|\theta_t\|_1}. \qquad \left(r \in [0, 1]^K\right) \tag{C.153}$$

According to triangle inequality, we have,

$$\|\theta_{\zeta_t}\|_1 \geq \|\theta_t\|_1 - \zeta_t \cdot \eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_1 \tag{C.154}$$

$$\geq \|\theta_t\|_1 - \zeta_t \cdot \eta_t \cdot \frac{K}{\|\theta_t\|_1}. \qquad \text{(by Eq. (C.151))} \tag{C.155}$$

$$= \|\theta_t\|_1 \cdot \left(1 - \zeta_t \cdot \frac{2}{9}\right) \qquad \left(\eta_t = \frac{2 \cdot \|\theta_t\|_1^2}{9 \cdot K}\right) \tag{C.156}$$

$$\geq \frac{2}{3} \cdot \|\theta_t\|_1. \qquad (\zeta_t \in [0, 1]) \tag{C.157}$$

Combining Eqs. (C.149) and (C.154), we have,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \tag{C.158}$$

$$\leq \frac{K}{\|\theta_{\zeta_t}\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_1^2 \tag{C.159}$$

$$\leq \frac{9 \cdot K}{4 \cdot \|\theta_t\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_1^2, \tag{C.160}$$

which implies,

$$\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r \leq -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{9 \cdot K}{4 \cdot \|\theta_t\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_1^2 \tag{C.161}$$

$$= -\eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{9 \cdot K}{4 \cdot \|\theta_t\|_1^2} \cdot \eta_t^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \tag{C.162}$$

$$\left(\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \tag{C.163}$$

$$= -\frac{\|\theta_t\|_1^2}{9 \cdot K} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\eta_t = \frac{2 \cdot \|\theta_t\|_1^2}{9 \cdot K}\right) \tag{C.164}$$

$$\leq -\frac{\cancel{\|\theta_t\|_1^2}}{9 \cdot K} \cdot \left[ \frac{1}{\cancel{\|\theta_t\|_1}} \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \text{(by Lemma 20)} \tag{C.165}$$

$$= -\frac{1}{9 \cdot K} \cdot \left[ (\pi^* - \pi_{\theta_t})^\top r \right]^2. \tag{C.166}$$

Using a similar induction argument as in Eq. (C.142), we have

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{9 \cdot K}{t}. \tag{C.167}$$

**Third part.** For the gradient flow, we have,

$$\frac{d\{(\pi^* - \pi_{\theta_t})^\top r\}}{dt} = -\frac{d\pi_{\theta_t}^\top r}{dt} \tag{C.168}$$

$$= -\left(\frac{d\theta_t}{dt}\right)^\top \left(\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \tag{C.169}$$

$$= -\eta_t \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 \qquad \left(\frac{d\theta_t}{dt} = \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \tag{C.170}$$

$$\leq -\eta_t \cdot \left[\frac{p}{\|\theta_t\|_p} \cdot \pi_{\theta_t}(a^*)^{1-1/p} \cdot (\pi^* - \pi_{\theta_t})^\top r\right]^2 \quad \text{(by Lemma 20)} \tag{C.171}$$

$$= -\pi_{\theta_t}(a^*)^{2-2/p} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r\right]^2 \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{p^2}\right) \tag{C.172}$$

$$\leq -c^{2-2p} \cdot [(\pi^* - \pi_{\theta_t})^\top r]^2, \tag{C.173}$$

which implies,

$$\frac{d}{dt}\left\{\frac{1}{(\pi^* - \pi_{\theta_t})^\top r}\right\} = -\frac{1}{[(\pi^* - \pi_{\theta_t})^\top r]^2} \cdot \frac{d\{(\pi^* - \pi_{\theta_t})^\top r\}}{dt} \tag{C.174}$$

$$= c^{2-2p}. \tag{C.175}$$

Taking integral, we have,

$$\frac{1}{(\pi^* - \pi_{\theta_t})^\top r} = \frac{1}{(\pi^* - \pi_{\theta_1})^\top r} + c^{2-2p} \cdot (t-1) \tag{C.176}$$

$$\geq 1 + c^{2-2p} \cdot (t-1), \qquad \left((\pi^* - \pi_{\theta_1})^\top r \in (0, 1]\right) \tag{C.177}$$

which is equivalent to

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

### C.2.3 General MDPs

**Lemma 43.** *The escort policy gradient w.r.t. $\theta$ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot A^{\pi_\theta}(s,a), \tag{C.178}$$

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

$$A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s), \tag{C.179}$$

$$Q^{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) V^{\pi_\theta}(s'). \tag{C.180}$$

*Proof.* According to Theorem 1, we have,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s' \sim d_\mu^{\pi_\theta}} \left[ \sum_a \frac{\partial \pi_\theta(a|s')}{\partial \theta} \cdot Q^{\pi_\theta}(s', a) \right]. \tag{C.181}$$

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[ \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right] \tag{C.182}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left( \frac{d\pi(\cdot|s)}{d\theta(s,\cdot)} \right)^\top Q^{\pi_\theta}(s,\cdot) \tag{C.183}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \operatorname{diag}\left( \frac{\pi(\cdot|s)}{\theta(s,\cdot)} \right) \left( \mathbf{Id} - \mathbf{1}\pi_\theta^\top \right) Q^{\pi_\theta}(s,\cdot). \tag{C.184}$$

For each component $a$, we have

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \tag{C.185}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \left[ Q^{\pi_\theta}(s,a) - \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right] \tag{C.186}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot (Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)) \tag{C.187}$$

$$\left( V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right) \tag{C.188}$$

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

**Lemma 44** (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 $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. Denote $A \coloneqq |\mathcal{A}|$ as the total number of actions. Then we have,*

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

$$\left| V^{\pi_{\theta'}}(\rho) - V^{\pi_\theta}(\rho) - \left\langle \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta}, \theta' - \theta \right\rangle \right| \tag{C.189}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{\|\theta' - \theta\|_2^2}{\min_s \|\theta_\zeta(s,\cdot)\|_p^2}. \tag{C.190}$$

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

$$\left| V^{\pi_{\theta'}}(\rho) - V^{\pi_\theta}(\rho) - \left\langle \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta}, \theta' - \theta \right\rangle \right| \tag{C.191}$$

$$\leq \frac{4 \cdot A^2}{(1-\gamma)^3} \cdot \frac{\|\theta' - \theta\|_2^2}{\min_s \|\theta_\zeta(s,\cdot)\|_1^2}. \tag{C.192}$$

*Proof.* Denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. For any $s \in \mathcal{S}$,

$$\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{C.193}$$

$$= \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta}, u \right\rangle \right|. \tag{C.194}$$

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

$$\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \right\rangle \right| \tag{C.195}$$

$$= \sum_a p \cdot \frac{\pi_\theta(a|s)}{|\theta(s,a)|} \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \tag{C.196}$$

$$= \sum_a \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \pi_\theta(a|s)^{1-1/p} \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \tag{C.197}$$

$$\leq \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \max_a \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \cdot \sum_a \pi_\theta(a|s)^{1-1/p} \tag{C.198}$$

$$\leq \frac{p}{\|\theta(s,\cdot)\|_p} \cdot 2 \cdot \|u\|_\infty \cdot A^{1/p} \qquad \text{(by Eq. (C.100))} \tag{C.199}$$

$$\leq \frac{2 \cdot p \cdot A^{1/p}}{\|\theta(s,\cdot)\|_p} \cdot \|u\|_2. \tag{C.200}$$

Similarly, the second derivative is,

$$\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{C.201}$$

$$= \sum_a \left| \left\langle \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{C.202}$$

$$= \sum_a \left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right|. \tag{C.203}$$

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

$$S_{i,j} = p \cdot \frac{\partial\{\delta_{ia} \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} - \pi_\theta(a|s) \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)}\}}{\partial \theta(s,j)} \tag{C.204}$$

$$= p \cdot \delta_{ia} \cdot \frac{\frac{p}{\theta(s,j)} \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] \cdot \theta(s,a) - \delta_{ja}\pi_\theta(a|s)}{\theta(s,a)^2} \tag{C.205}$$

$$- \frac{p^2}{\theta(s,j)} \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \tag{C.206}$$

$$- p \cdot \pi_\theta(a|s) \cdot \frac{\frac{p}{\theta(s,j)} \cdot [\delta_{ij}\pi_\theta(i|s) - \pi_\theta(i|s)\pi_\theta(j|s)] \cdot \theta(s,i) - \delta_{ij}\pi_\theta(i|s)}{\theta(s,i)^2} \tag{C.207}$$

$$= \delta_{ia} \cdot \delta_{ja} \cdot p \cdot (p-1) \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)^2} \tag{C.208}$$

$$- \delta_{ia} \cdot p^2 \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \frac{\pi_\theta(j|s)}{\theta(s,j)} - \delta_{ja} \cdot p^2 \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \tag{C.209}$$

$$+ p \cdot \pi_\theta(a|s) \cdot \left[2 \cdot p \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \cdot \frac{\pi_\theta(j|s)}{\theta(s,j)} - \delta_{ij} \cdot (p-1) \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)^2}\right], \tag{C.210}$$

where the $\delta$ notation is as defined in Eq. (C.47). Then we have,

$$\left|\left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot)\right\rangle\right| = \left|\sum_{i=1}^{A}\sum_{j=1}^{A} S_{i,j} u(s,i)u(s,j)\right| \tag{C.211}$$

$$\leq p \cdot (p-1) \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)^2} \cdot u(s,a)^2 \tag{C.212}$$

$$+ 2 \cdot p^2 \cdot \frac{\pi_\theta(a|s)}{|\theta(s,a)|} \cdot |u(s,a)| \cdot \left|\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)}\right)^\top u(s,\cdot)\right| \tag{C.213}$$

$$+ \pi_\theta(a|s) \cdot \left[2 \cdot p^2 \cdot \left|\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)}\right)^\top u(s,\cdot)\right|^2 \right. \tag{C.214}$$

$$\left. + p \cdot (p-1) \cdot \left|\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)^2}\right)^\top (u(s,\cdot) \odot u(s,\cdot))\right|\right] \tag{C.215}$$

$$= \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s)^{1-2/p} \cdot u(s,a)^2 \tag{C.216}$$

$$+ \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s)^{1-1/p} \cdot |u(s,a)| \cdot \left|\left(\pi_\theta(\cdot|s)^{1-1/p}\right)^\top u(s,\cdot)\right| \tag{C.217}$$

$$+ \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s) \cdot \left|\left(\pi_\theta(\cdot|s)^{1-1/p}\right)^\top u(s,\cdot)\right|^2 \tag{C.218}$$

$$+ \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s) \cdot \left|\left(\pi_\theta(\cdot|s)^{1-2/p}\right)^\top (u(s,\cdot) \odot u(s,\cdot))\right|. \tag{C.219}$$

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

$$\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \leq \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \sum_a u(s,a)^2 \tag{C.220}$$

$$+ \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \cdot \|\pi_\theta(\cdot|s)^{1-1/p}\|_2^2 \tag{C.221}$$

$$+ \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \cdot \|\pi_\theta(\cdot|s)^{1-1/p}\|_2^2 \tag{C.222}$$

$$+ \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \| \cdot \|\pi_\theta(\cdot|s)^{1-2/p}\|_\infty \cdot \|u(s,\cdot) \odot u(s,\cdot)\|_1 \tag{C.223}$$

$$\leq \frac{2 \cdot p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 + \frac{4 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot A^{1/p} \cdot \|u(s,\cdot)\|_2^2 \tag{C.224}$$

$$\text{(by Eq. (C.100))} \tag{C.225}$$

$$\leq \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p})}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \tag{C.226}$$

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

$$[P(\alpha)]_{(s,s')} = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \mathcal{P}(s'|s,a). \tag{C.227}$$

The derivative w.r.t. $\alpha$ is

$$\left[ \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{C.228}$$

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

$$\left[ \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} x \right]_{(s)} = \sum_{s'} \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s'). \tag{C.229}$$

The $\ell_\infty$ norm is upper bounded as,

$$\left\| \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} x \right\|_\infty = \max_s \left| \sum_{s'} \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \tag{C.230}$$

$$\leq \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{C.231}$$

$$= \max_s \sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{C.232}$$

$$\leq \max_s \frac{2 \cdot p \cdot A^{1/p}}{\|\theta(s,\cdot)\|_p} \cdot \|u\|_2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (C.195))} \tag{C.233}$$

Similarly, taking second derivative w.r.t. $\alpha$,

$$\left[\frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0}\right]_{(s,s')} = \sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0}\right] \cdot \mathcal{P}(s'|s,a). \tag{C.234}$$

The $\ell_\infty$ norm is upper bounded as,

$$\begin{aligned}
&\left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0} x \right\|_\infty && \text{(C.235)}\\
&\quad = \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0}\right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| && \text{(C.236)}\\
&\quad \le \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty && \text{(C.237)}\\
&\quad = \max_s \sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty && \text{(C.238)}\\
&\quad \le \max_s \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p})}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2 \cdot \|x\|_\infty. && \text{(C.239)}\\
&\qquad \text{(by Eq. (C.220))} && \text{(C.240)}
\end{aligned}$$

Next, consider the state value function of $\pi_{\theta_\alpha}$,

$$\begin{aligned}
V^{\pi_{\theta_\alpha}}(s) &= \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a) && \text{(C.241)}\\
&\quad + \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot V^{\pi_{\theta_\alpha}}(s'), && \text{(C.242)}
\end{aligned}$$

which implies,

$$V^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) r_{\theta_\alpha}, \tag{C.243}$$

where

$$M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1}, \tag{C.244}$$

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

$$r_{\theta_\alpha}(s) = \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a). \tag{C.245}$$

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

$$M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1} = \sum_{t=0}^{\infty} \gamma^t \left[P(\alpha)\right]^t, \tag{C.246}$$

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

$$\mathbf{1} = \frac{1}{1-\gamma} \cdot (\mathbf{Id} - \gamma P(\alpha))\, \mathbf{1} \Longrightarrow M(\alpha)\mathbf{1} = \frac{1}{1-\gamma} \cdot \mathbf{1}, \tag{C.247}$$

which implies, $\forall i$,

$$\left\| [M(\alpha)]_{i,:} \right\|_1 = \sum_j [M(\alpha)]_{(i,j)} = \frac{1}{1-\gamma}. \tag{C.248}$$

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

$$\begin{aligned} \|M(\alpha)x\|_\infty &= \max_i \left| [M(\alpha)]_{i,:}^\top x \right| && \text{(C.249)} \\ &\leq \max_i \left\| [M(\alpha)]_{i,:} \right\|_1 \cdot \|x\|_\infty && \text{(C.250)} \\ &= \frac{1}{1-\gamma} \cdot \|x\|_\infty. && \text{(C.251)} \end{aligned}$$

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

$$\|r_{\theta_\alpha}\|_\infty = \max_s |r_{\theta_\alpha}(s)| = \max_s \left| \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s, a) \right| \leq 1. \tag{C.252}$$

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

$$\begin{aligned} \left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| &= \left| \left( \frac{\partial r_{\theta_\alpha}(s)}{\partial \theta_\alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| && \text{(C.253)} \\ &= \left| \left( \frac{\partial \{\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\}}{\partial \theta_\alpha(s,\cdot)} \right)^\top u(s,\cdot) \right| && \text{(C.254)} \\ &= \left| p \cdot \left( \mathrm{diag}(1/\theta_\alpha(s,\cdot)) \left( \mathrm{diag}(\pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s)\pi_{\theta_\alpha}(\cdot|s)^\top \right) r(s,\cdot) \right)^\top u(s,\cdot) \right| && \text{(C.255)} \\ &\leq p \cdot \left\| \mathrm{diag}(1/\theta_\alpha(s,\cdot)) \left( \mathrm{diag}(\pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s)\pi_{\theta_\alpha}(\cdot|s)^\top \right) r(s,\cdot) \right\|_1 && \text{(C.256)} \\ &\quad \cdot \|u(s,\cdot)\|_\infty . && \text{(C.257)} \end{aligned}$$

The $\ell_1$ norm is upper bounded as,

$$\left\|\mathrm{diag}(1/\theta_\alpha(s,\cdot))\left(\mathrm{diag}(\pi_{\theta_\alpha}(\cdot|s))-\pi_{\theta_\alpha}(\cdot|s)\pi_{\theta_\alpha}(\cdot|s)^\top\right)r(s,\cdot)\right\|_1 \tag{C.258}$$

$$=\sum_a \frac{\pi_{\theta_\alpha}(a|s)^{1-1/p}}{\|\theta_\alpha(s,\cdot)\|_p}\cdot\left|r(s,a)-\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\right| \tag{C.259}$$

$$\le \frac{1}{\|\theta_\alpha(s,\cdot)\|_p}\cdot\max_a\left|r(s,a)-\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\right|\cdot\sum_a\pi_{\theta_\alpha}(a|s)^{1-1/p} \tag{C.260}$$

$$\le \frac{1}{\|\theta_\alpha(s,\cdot)\|_p}\cdot\sum_a\pi_{\theta_\alpha}(a|s)^{1-1/p} \qquad (r(s,a)\in[0,1]) \tag{C.261}$$

$$\le \frac{A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p}. \qquad \text{(by Eq. (C.100))} \tag{C.262}$$

Combining Eqs. (C.253) and (C.258), we have,

$$\left\|\frac{\partial r_{\theta_\alpha}}{\partial\alpha}\right\|_\infty=\max_s\left|\frac{\partial r_{\theta_\alpha}(s)}{\partial\alpha}\right| \tag{C.263}$$

$$\le \max_s p\cdot\left\|\mathrm{diag}(1/\theta_\alpha(s,\cdot))\left(\mathrm{diag}(\pi_{\theta_\alpha}(\cdot|s))-\pi_{\theta_\alpha}(\cdot|s)\pi_{\theta_\alpha}(\cdot|s)^\top\right)r(s,\cdot)\right\|_1 \tag{C.264}$$

$$\cdot\,\|u(s,\cdot)\|_\infty \tag{C.265}$$

$$\le \max_s\frac{p\cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p}\cdot\|u\|_2. \tag{C.266}$$

Similarly, for the second derivative, we have,

$$\left\|\frac{\partial^2 r_{\theta_\alpha}}{\partial\alpha^2}\right\|_\infty=\max_s\left|\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial\alpha^2}\right| \tag{C.267}$$

$$=\max_s\left|\left(\frac{\partial}{\partial\theta_\alpha}\left\{\frac{\partial r_{\theta_\alpha}(s)}{\partial\alpha}\right\}\right)^\top\frac{\partial\theta_\alpha}{\partial\alpha}\right| \tag{C.268}$$

$$=\max_s\left|\left(\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial\theta_\alpha^2}\frac{\partial\theta_\alpha}{\partial\alpha}\right)^\top\frac{\partial\theta_\alpha}{\partial\alpha}\right| \tag{C.269}$$

$$=\max_s\left|u(s,\cdot)^\top\frac{\partial^2\{\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\}}{\partial\theta_\alpha(s,\cdot)^2}u(s,\cdot)\right| \tag{C.270}$$

$$\le \max_s\frac{3\cdot p^2\cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2}\cdot\|u(s,\cdot)\|_2^2 \qquad \text{(by Eq. (C.103))} \tag{C.271}$$

$$\le \max_s\frac{3\cdot p^2\cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2}\cdot\|u\|_2^2. \tag{C.272}$$

Taking derivative w.r.t. $\alpha$ in Eq. (C.243), we have,

$$\frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial\alpha}=\gamma\cdot e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial\alpha}M(\alpha)r_{\theta_\alpha}+e_s^\top M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial\alpha}. \tag{C.273}$$

Taking second derivative w.r.t. $\alpha$, we have,

$$\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} = 2\gamma^2 \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \tag{C.274}$$

$$+ \gamma \cdot e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \tag{C.275}$$

$$+ 2\gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}. \tag{C.276}$$

For the last term,

$$\left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right| \le \|e_s\|_1 \cdot \left\| M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \tag{C.277}$$

$$\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (C.249))} \tag{C.278}$$

$$\le \frac{1}{1-\gamma} \cdot \max_s \frac{3 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (C.267))} \tag{C.279}$$

For the second last term,

$$\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right| \le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \tag{C.280}$$

$$\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (C.249))} \tag{C.281}$$

$$\le \frac{2 \cdot p \cdot A^{1/p} \cdot \|u\|_2}{1-\gamma} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left\| M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \tag{C.282}$$

$$\text{(by Eq. (C.230))} \tag{C.283}$$

$$\le \frac{2 \cdot p \cdot A^{1/p} \cdot \|u\|_2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \tag{C.284}$$

$$\text{(by Eq. (C.249))} \tag{C.285}$$

$$\le \frac{2 \cdot p^2 \cdot A^{2/p} \cdot \|u\|_2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2. \tag{C.286}$$

$$\text{(by Eq. (C.263))} \tag{C.287}$$

For the second term,

$$\left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right| \le \left\|M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{C.288}$$

$$\le \frac{1}{1-\gamma}\cdot\left\|\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \quad \text{(by Eq. (C.249))} \tag{C.289}$$

$$\le \frac{2\cdot p^2\cdot(1+2\cdot A^{1/p})\cdot\|u\|_2^2}{1-\gamma} \tag{C.290}$$

$$\cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\left\|M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \quad \text{(by Eq. (C.235))} \tag{C.291}$$

$$\le \frac{2\cdot p^2\cdot(1+2\cdot A^{1/p})\cdot\|u\|_2^2}{(1-\gamma)^2}\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\left\|r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{C.292}$$

$$\text{(by Eq. (C.249))} \tag{C.293}$$

$$\le \frac{2\cdot p^2\cdot(1+2\cdot A^{1/p})}{(1-\gamma)^2}\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\|u\|_2^2. \tag{C.294}$$

$$\text{(by Eq. (C.252))} \tag{C.295}$$

For the first term, according to Eq. (C.230), Eqs. (C.249) and (C.252),

$$\left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{C.296}$$

$$\le \left\|M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right\|_\infty \tag{C.297}$$

$$\le \frac{1}{1-\gamma}\cdot\max_s\frac{2\cdot p\cdot A^{1/p}}{\|\theta(s,\cdot)\|_p}\cdot\|u\|_2 \tag{C.298}$$

$$\cdot\frac{1}{1-\gamma}\cdot\max_s\frac{2\cdot p\cdot A^{1/p}}{\|\theta(s,\cdot)\|_p}\cdot\|u\|_2\cdot\frac{1}{1-\gamma}\cdot 1 \tag{C.299}$$

$$= \frac{4\cdot p^2\cdot A^{2/p}}{(1-\gamma)^3}\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\|u\|_2^2. \tag{C.300}$$

Combining Eqs. (C.277), (C.280), (C.288) and (C.296) with Eq. (C.274), we

have,

$$\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \leq 2\gamma^2 \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{C.301}$$

$$+\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha)r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{C.302}$$

$$+2\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| \tag{C.303}$$

$$+\left|e_s^\top M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| \tag{C.304}$$

$$\leq \left(\frac{8\cdot\gamma^2\cdot p^2\cdot A^{2/p}}{(1-\gamma)^3}+\frac{2\cdot\gamma\cdot p^2\cdot(1+2\cdot A^{1/p})}{(1-\gamma)^2}\right. \tag{C.305}$$

$$\left.+\frac{4\cdot\gamma\cdot p^2\cdot A^{2/p}}{(1-\gamma)^2}+\frac{3\cdot p^2\cdot A^{1/p}}{1-\gamma}\right)\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\|u\|_2^2 \tag{C.306}$$

$$\leq \frac{8\cdot p^2\cdot A^{2/p}}{(1-\gamma)^3}\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\|u\|_2^2, \tag{C.307}$$

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

$$\left|y^\top\frac{\partial^2 V^{\pi_\theta}(s)}{\partial\theta^2}y\right| = \left|\left(\frac{y}{\|y\|_2}\right)^\top\frac{\partial^2 V^{\pi_\theta}(s)}{\partial\theta^2}\left(\frac{y}{\|y\|_2}\right)\right|\cdot\|y\|_2^2 \tag{C.308}$$

$$\leq \max_{\|u\|_2=1}\left|\left\langle\frac{\partial^2 V^{\pi_\theta}(s)}{\partial\theta^2}u,u\right\rangle\right|\cdot\|y\|_2^2 \tag{C.309}$$

$$= \max_{\|u\|_2=1}\left|\left\langle\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial\theta_\alpha^2}\Big|_{\alpha=0}\frac{\partial\theta_\alpha}{\partial\alpha},\frac{\partial\theta_\alpha}{\partial\alpha}\right\rangle\right|\cdot\|y\|_2^2 \tag{C.310}$$

$$= \max_{\|u\|_2=1}\left|\left\langle\frac{\partial}{\partial\theta_\alpha}\left\{\frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial\alpha}\right\}\Big|_{\alpha=0},\frac{\partial\theta_\alpha}{\partial\alpha}\right\rangle\right|\cdot\|y\|_2^2 \tag{C.311}$$

$$= \max_{\|u\|_2=1}\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial\alpha^2}\Big|_{\alpha=0}\right|\cdot\|y\|_2^2 \tag{C.312}$$

$$\leq \frac{8\cdot p^2\cdot A^{2/p}}{(1-\gamma)^3}\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot\|y\|_2^2. \qquad \text{(by Eq. (C.301))} \tag{C.313}$$

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

$\forall \theta,\ \theta'$,

$$\left|V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta \right\rangle\right| \tag{C.314}$$

$$= \frac{1}{2} \cdot \left|(\theta' - \theta)^\top \frac{\partial^2 V^{\pi_{\theta_\zeta}}(s)}{\partial \theta_\zeta^2} (\theta' - \theta)\right| \tag{C.315}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta_\zeta(s,\cdot)\|_p^2} \cdot \|\theta' - \theta\|_2^2 \tag{C.316}$$

$$\text{(by Eq. (C.308))} \tag{C.317}$$

$$= \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\zeta(s,\cdot)\|_p^2} \cdot \|\theta' - \theta\|_2^2. \tag{C.318}$$

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

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

$$\left|\left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot)\right\rangle\right| \tag{C.319}$$

$$\leq \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s)^{1-1/p} \cdot |u(s,a)| \cdot \left|\left(\pi_\theta(\cdot|s)^{1-1/p}\right)^\top u(s,\cdot)\right| \tag{C.320}$$

$$+ \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s) \cdot \left|\left(\pi_\theta(\cdot|s)^{1-1/p}\right)^\top u(s,\cdot)\right|^2. \tag{C.321}$$

$$\text{(by Eq. (C.211))} \tag{C.322}$$

Therefore we have,

$$\sum_a \left|\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \leq \frac{4 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \cdot \|\pi_\theta(\cdot|s)^{1-1/p}\|_2^2 \tag{C.323}$$

$$\leq \frac{4 \cdot p^2 \cdot A}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \tag{C.324}$$

Similar to Eq. (C.235), we have,

$$\left\|\frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0} x\right\|_\infty \leq \max_s \sum_a \left|\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \cdot \|x\|_\infty \tag{C.325}$$

$$\leq \max_s \frac{4 \cdot p^2 \cdot A}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (C.323))} \tag{C.326}$$

Similar to Eq. (C.267), for the second derivative, we have,

$$\left\|\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\right\|_\infty = \max_s \left| u(s,\cdot)^\top \frac{\partial^2 \{\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\}}{\partial \theta_\alpha(s,\cdot)^2} u(s,\cdot) \right| \tag{C.327}$$

$$\leq \max_s \frac{2 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \qquad \text{(by Eq. (C.107))} \tag{C.328}$$

$$\leq \max_s \frac{2 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \tag{C.329}$$

Similar to Eq. (C.277), we have,

$$\left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right| \leq \|e_s\|_1 \cdot \left\| M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \tag{C.330}$$

$$\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (C.249))} \tag{C.331}$$

$$\leq \frac{1}{1-\gamma} \cdot \max_s \frac{2 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (C.327))} \tag{C.332}$$

Similar to Eq. (C.288), we have,

$$\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right| \leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \tag{C.333}$$

$$\text{(by Eq. (C.249))} \tag{C.334}$$

$$\leq \frac{4 \cdot p^2 \cdot A \cdot \|u\|_2^2}{1-\gamma} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \left\| M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \tag{C.335}$$

$$\text{(by Eq. (C.325))} \tag{C.336}$$

$$\leq \frac{4 \cdot p^2 \cdot A \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \left\| r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \tag{C.337}$$

$$\text{(by Eq. (C.249))} \tag{C.338}$$

$$\leq \frac{4 \cdot p^2 \cdot A}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (C.252))} \tag{C.339}$$

Combining Eqs. (C.280) and (C.296), Eqs. (C.330) and (C.333) with Eq. (C.274),

we have,

$$\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| \le 2\gamma^2 \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{C.340}$$

$$+ \gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial \alpha^2}M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{C.341}$$

$$+ 2\gamma \cdot \left|e_s^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0}\right| \tag{C.342}$$

$$+ \left|e_s^\top M(\alpha)\frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0}\right| \tag{C.343}$$

$$\le \left(\frac{8\cdot\gamma^2\cdot p^2\cdot A^{2/p}}{(1-\gamma)^3} + \frac{4\cdot\gamma\cdot p^2\cdot A}{(1-\gamma)^2}\right. \tag{C.344}$$

$$\left. + \frac{4\cdot\gamma\cdot p^2\cdot A^{2/p}}{(1-\gamma)^2} + \frac{2\cdot p^2\cdot A^{1/p}}{1-\gamma}\right)\cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2}\cdot \|u\|_2^2 \tag{C.345}$$

$$\le \frac{8\cdot A^2}{(1-\gamma)^3}\cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_1^2}\cdot \|u\|_2^2. \qquad (p=1) \tag{C.346}$$

Similar to Eq. (C.308), Eq. (C.340) implies for all $y \in \mathbb{R}^{SA}$ and $\theta$,

$$\left|y^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} y\right| \le \max_{\|u\|_2=1}\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right|\cdot \|y\|_2^2 \tag{C.347}$$

$$\le \frac{8\cdot A^2}{(1-\gamma)^3}\cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_1^2}\cdot \|y\|_2^2. \qquad \text{(Eq. (C.340))} \tag{C.348}$$

Similar to Eq. (C.314), we have, $\forall s$, $\forall \theta,\ \theta'$,

$$\left|V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta\right\rangle\right| \tag{C.349}$$

$$\le \frac{4\cdot A^2}{(1-\gamma)^3}\cdot \max_s \frac{\|\theta'-\theta\|_2^2}{\|\theta_\zeta(s,\cdot)\|_1^2} \qquad \text{(Eq. (C.347))} \tag{C.350}$$

$$= \frac{4\cdot A^2}{(1-\gamma)^3}\cdot \frac{\|\theta'-\theta\|_2^2}{\min_s \|\theta_\zeta(s,\cdot)\|_1^2}. \tag{C.351}$$

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

**Lemma 45** (Non-uniform Łojasiewicz)**.** *Suppose $\mu(s) > 0$ for all state $s$ and $\pi_\theta \coloneqq f_p(\theta)$. Then,*

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \ge \frac{p}{\sqrt{S}}\cdot \left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1} \tag{C.352}$$

$$\cdot \frac{\min_s \pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s \|\theta(s,\cdot)\|_p}\cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho)\right], \tag{C.353}$$

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

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

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \left[ \sum_{s,a} \left( \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \tag{C.354}$$

$$\geq \left[ \sum_{s} \left( \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))} \right)^2 \right]^{\frac{1}{2}} \tag{C.355}$$

$$\geq \frac{1}{\sqrt{S}} \sum_{s} \left| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))} \right| \tag{C.356}$$

$$\text{(by Cauchy-Schwarz, } \|x\|_1 = |\langle \mathbf{1}, |x| \rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2) \tag{C.357}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_{s} \left| d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a^*(s)|s)}{\theta(s,a^*(s))} \cdot A^{\pi_\theta}(s,a^*(s)) \right| \tag{C.358}$$

$$\text{(by Lemma 43)} \tag{C.359}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_{s} d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a^*(s)|s)}{|\theta(s,a^*(s))|} \cdot \left| A^{\pi_\theta}(s,a^*(s)) \right|. \tag{C.360}$$

$$\left( d_\mu^{\pi_\theta}(s) \geq 0,\ \pi_\theta(a^*(s)|s) \geq 0 \right) \tag{C.361}$$

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

have,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2 \geq \frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)}\cdot d_\rho^{\pi^*}(s) \tag{C.362}$$

$$\cdot\, p\cdot\frac{\pi_\theta(a^*(s)|s)}{|\theta(s,a^*(s))|}\cdot|A^{\pi_\theta}(s,a^*(s))| \tag{C.363}$$

$$=\frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)}\cdot d_\rho^{\pi^*}(s)\cdot p\cdot\frac{1}{\|\theta(s,\cdot)\|_p} \tag{C.364}$$

$$\cdot\,(\pi_\theta(a^*(s)|s))^{1-1/p}\cdot|A^{\pi_\theta}(s,a^*(s))| \tag{C.365}$$

$$\geq\frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot p\cdot\min_s\frac{1}{\|\theta(s,\cdot)\|_p} \tag{C.366}$$

$$\cdot\min_s\pi_\theta(a^*(s)|s)^{1-1/p}\cdot\sum_s d_\rho^{\pi^*}(s)\cdot|A^{\pi_\theta}(s,a^*(s))| \tag{C.367}$$

$$\geq\frac{1}{1-\gamma}\cdot\frac{1}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot p\cdot\min_s\frac{1}{\|\theta(s,\cdot)\|_p} \tag{C.368}$$

$$\cdot\min_s\pi_\theta(a^*(s)|s)^{1-1/p}\cdot\sum_s d_\rho^{\pi^*}(s)\cdot A^{\pi_\theta}(s,a^*(s)) \tag{C.369}$$

$$=\frac{p}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\frac{\min_s\pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s\|\theta(s,\cdot)\|_p} \tag{C.370}$$

$$\cdot\frac{1}{1-\gamma}\sum_s d_\rho^{\pi^*}(s)\sum_a\pi^*(a|s)\cdot A^{\pi_\theta}(s,a) \tag{C.371}$$

$$=\frac{p}{\sqrt{S}}\cdot\left\|\frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-1}\cdot\frac{\min_s\pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s\|\theta(s,\cdot)\|_p}\cdot\left[V^*(\rho)-V^{\pi_\theta}(\rho)\right], \tag{C.372}$$

where the last equation is according to the performance difference lemma of Lemma 34. □

### C.2.4 An Equivalent Algorithm with Parameter Normalization

For convenience of analysis, we introduce Algorithm 3, which is equivalent to Algorithm 1 as shown in Lemma 46.

**Lemma 46.** *Using the escort transform $\pi_\theta = f_p(\theta)$, Algorithm 3 with constant learning rate $\eta$ and Algorithm 1 with learning rate $\eta_t(s) = \eta\cdot\|\theta_t(s,\cdot)\|_p^2$ are*

**Algorithm 3** 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**

*equivalent, i.e., for all* $(s,a)$,

$$\tilde{\theta}_t(s,a) = \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p}, \quad \textit{and} \tag{C.373}$$

$$\pi_{\tilde{\theta}_t}(a|s) = \pi_{\theta_t}(a|s). \tag{C.374}$$

*Proof.* For $t = 1$, according to Algorithm 3, we have, for all $(s,a)$, $\tilde{\theta}_1(s,a) = \frac{\theta_1(s,a)}{\|\theta_1(s,\cdot)\|_p}$, and,

$$\pi_{\tilde{\theta}_1}(a|s) = \frac{|\tilde{\theta}_1(s,a)|^p}{\sum_{a'} |\tilde{\theta}_1(s,a')|^p} \tag{C.375}$$

$$= \frac{|\theta_1(s,a)|^p}{\sum_{a'} |\theta_1(s,a')|^p} \cdot \frac{1}{\cancel{\|\theta_1(s,\cdot)\|_p^p}} \cdot \cancel{\|\theta_1(s,\cdot)\|_p^p} = \pi_{\theta_1}(a|s). \tag{C.376}$$

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. (C.375), we have, for all $(s,a)$, $\pi_{\tilde{\theta}_t}(a|s) = \pi_{\theta_t}(a|s)$, and,

$$\tilde{\zeta}_{t+1}(s,a) \leftarrow \tilde{\theta}_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s,a)} \qquad \text{(Algorithm 3)} \tag{C.377}$$

$$= \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p} + \eta \cdot \|\theta_t(s,\cdot)\|_p \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \tag{C.378}$$

$$\left(\text{induction hypothesis and } \pi_{\tilde{\theta}_t} = \pi_{\theta_t}\right) \tag{C.379}$$

$$= \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p} + \eta_t(s) \cdot \frac{1}{\|\theta_t(s,\cdot)\|_p} \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \tag{C.380}$$

$$\left(\eta_t(s) = \eta \cdot \|\theta_t(s,\cdot)\|_p^2\right) \tag{C.381}$$

$$= \frac{1}{\|\theta_t(s,\cdot)\|_p} \cdot \left(\theta_t(s,a) + \eta_t(s) \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}\right) \tag{C.382}$$

$$= \frac{1}{\|\theta_t(s,\cdot)\|_p} \cdot \theta_{t+1}(s,a). \qquad \text{(Algorithm 1)} \tag{C.383}$$

Therefore we have,

$$\tilde{\theta}_{t+1}(s,a) \leftarrow \frac{\tilde{\zeta}_{t+1}(s,a)}{\|\tilde{\zeta}_{t+1}(s,\cdot)\|_p} \qquad \text{(Algorithm 3)} \tag{C.384}$$

$$= \frac{1}{\cancel{\|\theta_t(s,\cdot)\|_p}} \cdot \theta_{t+1}(s,a) \cdot \frac{\cancel{\|\theta_t(s,\cdot)\|_p}}{\|\theta_{t+1}(s,\cdot)\|_p} \qquad \text{(by Eq. (C.377))} \tag{C.385}$$

$$= \frac{\theta_{t+1}(s,a)}{\|\theta_{t+1}(s,\cdot)\|_p}. \tag{C.386}$$

Using similar calculation as in Eq. (C.375), we have, for all $(s,a)$, $\pi_{\tilde{\theta}_{t+1}}(a|s) = \pi_{\theta_{t+1}}(a|s)$. □

**Theorem 13.** Following the escort policy gradient with any initialization such that $|\theta_1(s,a)| > 0$, $\forall(s,a)$ to get $\{\theta_t\}_{t\geq 1}$, for any $t \geq 1$, the following upper bounds hold for $\pi_{\theta_t}$,

**(i)** for $p \geq 2$, with $\eta_t = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{1/p}}$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{20 \cdot A^{1/p} \cdot S}{c^{2-2/p} \cdot (1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty, \tag{C.387}$$

**(ii)** for $p = 1$, with $\eta_t = \frac{(1-\gamma)^3}{10 \cdot A}$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{20 \cdot A \cdot S}{(1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty, \tag{C.388}$$

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 method.

*Proof.* Note that for any $\theta$ and $\mu$,

$$d_\mu^{\pi_\theta}(s) = \mathop{\mathbb{E}}_{s_0\sim\mu}\left[d_\mu^{\pi_\theta}(s)\right] \tag{C.389}$$

$$= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s|s_0,\pi_\theta,\mathcal{P})\right] \tag{C.390}$$

$$\geq \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\Pr(s_0 = s|s_0)\right] \tag{C.391}$$

$$= (1-\gamma)\cdot\mu(s). \tag{C.392}$$

According to the value sub-optimality lemma of Lemma 36,

$$V^*(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{C.393}$$

$$= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{C.394}$$

$$\leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{C.395}$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{C.396}$$

$$\left( \text{by Eq. (C.389) and } \min_s \mu(s) > 0 \right) \tag{C.397}$$

$$= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[ V^*(\mu) - V^{\pi_\theta}(\mu) \right], \tag{C.398}$$

where the first inequality is because of

$$\sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \geq 0, \tag{C.399}$$

and the last equation is again by Lemma 36.

For $p \geq 2$ and $p = 1$, according to Lemma 44, $V^{\pi_\theta}(\mu)$ is $\beta$-smooth with $\beta = \frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_{\lambda_t}(s,\cdot)\|_p^2}$, i.e., we have, in Algorithm 3,

$$\left| V^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| \tag{C.400}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p^2}, \tag{C.401}$$

where

$$\tilde{\theta}_{\lambda_t} \coloneqq \tilde{\theta}_t + \lambda_t \cdot (\tilde{\zeta}_{t+1} - \tilde{\theta}_t) \tag{C.402}$$

$$= \tilde{\theta}_t + \lambda_t \cdot \eta \cdot \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \qquad \text{(Algorithm 3)} \tag{C.403}$$

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

$$\|\tilde{\theta}_{\lambda_t}(s_{\lambda_t},\cdot)\|_p \geq \|\tilde{\theta}_t(s_{\lambda_t},\cdot)\|_p - \lambda_t \cdot \eta \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_{\lambda_t},\cdot)} \right\|_p \tag{C.404}$$

$$\text{(by triangle inequality)} \tag{C.405}$$

$$\geq \min_s \|\tilde{\theta}_t(s,\cdot)\|_p - \lambda_t \cdot \eta \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_{\lambda_t},\cdot)} \right\|_p. \tag{C.406}$$

The $\ell_p$ gradient norm can be upper bounded as,

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)}\right\|_p = \left[\sum_a \left|\frac{1}{1-\gamma}\cdot d_\mu^{\pi_\theta}(s)\cdot p\cdot \frac{\pi_\theta(a|s)}{\theta(s,a)}\cdot A^{\pi_\theta}(s,a)\right|^p\right]^{\frac{1}{p}} \tag{C.407}$$

$$\text{(by Lemma 43)} \tag{C.408}$$

$$\leq \frac{p}{1-\gamma}\cdot\left[\sum_a \left|\frac{\pi_\theta(a|s)}{\theta(s,a)}\cdot A^{\pi_\theta}(s,a)\right|^p\right]^{\frac{1}{p}} \tag{C.409}$$

$$= \frac{p}{1-\gamma}\cdot\frac{1}{\|\theta(s,\cdot)\|_p}\cdot\left[\sum_a \left(\pi_\theta(a|s)^{1-1/p}\cdot|A^{\pi_\theta}(s,a)|\right)^p\right]^{\frac{1}{p}} \tag{C.410}$$

$$\leq \frac{p}{1-\gamma}\cdot\frac{1}{\|\theta(s,\cdot)\|_p}\cdot\left[\sum_a \left(1\cdot\frac{1}{1-\gamma}\right)^p\right]^{\frac{1}{p}} \tag{C.411}$$

$$\leq \frac{p\cdot A^{1/p}}{(1-\gamma)^2}\cdot\max_s\frac{1}{\|\theta(s,\cdot)\|_p}. \tag{C.412}$$

Combining Eqs. (C.404) and (C.407), we have,

$$\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p \tag{C.413}$$

$$\geq \min_s \|\tilde{\theta}_t(s,\cdot)\|_p - \lambda_t\cdot\eta\cdot\frac{p\cdot A^{1/p}}{(1-\gamma)^2}\cdot\frac{1}{\min_s\|\tilde{\theta}_t(s,\cdot)\|_p} \tag{C.414}$$

$$= 1-\lambda_t\cdot\eta\cdot\frac{p\cdot A^{1/p}}{(1-\gamma)^2} \tag{C.415}$$

$$\left(\|\tilde{\theta}_t(s,\cdot)\|_p = 1,\ \text{for all } s,\ \text{Algorithm 3}\right) \tag{C.416}$$

$$= 1-\lambda_t\cdot\frac{1-\gamma}{10\cdot p\cdot A^{1/p}} \tag{C.417}$$

$$\left(\eta = \frac{(1-\gamma)^3}{10\cdot p^2\cdot A^{2/p}},\ \text{by Lemma 46}\right) \tag{C.418}$$

$$\geq 1-\frac{1-\gamma}{10\cdot p\cdot A^{1/p}} \qquad (\lambda_t\in[0,1]) \tag{C.419}$$

$$= \left(1-\frac{2}{\sqrt{5}}\right)\cdot\left(1-\frac{5+2\sqrt{5}}{10}\cdot\frac{1-\gamma}{p\cdot A^{1/p}}\right)+\frac{2}{\sqrt{5}} \tag{C.420}$$

$$\geq \frac{2}{\sqrt{5}}. \qquad \left(p\geq 2,\ A^{1/p}\geq 1,\ 1-\gamma\in(0,1]\right) \tag{C.421}$$

Combining Eqs. (C.400) and (C.413), we have,

$$\left|V^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle\right| \tag{C.422}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p^2} \tag{C.423}$$

$$\leq \frac{5 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2, \tag{C.424}$$

which implies,

$$V^{\pi_{\tilde{\theta}_t}}(\mu) - V^{\pi_{\tilde{\theta}_{t+1}}}(\mu) = V^{\pi_{\tilde{\theta}_t}}(\mu) - V^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) \tag{C.425}$$

$$\left(\tilde{\theta}_{t+1}(s,a) = \frac{\tilde{\zeta}_{t+1}(s,a)}{\|\tilde{\zeta}_{t+1}(s,\cdot)\|_p}, \text{ Algorithm 3}\right) \tag{C.426}$$

$$\leq -\left\langle \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle + \frac{5 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2 \tag{C.427}$$

$$\text{(by Eq. (C.422))} \tag{C.428}$$

$$= -\eta \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 + \frac{5 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \eta^2 \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \tag{C.429}$$

$$\left(\tilde{\zeta}_{t+1} = \tilde{\theta}_t + \eta \cdot \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \text{ Algorithm 3}\right) \tag{C.430}$$

$$= -\frac{(1-\gamma)^3}{20 \cdot p^2 \cdot A^{2/p}} \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \qquad \left(\eta = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p}}\right) \tag{C.431}$$

$$\leq -\frac{(1-\gamma)^3}{20 \cdot \cancel{p^2} \cdot A^{2/p}} \cdot \left[\frac{\cancel{p}}{\sqrt{S}} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-1} \cdot \frac{\min_s \pi_{\tilde{\theta}_t}(a^*(s)|s)^{1-1/p}}{\max_s \|\tilde{\theta}_t(s,\cdot)\|_p}\right. \tag{C.432}$$

$$\left. \cdot \left[V^*(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu)\right]\right]^2 \qquad \text{(Lemma 45)} \tag{C.433}$$

$$= -\frac{(1-\gamma)^3}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-2} \cdot \min_s \pi_{\tilde{\theta}_t}(a^*(s)|s)^{2-2/p} \tag{C.434}$$

$$\cdot \left[V^*(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu)\right]^2 \qquad \left(\|\tilde{\theta}_t(s,\cdot)\|_p = 1, \text{ for all } s\right) \tag{C.435}$$

$$\leq -\frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \tag{C.436}$$

$$\cdot \min_s \pi_{\tilde{\theta}_t}(a^*(s)|s)^{2-2/p} \cdot \left[V^*(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu)\right]^2, \tag{C.437}$$

where the last inequality is by Eq. (C.389). Then we have,

$$V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) = V^{\pi_{\tilde{\theta}_t}}(\mu) - V^{\pi_{\tilde{\theta}_{t+1}}}(\mu) \qquad \text{(by Lemma 46)} \tag{C.438}$$

$$\leq -\frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \min_s \pi_{\theta_t}(a^*(s)|s)^{2-2/p} \tag{C.439}$$

$$\cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu)\right]^2 \qquad \text{(by Eq. (C.425) and Lemma 46)} \tag{C.440}$$

$$\leq -\frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot c^{2-2/p} \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu)\right]^2, \tag{C.441}$$

which is equivalent to,

$$\delta_{t+1} - \delta_t \leq -\frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot c^{2-2/p} \cdot \delta_t^2, \tag{C.442}$$

where $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. Using the similar induction argument as in Eq. (C.142), we have,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \leq \frac{20 \cdot A^{2/p} \cdot S}{(1-\gamma)^5 \cdot t} \cdot \frac{1}{c^{2-2/p}} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2, \tag{C.443}$$

which leads to the final result,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu)\right] \tag{C.444}$$

$$\text{(by Eq. (C.393))} \tag{C.445}$$

$$\leq \frac{20 \cdot A^{2/p} \cdot S}{c^{2-2/p} \cdot (1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty. \qquad \square$$

### C.2.5 Entropy Regularized MDPs

**Lemma 47.** *The entropy regularized escort policy gradient w.r.t. $\theta$ is*

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \tilde{A}^{\pi_\theta}(s,a), \tag{C.446}$$

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \tag{C.447}$$

$$\cdot \, diag\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)}\right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta(\cdot|s)^\top\right) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s)\right]. \tag{C.448}$$

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

$$\tilde{A}^{\pi_\theta}(s,a) = \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \tag{C.449}$$

$$\tilde{Q}^{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s'). \tag{C.450}$$

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

$$\tilde{V}^{\pi_\theta}(\mu) = \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right]. \tag{C.451}$$

Taking derivative w.r.t. $\theta$,

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} = \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{C.452}$$

$$+ \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} - \tau \frac{1}{\pi_\theta(a|s)} \frac{\partial \pi_\theta(a|s)}{\partial \theta}\right] \tag{C.453}$$

$$= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{C.454}$$

$$+ \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} \tag{C.455}$$

$$= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{C.456}$$

$$+ \gamma \cdot \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \frac{\partial \tilde{V}^{\pi_\theta}(s')}{\partial \theta} \tag{C.457}$$

$$= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right], \tag{C.458}$$

where the second equation is because of

$$\sum_a \pi_\theta(a|s) \cdot \left[\frac{1}{\pi_\theta(a|s)} \frac{\partial \pi_\theta(a|s)}{\partial \theta}\right] = \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \tag{C.459}$$

$$= \frac{\partial}{\partial \theta} \sum_a \pi_\theta(a|s) = \frac{\partial 1}{\partial \theta} = 0. \tag{C.460}$$

Using similar arguments as in the proof for Lemma 43, i.e., for $s' \neq s$, $\frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = \mathbf{0}$,

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{C.461}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d\pi_\theta(\cdot|s)}{d\theta(s,\cdot)}\right)^\top \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s)\right] \tag{C.462}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \tag{C.463}$$

$$\cdot \operatorname{diag}\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)}\right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta(\cdot|s)^\top\right) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s)\right]. \tag{C.464}$$

For each component $a$, we have

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \Bigg[ \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \tag{C.465}$$

$$- \sum_a \pi_\theta(a|s) \cdot \Big[ \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \Big] \Bigg] \tag{C.466}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \tag{C.467}$$

$$\cdot \Big[ \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \Big] \tag{C.468}$$

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

**Lemma 48** (Non-uniform Łojasiewicz)**.** *Suppose $\mu(s) > 0$ for all $s \in \mathcal{S}$ and $\pi_\theta = f_p(\theta)$. Then,*

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \tag{C.469}$$

$$\cdot \frac{p \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p}}{\max_s \|\theta(s,\cdot)\|_p} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \Big[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \Big]^{\frac{1}{2}}. \tag{C.470}$$

*Proof.* According to the definition of soft value functions,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \tag{C.471}$$

$$= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi_\tau^*(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau\log\pi_\tau^*(a_t|s_t))\right] - \tilde{V}^{\pi_\theta}(\rho) \tag{C.472}$$

$$= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi_\tau^*(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \Bigg[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau\log\pi_\tau^*(a_t|s_t) \tag{C.473}$$

$$+ \tilde{V}^{\pi_\theta}(s_t) - \tilde{V}^{\pi_\theta}(s_t))\Bigg] - \tilde{V}^{\pi_\theta}(\rho) \tag{C.474}$$

$$= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi_\tau^*(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \Bigg[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau\log\pi_\tau^*(a_t|s_t) \tag{C.475}$$

$$+ \gamma\tilde{V}^{\pi_\theta}(s_{t+1}) - \tilde{V}^{\pi_\theta}(s_t))\Bigg] \tag{C.476}$$

$$= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_\tau^*}(s) \cdot \Bigg[\sum_a \pi_\tau^*(a|s) \cdot \Big(r(s,a) - \tau\log\pi_\tau^*(a|s) \tag{C.477}$$

$$+ \gamma\sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s') - \tilde{V}^{\pi_\theta}(s)\Big)\Bigg] \tag{C.478}$$

$$= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_\tau^*}(s) \cdot \Bigg[\sum_a \pi_\tau^*(a|s) \cdot \Big(\tilde{Q}^{\pi_\theta}(s,a) - \tau\log\pi_\tau^*(a|s)\Big) \tag{C.479}$$

$$- \tilde{V}^{\pi_\theta}(s)\Bigg]. \tag{C.480}$$

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

$$\bar{\pi}_\theta(a|s) = \frac{\exp\left\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\right\}}{\sum_{a'}\exp\left\{\tilde{Q}^{\pi_\theta}(s,a')/\tau\right\}},\ \forall a. \tag{C.481}$$

We have, $\forall s$,

$$\sum_a \pi_\tau^*(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau\log\pi_\tau^*(a|s)\Big] \tag{C.482}$$

$$\leq \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau\log\pi(a|s)\Big] \tag{C.483}$$

$$= \sum_a \bar{\pi}_\theta(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau\log\bar{\pi}_\theta(a|s)\Big] \tag{C.484}$$

$$= \tau\log\sum_a \exp\Big\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\Big\}. \tag{C.485}$$

Also note that,

$$\tilde{V}^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s)\right] \tag{C.486}$$

$$= \sum_a \pi_\theta(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) \tag{C.487}$$

$$+ \tau \log \bar{\pi}_\theta(a|s) - \tau \log \pi_\theta(a|s)\Big] \tag{C.488}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s)\right] \tag{C.489}$$

$$- \tau D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \tag{C.490}$$

$$= \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\theta}(s,a)/\tau\right\} - \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)). \tag{C.491}$$

Combining Eq. (C.471), Eqs. (C.482) and (C.486), we have,

$$\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \tag{C.492}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*_\tau}(s) \tag{C.493}$$

$$\cdot \left[\sum_a \pi^*_\tau(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi^*_\tau(a|s)\right] - \tilde{V}^{\pi_\theta}(s)\right] \tag{C.494}$$

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

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

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*_\tau}(s) \cdot \frac{\tau}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \log \pi_\theta(s,\cdot) - \frac{c_\theta(s)}{\tau} \cdot \mathbf{1} \right\|_\infty^2 \tag{C.497}$$

(by Lemma 42) (C.498)

$$= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi^*_\tau}(s)}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2, \tag{C.499}$$

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. (C.492), we

have,

$$\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}} \tag{C.500}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \left[ \sum_s \frac{d_\rho^{\pi_\tau^*}(s)}{2\tau} \right. \tag{C.501}$$

$$\left. \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2 \right]^{\frac{1}{2}} \tag{C.502}$$

$$= \frac{1}{\sqrt{1-\gamma}} \cdot \left[ \sum_s \left( \sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \right. \right. \tag{C.503}$$

$$\left. \left. \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty \right)^2 \right]^{\frac{1}{2}} \tag{C.504}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \sum_s \sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \tag{C.505}$$

$$\cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty \qquad (\|x\|_2 \leq \|x\|_1) \tag{C.506}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \tag{C.507}$$

$$\cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty . \tag{C.508}$$

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

bounded as

$$\left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 = \left[\sum_s \left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)}\right\|_2^2\right]^{\frac{1}{2}} \tag{C.509}$$

$$\geq \frac{1}{\sqrt{S}} \sum_s \left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)}\right\|_2 \tag{C.510}$$

$$\text{(by Cauchy-Schwarz , } \|x\|_1 = |\langle \mathbf{1},\ |x|\rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2) \tag{C.511}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \tag{C.512}$$

$$\cdot \left\| p \cdot \operatorname{diag}\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)}\right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta(\cdot|s)^\top\right) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s)\cdot \mathbf{1}\right]\right\|_2 \tag{C.513}$$

$$\text{(by Lemma 47)} \tag{C.514}$$

$$= \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s,\cdot)\|_p} \tag{C.515}$$

$$\cdot \left\| \operatorname{diag}\left(\pi_\theta(\cdot|s)^{1-1/p}\right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta(\cdot|s)^\top\right) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s)\cdot \mathbf{1}\right]\right\|_2 \tag{C.516}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \min_a \pi_\theta(a|s)^{1-1/p} \tag{C.517}$$

$$\cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s)\cdot \mathbf{1}\right\|_2 . \text{ (by Lemma 51)} \tag{C.518}$$

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

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \tag{C.519}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \min_a \pi_\theta(a|s)^{1-1/p} \cdot \|\zeta_\theta(s)\|_2 \tag{C.520}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \min_a \pi_\theta(a|s)^{1-1/p} \cdot \|\zeta_\theta(s)\|_\infty \tag{C.521}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_s \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p} \tag{C.522}$$

$$\cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[ \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \|\zeta_\theta(s)\|_\infty \right] \tag{C.523}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_s \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p} \tag{C.524}$$

$$\cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} \qquad \text{(by Eq. (C.500))} \tag{C.525}$$

$$\geq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_s \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p} \tag{C.526}$$

$$\cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}}, \tag{C.527}$$

where the last inequality is by $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\cdot\mu(s)$ (cf. Eq. (C.389)). 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 Lemma 15. □

**Theorem 14.** 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

$$\eta_t = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{1/p} + c_\tau}, \tag{C.528}$$

to get $\{\theta_t\}_{t\geq 1}$, for all $t \geq 1$, the following sub-optimality upper bounds hold

for $\pi_{\theta_t}$, for $p \geq 2$:

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \frac{\|1/\mu\|_\infty}{\exp\{C_\tau \cdot c'^2 \cdot t\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2}, \tag{C.529}$$

where $c' > c \coloneqq \inf_{(s,a)} \inf_{t \geq 1} \pi_{\theta_t}(a|s) > 0$, $\tau$ is the temperature for entropy regularization, $\pi_\tau^*$ is the softmax optimal policy, and $c_\tau$, $C_\tau$ are problem-dependent constants.

*Proof.* According to the soft sub-optimality lemma of Lemma 41,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) = \frac{1}{1-\gamma} \sum_s \left[ d_\rho^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \tag{C.530}$$

$$= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_{\theta_t}}(s)}{d_\mu^{\pi_{\theta_t}}(s)} \cdot \left[ d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \tag{C.531}$$

$$\leq \frac{1}{(1-\gamma)^2} \sum_s \frac{1}{\mu(s)} \cdot \left[ d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \tag{C.532}$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s \left[ d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \tag{C.533}$$

$$= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right], \tag{C.534}$$

where the last equation is again by Lemma 41, and the first inequality is according to $d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (C.389)).

According to Lemma 46, 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 3 with learning rate $\eta$. We have, in Algorithm 3,

$$\left| \tilde{V}^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| \tag{C.535}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p^2}, \tag{C.536}$$

where

$$\tilde{\theta}_{\lambda_t} \coloneqq \tilde{\theta}_t + \lambda_t \cdot (\tilde{\zeta}_{t+1} - \tilde{\theta}_t) \tag{C.537}$$

$$= \tilde{\theta}_t + \lambda_t \cdot \eta \cdot \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \qquad \text{(Algorithm 3)} \tag{C.538}$$

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

$$\|\tilde{\theta}_{\lambda_t}(s_{\lambda_t},\cdot)\|_p \geq \|\tilde{\theta}_t(s_{\lambda_t},\cdot)\|_p - \lambda_t \cdot \eta \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_{\lambda_t},\cdot)} \right\|_p \tag{C.539}$$

$$\text{(by triangle inequality)} \tag{C.540}$$

$$\geq \min_s \|\tilde{\theta}_t(s,\cdot)\|_p - \lambda_t \cdot \eta \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_{\lambda_t},\cdot)} \right\|_p. \tag{C.541}$$

The $\ell_p$ gradient norm can be upper bounded as,

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_p = \left[ \sum_a \left| \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \tilde{A}^{\pi_\theta}(s,a) \right|^p \right]^{\frac{1}{p}} \tag{C.542}$$

$$\text{(by Lemma 47)} \tag{C.543}$$

$$\leq \frac{p}{1-\gamma} \cdot \left[ \sum_a \left| \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \tilde{A}^{\pi_\theta}(s,a) \right|^p \right]^{\frac{1}{p}} \tag{C.544}$$

$$= \frac{p}{1-\gamma} \cdot \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left[ \sum_a \left( \pi_\theta(a|s)^{1-1/p} \cdot |\tilde{A}^{\pi_\theta}(s,a)| \right)^p \right]^{\frac{1}{p}} \tag{C.545}$$

$$\leq \frac{p}{1-\gamma} \cdot \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left[ \sum_a \left( 1 \cdot \frac{1+\tau \log A}{1-\gamma} \right)^p \right]^{\frac{1}{p}} \tag{C.546}$$

$$\leq \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p}. \tag{C.547}$$

Combining Eqs. (C.539) and (C.542), we have,

$$\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p \geq \min_s \|\tilde{\theta}_t(s,\cdot)\|_p \tag{C.548}$$

$$- \xi \cdot \eta \cdot \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} \cdot \frac{1}{\min_s \|\tilde{\theta}_t(s,\cdot)\|_p} \tag{C.549}$$

$$= 1 - \xi \cdot \eta \cdot \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2}. \tag{C.550}$$

$$\left( \|\tilde{\theta}_t(s,\cdot)\|_p = 1, \text{ for all } s, \text{ Algorithm 3} \right) \tag{C.551}$$

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

$$\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p \geq 1 - \lambda_t \cdot \eta \cdot \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} \tag{C.552}$$

$$\text{(by Eq. (C.548))} \tag{C.553}$$

$$= 1 - \lambda_t \cdot \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p}} \cdot \frac{p \cdot A^{1/p}}{(1-\gamma)^2} \tag{C.554}$$

$$\geq 1 - \frac{1-\gamma}{10 \cdot p \cdot A^{1/p}} \tag{C.555}$$

$$= \left(1 - \frac{2}{\sqrt{5}}\right) \cdot \left(1 - \frac{5+2\sqrt{5}}{10} \cdot \frac{1-\gamma}{p \cdot A^{1/p}}\right) + \frac{2}{\sqrt{5}} \tag{C.556}$$

$$\geq \frac{2}{\sqrt{5}}. \qquad \left(p \geq 2,\ A^{1/p} \geq 1,\ 1-\gamma \in (0,1]\right) \tag{C.557}$$

Combining Eqs. (C.535) and (C.552), we have,

$$\left| \tilde{V}^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| \tag{C.558}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{\lambda_t}(s,\cdot)\|_p^2} \tag{C.559}$$

$$\leq \frac{4 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \frac{5}{4} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2 \tag{C.560}$$

$$= \frac{5 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2, \tag{C.561}$$

which implies,

$$\tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_{t+1}}}(\mu) = \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \tilde{V}^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) \tag{C.562}$$

$$\left(\tilde{\theta}_{t+1}(s,a) = \frac{\tilde{\zeta}_{t+1}(s,a)}{\|\tilde{\zeta}_{t+1}(s,\cdot)\|_p}, \text{ Algorithm 3}\right) \tag{C.563}$$

$$\leq -\left\langle \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle + \frac{5 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2 \tag{C.564}$$

(by Eq. (C.558)) (C.565)

$$= -\eta \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 + \frac{5 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \eta^2 \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \tag{C.566}$$

$$\left(\tilde{\zeta}_{t+1} = \tilde{\theta}_t + \eta \cdot \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \text{ Algorithm 3}\right) \tag{C.567}$$

$$= -\frac{(1-\gamma)^3}{20 \cdot p^2 \cdot A^{2/p} + c_\tau} \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \tag{C.568}$$

$$\left(\eta = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p} + c_\tau}\right) \tag{C.569}$$

$$\leq -\frac{(1-\gamma)^3}{20 \cdot \cancel{p^2} \cdot A^{2/p} + c_\tau} \cdot \frac{2\tau}{S} \cdot \min_s \mu(s) \cdot \frac{\cancel{p^2} \cdot \min_{s,a} \pi_{\tilde{\theta}_t}(a|s)^{2-2/p}}{\max_s \|\tilde{\theta}_t(s,\cdot)\|_p^2} \tag{C.570}$$

$$\cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-1} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)\right] \quad \text{(Lemma 48)} \tag{C.571}$$

$$= -\frac{(1-\gamma)^3 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-1} \cdot \min_{s,a} \pi_{\tilde{\theta}_t}(a|s)^{2-2/p} \tag{C.572}$$

$$\cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)\right] \qquad \left(\|\tilde{\theta}_t(s,\cdot)\|_p = 1, \text{ for all } s\right) \tag{C.573}$$

$$\leq -\frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot \min_{s,a} \pi_{\tilde{\theta}_t}(a|s)^{2-2/p} \tag{C.574}$$

$$\cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)\right], \qquad \text{(by Eq. (C.389))} \tag{C.575}$$

which implies,

$$\tilde{V}^{\pi_{\theta_t}}(\mu) - \tilde{V}^{\pi_{\theta_{t+1}}}(\mu) = \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_{t+1}}}(\mu) \qquad \text{(by Lemma 46)} \tag{C.576}$$

$$\leq -\frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot \min_{s,a} \pi_{\theta_t}(a|s)^{2-2/p} \tag{C.577}$$

$$\cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right], \quad \text{(by Eq. (C.562) and Lemma 46)} \tag{C.578}$$

$$\leq -\frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot c^{2-2/p} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right], \tag{C.579}$$

which is equivalent to,

$$\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \tag{C.580}$$

$$\leq \left( 1 - \frac{(1-\gamma)^4 \cdot \tau \cdot c^{2-2/p}}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \right) \cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_{t-1}}}(\mu) \right] \tag{C.581}$$

$$\leq \exp\left\{ -\frac{(1-\gamma)^4 \cdot \tau \cdot c^{2-2/p}}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \right\} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_{t-1}}}(\mu) \right] \tag{C.582}$$

$$\leq \exp\left\{ -\frac{(1-\gamma)^4 \cdot \tau \cdot c^{2-2/p}}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1) \right\} \cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_1}}(\mu) \right] \tag{C.583}$$

$$\leq \exp\left\{ -\frac{(1-\gamma)^4 \cdot \tau \cdot c^{2-2/p}}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1) \right\} \cdot \frac{1 + \tau \log A}{1-\gamma}, \tag{C.584}$$

which leads to the final result,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[ \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] \qquad \text{(by Eq. (C.530))} \tag{C.585}$$

$$\leq \left\| \frac{1}{\mu} \right\|_\infty \cdot \exp\left\{ -\frac{(1-\gamma)^4 \cdot \tau \cdot c'^2}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1) \right\} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2}, \tag{C.586}$$

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

## C.3 Proofs for Section 3.5: Escort Cross Entropy

**Lemma 49** (Non-uniform Smoothness)**.** *Let* $\pi_\theta \coloneqq f_p(\theta)$*, and* $\pi_{\theta'} \coloneqq f_p(\theta')$*. Denote* $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ *with some* $\zeta \in [0, 1]$*. Then for* $p = 2$*, we have* $D_{\mathrm{KL}}(y \| \pi_\theta)$ *is* $\beta$*-smooth, i.e.,*

$$\left| D_{\mathrm{KL}}(y\|\pi_{\theta'}) - D_{\mathrm{KL}}(y\|\pi_\theta) - \left\langle \frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2, \tag{C.587}$$

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

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

$$\frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta} = \frac{d\{-y^\top \log \pi_\theta\}}{d\theta} \tag{C.588}$$

$$= \left( \frac{d\pi_\theta}{d\theta} \right)^\top \left( \frac{d\{-y^\top \log \pi_\theta\}}{d\pi_\theta} \right) \tag{C.589}$$

$$= p \cdot \mathrm{diag}\left( \frac{1}{\theta} \right) \left( \mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) \mathrm{diag}\left( \frac{1}{\pi_\theta} \right) (-y) \tag{C.590}$$

$$= p \cdot \mathrm{diag}\left( \frac{1}{\theta} \right) (\pi_\theta - y). \tag{C.591}$$

Denote the second derivative w.r.t. $\theta$ (i.e., Hessian) as

$$K(y, \theta) = \frac{d}{d\theta} \left\{ \frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta} \right\} \tag{C.592}$$

$$= p \cdot \frac{d}{d\theta} \left\{ \mathrm{diag}\left( \frac{1}{\theta} \right) (\pi_\theta - y) \right\}. \tag{C.593}$$

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

$$K_{i,j} = p \cdot \frac{d\{\frac{\pi_\theta(i) - y(i)}{\theta(i)}\}}{d\theta(j)} \tag{C.594}$$

$$= p \cdot \frac{\frac{p}{\theta(j)} \cdot [\delta_{ij} \pi_\theta(j) - \pi_\theta(i)\pi_\theta(j)] \cdot \theta(i) - (\pi_\theta(i) - y(i)) \cdot \delta_{ij}}{\theta(i)^2} \tag{C.595}$$

$$= p \cdot (p-1) \cdot \delta_{ij} \cdot \frac{\pi_\theta(i)}{\theta(i)^2} - p^2 \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{\pi_\theta(j)}{\theta(j)} + \delta_{ij} \cdot p \cdot \frac{y(i)}{\theta(i)^2}, \tag{C.596}$$

where the $\delta$ notation is defined in Eq. (C.47). For any $x \in \mathbb{R}^K$,

$$\left|x^\top K(y,\theta)x\right| = \left|\sum_{i=1}^K \sum_{j=1}^K K_{i,j}x(i)x(j)\right| \tag{C.597}$$

$$= \Bigg|p\cdot(p-1)\sum_i \frac{\pi_\theta(i)}{\theta(i)^2}\cdot x(i)^2 \tag{C.598}$$

$$- p^2\sum_i \frac{\pi_\theta(i)}{\theta(i)}\cdot x(i)\sum_j \frac{\pi_\theta(j)}{\theta(j)}\cdot x(j) + p\sum_i \frac{y(i)}{\theta(i)^2}\cdot x(i)^2\Bigg| \tag{C.599}$$

$$\le p\cdot(p-1)\cdot\left[\sum_i \frac{\pi_\theta(i)}{\theta(i)^2}\cdot x(i)^2\right] \tag{C.600}$$

$$+ p^2\cdot\left[\sum_i \frac{\pi_\theta(i)}{\theta(i)}\cdot x(i)\right]^2 + p\sum_i \frac{y(i)}{\theta(i)^2}\cdot x(i)^2, \tag{C.601}$$

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

$$\sum_i \frac{\pi_\theta(i)}{\theta(i)^2}\cdot x(i)^2 = \frac{1}{\|\theta\|_p^2}\sum_i \pi_\theta(i)^{1-2/p}\cdot x(i)^2 \tag{C.602}$$

$$\le \frac{1}{\|\theta\|_p^2}\sum_i 1\cdot x(i)^2 \qquad (p=2) \tag{C.603}$$

$$= \frac{1}{\|\theta\|_p^2}\cdot\|x\|_2^2. \tag{C.604}$$

The second term is upper bounded as,

$$\left[\sum_i \frac{\pi_\theta(i)}{\theta(i)}\cdot x(i)\right]^2 \le \sum_i\left(\frac{\pi_\theta(i)}{\theta(i)}\right)^2\cdot\|x\|_2^2 \qquad \text{(by Cauchy-Schwarz)} \tag{C.605}$$

$$= \frac{1}{\|\theta\|_p^2}\cdot\sum_i\left(\pi_\theta(i)^{1-1/p}\right)^2\cdot\|x\|_2^2 \tag{C.606}$$

$$\le \frac{1}{\|\theta\|_p^2}\cdot\left[\sum_i \pi_\theta(i)\right]\cdot\|x\|_2^2 \qquad (p=2) \tag{C.607}$$

$$= \frac{1}{\|\theta\|_p^2}\cdot\|x\|_2^2. \tag{C.608}$$

The last term is upper bounded as,

$$\sum_i \frac{y(i)}{\theta(i)^2}\cdot x(i)^2 \le \left(\max_i \frac{y(i)}{\theta(i)^2}\right)\cdot\|x\|_2^2. \tag{C.609}$$

Combining Eqs. (C.597), (C.602), (C.605) and (C.609), for $p = 2$, for any $x \in \mathbb{R}^K$, we have,

$$\left|x^\top K(y,\theta)x\right| \leq p\cdot(p-1)\cdot\frac{1}{\|\theta\|_p^2}\cdot\|x\|_2^2 \tag{C.610}$$

$$+\,p^2\cdot\frac{1}{\|\theta\|_p^2}\cdot\|x\|_2^2 + p\cdot\left\|\frac{y}{\theta\odot\theta}\right\|_\infty\cdot\|x\|_2^2 \tag{C.611}$$

$$= \frac{6}{\|\theta\|_p^2}\cdot\|x\|_2^2 + 2\cdot\left(\max_i\frac{y(i)}{\theta(i)^2}\right)\cdot\|x\|_2^2. \tag{C.612}$$

According to Taylor's theorem, we have,

$$\left|D_{\mathrm{KL}}(y\|\pi_{\theta'}) - D_{\mathrm{KL}}(y\|\pi_\theta) - \left\langle\frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta}, \theta'-\theta\right\rangle\right| \tag{C.613}$$

$$= \frac{1}{2}\cdot\left|(\theta'-\theta)^\top K(y,\theta_\zeta)(\theta'-\theta)\right| \tag{C.614}$$

$$\leq \left[\frac{3}{\|\theta_\zeta\|_p^2} + \max_i\frac{y(i)}{\theta_\zeta(i)^2}\right]\cdot\|\theta'-\theta\|_2^2. \qquad \square$$

**Lemma 50** (Non-uniform Łojasiewicz)**.** *Let $\pi_\theta = f_p(\theta)$. For any $p \geq 2$, we have,*

$$\left\|\frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta}\right\|_2^2 \geq \frac{p^2}{\|\theta\|_p^2}\cdot\min_a\pi_\theta(a)^{1-2/p}\cdot D_{\mathrm{KL}}(y\|\pi_\theta). \tag{C.615}$$

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

$$D_{\mathrm{KL}}(y\|\pi_\theta) = \sum_a y(a)\cdot\log\left(\frac{y(a)}{\pi_\theta(a)}\right) \tag{C.616}$$

$$\leq \sum_a y(a)\cdot\left(\frac{y(a)}{\pi_\theta(a)}-1\right) \qquad (\log x\leq x-1) \tag{C.617}$$

$$= \sum_a \left(y(a)-\pi_\theta(a)+\pi_\theta(a)\right)\cdot\frac{y(a)-\pi_\theta(a)}{\pi_\theta(a)} \tag{C.618}$$

$$= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\pi_\theta(a)} \tag{C.619}$$

$$= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\pi_\theta(a)^{2/p}}\cdot\frac{1}{\pi_\theta(a)^{1-2/p}} \tag{C.620}$$

$$= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\theta(a)^2}\cdot\|\theta\|_p^2\cdot\frac{1}{\pi_\theta(a)^{1-2/p}} \tag{C.621}$$

$$\left(\pi_\theta(a)=\frac{|\theta(a)|^p}{\sum_{a'}|\theta(a')|^p}\right) \tag{C.622}$$

$$\leq \|\theta\|_p^2\cdot\frac{1}{\min_a \pi_\theta(a)^{1-2/p}}\cdot\sum_a\frac{\left(y(a)-\pi_\theta(a)\right)^2}{\theta(a)^2} \tag{C.623}$$

$$= \|\theta\|_p^2\cdot\frac{1}{\min_a \pi_\theta(a)^{1-2/p}}\cdot\frac{1}{p^2}\cdot\left\|p\cdot\mathrm{diag}\left(\frac{1}{\theta}\right)(y-\pi_\theta)\right\|_2^2. \tag{C.624}$$

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 15.** Using the escort transform with $p=2$ on the cross entropy objective, we obtain for all $t\geq 1$,
**(gradient flow)** with $\eta_t = \frac{\|\theta_t\|_p^2}{p^2}$,

$$D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \leq D_{\mathrm{KL}}(y\|\pi_{\theta_1})\cdot e^{-(t-1)}, \tag{C.625}$$

**(gradient ascent)** with $\eta_t = \frac{\|\theta_t\|_p^2}{4\cdot(3+c_1^2)}$,

$$-\log\pi_{\theta_t}(a_y) = D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \tag{C.626}$$

$$\leq D_{\mathrm{KL}}(y\|\pi_{\theta_1})\cdot\exp\left\{-\frac{(t-1)}{2\cdot(3+c_1^2)}\right\}, \tag{C.627}$$

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

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

$$\frac{d\theta_t}{dt} = -\eta_t \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}. \tag{C.628}$$

Then we have,

$$\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{dt} = \left(\frac{d\theta_t}{dt}\right)^\top \left(\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}\right) \tag{C.629}$$

$$= -\eta_t \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 \qquad \text{(by Eq. (C.628))} \tag{C.630}$$

$$\leq -\eta_t \cdot \frac{p^2}{\|\theta_t\|_p^2} \cdot \min_a \pi_{\theta_t}(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \qquad \text{(by Lemma 50)} \tag{C.631}$$

$$= -\min_a \pi_{\theta_t}(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{p^2}\right) \tag{C.632}$$

$$= -D_{\mathrm{KL}}(y\|\pi_{\theta_t}), \qquad (p=2) \tag{C.633}$$

which implies,

$$\frac{d\left\{\log D_{\mathrm{KL}}(y\|\pi_{\theta_t})\right\}}{dt} = \frac{1}{D_{\mathrm{KL}}(y\|\pi_{\theta_t})} \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{dt} \leq -1. \tag{C.634}$$

Taking integral, we have,

$$\log D_{\mathrm{KL}}(y\|\pi_{\theta_t}) - \log D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \leq -(t-1), \tag{C.635}$$

which is equivalent to

$$D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \leq D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \cdot e^{-(t-1)}. \tag{C.636}$$

**Second part.** For the gradient descent, according to Lemma 49, we have,

$$D_{\mathrm{KL}}(y\|\pi_{\theta_{t+1}}) - D_{\mathrm{KL}}(y\|\pi_{\theta_t}) - \left\langle \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{C.637}$$

$$\leq \frac{\beta}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{C.638}$$

where

$$\beta = \frac{6}{\|\theta_{\zeta_t}\|_p^2} + 2 \cdot \left(\max_i \frac{y(i)}{\theta_{\zeta_t}(i)^2}\right) \tag{C.639}$$

$$= \frac{6}{\|\theta_{\zeta_t}\|_p^2} + \frac{2}{\theta_{\zeta_t}(a_y)^2}, \qquad (y \text{ is one-hot}) \tag{C.640}$$

and

$$\theta_{\zeta_t} \coloneqq \theta_t + \zeta_t \cdot (\theta_{t+1} - \theta_t) = \theta_t - \zeta_t \cdot \eta_t \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}. \tag{C.641}$$

The $\ell_p$ gradient norm is upper bounded as,

$$\left\|\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}\right\|_p = \left[\sum_a \left|\frac{\pi_{\theta_t}(a) - y(a)}{\theta_t(a)}\right|^p\right]^{\frac{1}{p}} \tag{C.642}$$

$$= \left[\sum_{a\neq a_y} \left|\frac{\pi_{\theta_t}(a)}{\theta_t(a)}\right|^p + \left|\frac{\pi_{\theta_t}(a_y) - 1}{\theta_t(a_y)}\right|^p\right]^{\frac{1}{p}} \qquad (y(a_y) = 1) \tag{C.643}$$

$$\leq \left[\sum_{a\neq a_y} \left|\frac{\pi_{\theta_t}(a)}{\theta_t(a)}\right|^p + \frac{1}{|\theta_t(a_y)|^p}\right]^{\frac{1}{p}} \qquad (\pi_{\theta_t}(a_y) \in (0, 1]) \tag{C.644}$$

$$= \left[\frac{1}{\|\theta_t\|_p^p} \sum_{a\neq a_y} \pi_{\theta_t}(a)^{p-1} + \frac{1}{|\theta_t(a_y)|^p}\right]^{\frac{1}{p}} \tag{C.645}$$

$$\leq \left[\frac{1}{\|\theta_t\|_p^p} + \frac{1}{|\theta_t(a_y)|^p}\right]^{\frac{1}{p}} \qquad (p = 2) \tag{C.646}$$

$$\leq \frac{1}{\|\theta_t\|_p} + \frac{1}{|\theta_t(a_y)|}. \qquad \left(\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}\right) \tag{C.647}$$

Next, we have,

$$\theta_{t+1}(a_y) = \theta_t(a_y) - \eta_t \cdot \frac{p}{\theta_t(a_y)} \cdot (\pi_{\theta_t}(a_y) - 1) \tag{C.648}$$

$$\begin{cases} \geq \theta_t(a_y), & \text{if } \theta_t(a_y) > 0, \\ \leq \theta_t(a_y), & \text{if } \theta_t(a_y) < 0. \end{cases} \tag{C.649}$$

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,

$$\theta_{t+1}(a) = \theta_t(a) - \eta_t \cdot \frac{p}{\theta_t(a)} \cdot \pi_{\theta_t}(a) \tag{C.650}$$

$$\begin{cases} \leq \theta_t(a_y), & \text{if } \theta_t(a_y) > 0, \\ \geq \theta_t(a_y), & \text{if } \theta_t(a_y) < 0. \end{cases} \tag{C.651}$$

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

have, for all $t \geq 1$,

$$\frac{|\theta_t(a_y)|}{\|\theta_t\|_p} = \frac{|\theta_t(a_y)|}{\left(\sum_a |\theta_t(a)|^p\right)^{1/p}} \geq \frac{|\theta_t(a_y)|}{\left(\sum_{a \neq a_y} |\theta_1(a)|^p + |\theta_t(a_y)|^p\right)^{1/p}} \tag{C.652}$$

$$\geq \frac{|\theta_1(a_y)|}{\left(\sum_{a \neq a_y} |\theta_1(a)|^p + |\theta_1(a_y)|^p\right)^{1/p}} \tag{C.653}$$

$$= \frac{|\theta_1(a_y)|}{\|\theta_1\|_p} = \frac{1}{c_1}. \tag{C.654}$$

Combining Eqs. (C.642) and (C.652), we have,

$$\left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p \leq \frac{1}{\|\theta_t\|_p} + \frac{1}{|\theta_t(a_y)|} \leq \frac{1}{\|\theta_t\|_p} \cdot (1 + c_1)\,. \tag{C.655}$$

Then we have,

$$\|\theta_{\zeta_t}\|_p = \left\| \theta_t - \zeta_t \cdot \eta_t \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p \qquad \text{(by Eq. (C.641))} \tag{C.656}$$

$$\geq \|\theta_t\|_p - \zeta_t \cdot \eta_t \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p \qquad \text{(by triangle inequality)} \tag{C.657}$$

$$\geq \|\theta_t\|_p - \zeta_t \cdot \eta_t \cdot \frac{1}{\|\theta_t\|_p} \cdot (1 + c_1)\,. \qquad \text{(by Eq. (C.655))} \tag{C.658}$$

$$= \|\theta_t\|_p \cdot \left[1 - \zeta_t \cdot \frac{1 + c_1}{4 \cdot (3 + c_1^2)}\right] \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{4 \cdot (3 + c_1^2)}\right) \tag{C.659}$$

$$\geq \|\theta_t\|_p \cdot \left[1 - \frac{1 + c_1}{4 \cdot (3 + c_1^2)}\right] \qquad (\zeta_t \in [0, 1]) \tag{C.660}$$

$$= \|\theta_t\|_p \cdot \left[\left(1 - \frac{1}{\sqrt{2}}\right) \cdot \left(1 - \frac{\sqrt{2} + 1}{2\sqrt{2}} \cdot \frac{1 + c_1}{3 + c_1^2}\right) + \frac{1}{\sqrt{2}}\right] \tag{C.661}$$

$$\geq \frac{\|\theta_t\|_p}{\sqrt{2}}. \qquad (1/c_1 \in (0, 1],\ c_1 \geq 1) \tag{C.662}$$

Similar to Eq. (C.652), we have,

$$\beta = \frac{6}{\|\theta_{\zeta_t}\|_p^2} + \frac{2}{\theta_{\zeta_t}(a_y)^2} \qquad \text{(by Eq. (C.639))} \tag{C.663}$$

$$\leq \frac{1}{\|\theta_{\zeta_t}\|_p^2} \cdot \left(6 + 2 \cdot c_1^2\right). \tag{C.664}$$

Combining the results, we have,

$$D_{\mathrm{KL}}(y\|\pi_{\theta_{t+1}}) - D_{\mathrm{KL}}(y\|\pi_{\theta_t}) - \left\langle \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{C.665}$$

$$\leq \frac{1}{2} \cdot \frac{1}{\|\theta_{\zeta_t}\|_p^2} \cdot \left(6 + 2 \cdot c_1^2\right) \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{C.666}$$

$$\text{(by Eqs. (C.637) and (C.663))} \tag{C.667}$$

$$\leq \frac{2}{\|\theta_t\|_p^2} \cdot \left(3 + c_1^2\right) \cdot \|\theta_{t+1} - \theta_t\|_2^2, \qquad \text{(by Eq. (C.656))} \tag{C.668}$$

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

$$D_{\mathrm{KL}}(y\|\pi_{\theta_{t+1}}) - D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \tag{C.669}$$

$$\leq -\eta_t \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 \tag{C.670}$$

$$+ \frac{2 \cdot (3 + c_1^2)}{\|\theta_t\|_p^2} \cdot \eta_t^2 \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 \tag{C.671}$$

$$= -\frac{\|\theta_t\|_p^2}{8 \cdot (3 + c_1^2)} \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{4 \cdot (3 + c_1^2)}\right) \tag{C.672}$$

$$\leq -\frac{\cancel{\|\theta_t\|_p^2}}{8 \cdot (3 + c_1^2)} \cdot \frac{p^2}{\cancel{\|\theta_t\|_p^2}} \cdot \min_a \pi_{\theta_t}(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \tag{C.673}$$

$$\text{(by Lemma 50)} \tag{C.674}$$

$$= -\frac{1}{2 \cdot (3 + c_1^2)} \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}), \qquad (p = 2) \tag{C.675}$$

which is equivalent to,

$$D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \leq \left[1 - \frac{1}{2 \cdot (3 + c_1^2)}\right] \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_{t-1}}) \tag{C.676}$$

$$\leq D_{\mathrm{KL}}(y\|\pi_{\theta_{t-1}}) \cdot \exp\left\{ -\frac{1}{2 \cdot (3 + c_1^2)} \right\} \tag{C.677}$$

$$\leq D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \cdot \exp\left\{ -\frac{(t-1)}{2 \cdot (3 + c_1^2)} \right\}, \tag{C.678}$$

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

## C.4 Miscellaneous Extra Supporting Results

**Lemma 51.** *Let $\pi \in \Delta(\mathcal{A})$ and $q \geq 0$. For any vector $x \in \mathbb{R}^K$, we have,*

$$\left\| \mathit{diag}(q) \left(\mathbf{Id} - \mathbf{1}\pi^\top\right) \left(x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1}\right) \right\|_2 \geq \min_a q(a) \cdot \left\| x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2. \tag{C.679}$$

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

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_K. \tag{C.680}$$

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

$$G^\top G\mathbf{1} = G^\top \mathrm{diag}(q)\left(\mathbf{Id} - \mathbf{1}\pi^\top\right)\mathbf{1} \tag{C.681}$$

$$= G^\top \mathrm{diag}(q)\left(\mathbf{1} - \mathbf{1}\right) = 0 \cdot \mathbf{1}, \tag{C.682}$$

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,

$$x^\top G^\top G x = \|Gx\|_2^2 \ge 0, \tag{C.683}$$

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 $G^\top G$,

$$x = a_1 \cdot \frac{\mathbf{1}}{\sqrt{K}} + a_2 \cdot v_2 + \cdots + a_K \cdot v_K \tag{C.684}$$

$$= \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} + a_2 \cdot v_2 + \cdots + a_K \cdot v_K. \tag{C.685}$$

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

$$a_1 = x^\top \frac{\mathbf{1}}{\sqrt{K}} = \frac{x^\top \mathbf{1}}{\sqrt{K}}. \tag{C.686}$$

Denote

$$x' = x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} = a_2 \cdot v_2 + \cdots + a_K \cdot v_K. \tag{C.687}$$

We have,

$$\|x'\|_2^2 = a_2^2 + \cdots + a_K^2. \tag{C.688}$$

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

$$G^\top G x' = a_2 \cdot \lambda_2 \cdot v_2 + \cdots + a_K \cdot \lambda_K \cdot v_K. \tag{C.689}$$

Therefore we have,

$$\|Gx'\|_2 = \left(\|Gx'\|_2^2\right)^{\frac{1}{2}} = \left(x'^{\top} G^{\top} G x'\right)^{\frac{1}{2}} \tag{C.690}$$

$$= \left(a_2^2 \cdot \lambda_2 + \cdots + a_K^2 \cdot \lambda_K\right)^{\frac{1}{2}} \tag{C.691}$$

$$\geq \left(a_2^2 \cdot \lambda_2 + \cdots + a_K^2 \cdot \lambda_2\right)^{\frac{1}{2}} \tag{C.692}$$

$$= \sqrt{\lambda_2} \cdot \|x'\|_2. \qquad \text{(by Eq. (C.688))} \tag{C.693}$$

Next, we have,

$$\lambda_2 = \frac{v_2^{\top} G^{\top} G v_2}{v_2^{\top} v_2} = \frac{1}{v_2^{\top} v_2} \cdot \|G v_2\|_2^2 \tag{C.694}$$

$$= \frac{1}{v_2^{\top} v_2} \cdot \left\|\mathrm{diag}(q)\left(v_2 - \pi^{\top} v_2 \cdot \mathbf{1}\right)\right\|_2^2 \tag{C.695}$$

$$= \frac{1}{v_2^{\top} v_2} \cdot \left[\sum_{a=1}^{K} q(a)^2 \cdot \left(v_2(a) - \pi^{\top} v_2\right)^2\right] \tag{C.696}$$

$$\geq \frac{1}{v_2^{\top} v_2} \cdot \min_a q(a)^2 \cdot \left\|v_2 - \pi^{\top} v_2 \cdot \mathbf{1}\right\|_2^2 \tag{C.697}$$

$$= \min_a q(a)^2 \cdot \frac{v_2^{\top} v_2 + K \cdot (\pi^{\top} v_2)^2}{v_2^{\top} v_2} \qquad \left(v_2^{\top} \mathbf{1} = 0\right) \tag{C.698}$$

$$\geq \min_a q(a)^2. \tag{C.699}$$

Combining Eqs. (C.690) and (C.694), we have,

$$\left\|\mathrm{diag}(q)\left(\mathbf{Id} - \mathbf{1}\pi^{\top}\right)\left(x - \frac{x^{\top}\mathbf{1}}{K} \cdot \mathbf{1}\right)\right\|_2 = \|Gx'\|_2 \tag{C.700}$$

$$\geq \sqrt{\lambda_2} \cdot \|x'\|_2 \tag{C.701}$$

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

# Appendix D

# Proofs for Chapter 4: Non-uniform Analysis

## D.1 Proofs for Section 4.5: Non-uniform Analysis for General Optimization

**Theorem 16.** Suppose $f : \Theta \to \mathbb{R}$ satisfies NS with $\beta(\theta)$ and the NŁ inequality with $(C(\theta), \xi)$. Suppose $C \coloneqq \inf_{t\geq 1} C(\theta_t) > 0$ for GD and GNGD. Let $\delta(\theta) \coloneqq f(\theta) - f(\theta^*)$ be the sub-optimality gap. The following hold:

**(1a)** if $\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}$ with $\xi \in (-\infty, 1/2)$, then the conclusions of (1b) hold;

**(1b)** if $\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$ with $\xi \in (-\infty, 1/2)$, then GD with $\eta \in O(1)$ achieves $\delta(\theta_t) \in \Theta(1/t^{\frac{1}{1-2\xi}})$, and GNGD achieves $\delta(\theta_t) \in O(e^{-c' \cdot t})$.

**(2a)** if $\beta(\theta) \leq L_0 + L_1 \cdot \|\nabla f(\theta)\|_2$, then the conclusions of (2b) hold;

**(2b)** if $\beta(\theta) \leq L_0 \cdot \frac{\|\nabla f(\theta)\|^2}{\delta(\theta)^{2-2\xi}} + L_1 \cdot \|\nabla f(\theta)\|_2$, then GD and GNGD both achieve $\delta(\theta_t) \in O(1/t^{\frac{1}{1-2\xi}})$ when $\xi \in (-\infty, 1/2)$, and $O(e^{-c' \cdot t})$ when $\xi = 1/2$. GNGD has strictly better constant than GD ($1 > C > C^2$).

**(3a)** if $\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}$ with $\xi \in (1/2, 1)$, then the conclusions of (3b) hold;

**(3b)** if $\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}$ with $\xi \in (1/2, 1)$, then GD with $\eta \in \Theta(1)$ does not converge, while GNGD achieves $\delta(\theta_t) \in O(e^{-c' \cdot t})$.

*Proof.* **(1a) First part:** *$O(1/t^{\frac{1}{1-2\xi}})$ upper bound for GD update $\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla f(\theta_t)$ with $\eta \in O(1)$.*

We show that using GD with learning rate $\eta = \frac{1}{c \cdot \delta(\theta_1)^{1-2\xi}}$, the sub-optimality $\delta(\theta_t)$ is monotonically decreasing. And thus there exists a universal constant $\beta > 0$ such that $\beta(\theta_t) \leq \beta$, for all $t \geq 1$.

Denote $\beta \coloneqq c \cdot \delta(\theta_1)^{1-2\xi}$. We have $\beta \in (0, \infty)$, since $f(\theta^*) > -\infty$, and $f(\theta^*) < f(\theta_1) < \infty$. By assumption, we have $\beta(\theta_1) \leq \beta$. According to Lemma 33, using GD with $\eta = \frac{1}{\beta}$, we have,

$$\delta(\theta_2) - \delta(\theta_1) = f(\theta_2) - f(\theta_1) \leq 0. \tag{D.1}$$

Therefore, we have,

$$\begin{aligned}
\beta(\theta_2) &\leq c \cdot \delta(\theta_2)^{1-2\xi} && \text{(by assumption)} && \text{(D.2)} \\
&\leq c \cdot \delta(\theta_1)^{1-2\xi} && (0 < \delta(\theta_2) \leq \delta(\theta_1) \text{ and } \xi < 1/2) && \text{(D.3)} \\
&= \beta. && && \text{(D.4)}
\end{aligned}$$

Repeating similar arguments of Eqs. (D.1) and (D.2), we have, for all $t \geq 1$, $\beta(\theta_t) \leq \beta$ and,

$$0 < \delta(\theta_{t+1}) \leq \delta(\theta_t). \tag{D.5}$$

Therefore, we have, for all $t \geq 1$ (or using Lemma 33),

$$\begin{aligned}
\delta(\theta_{t+1}) - \delta(\theta_t) &= f(\theta_{t+1}) - f(\theta_t) && && \text{(D.6)} \\
&\leq \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{\beta(\theta_t)}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 && \text{(NS)} && \text{(D.7)} \\
&\leq \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{\beta}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 && (\beta(\theta_t) \leq \beta) && \text{(D.8)} \\
&= -\frac{1}{2\beta} \cdot \|\nabla f(\theta_t)\|_2^2 && \left(\theta_{t+1} \leftarrow \theta_t - \frac{1}{\beta} \cdot \nabla f(\theta_t)\right) && \text{(D.9)} \\
&\leq -\frac{1}{2\beta} \cdot C(\theta_t)^2 \cdot \delta(\theta_t)^{2-2\xi} && \text{(NŁ)} && \text{(D.10)} \\
&\leq -\frac{1}{2\beta} \cdot C^2 \cdot \delta(\theta_t)^{2-2\xi}. && \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) && \text{(D.11)}
\end{aligned}$$

According to Lemma 53, given any $\alpha > 0$, we have, for all $x \in [0, 1]$,

$$\frac{1}{\alpha} \cdot (1 - x^\alpha) \geq x^\alpha \cdot (1 - x). \tag{D.12}$$

Let $\alpha = 1-2\xi > 0$, since $\xi < 1/2$. Also let $x = \frac{\delta(\theta_{t+1})}{\delta(\theta_t)} \in (0, 1]$ due to Eq. (D.5). We have,

$$\frac{1}{1-2\xi} \cdot \left[1 - \frac{\delta(\theta_{t+1})^{1-2\xi}}{\delta(\theta_t)^{1-2\xi}}\right] \geq \frac{\delta(\theta_{t+1})^{1-2\xi}}{\delta(\theta_t)^{1-2\xi}} \cdot \left[1 - \frac{\delta(\theta_{t+1})}{\delta(\theta_t)}\right]. \tag{D.13}$$

Next, we have,

$$\begin{aligned}
\frac{1}{\delta(\theta_t)^{1-2\xi}} &= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \frac{1}{\delta(\theta_t)^{1-2\xi}} - \frac{1}{\delta(\theta_1)^{1-2\xi}} && \text{(D.14)}\\
&= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t-1}\left[\frac{1}{\delta(\theta_{s+1})^{1-2\xi}} - \frac{1}{\delta(\theta_s)^{1-2\xi}}\right] && \text{(D.15)}\\
&= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t-1}\frac{1-2\xi}{\delta(\theta_{s+1})^{1-2\xi}} \cdot \frac{1}{1-2\xi} \cdot \left[1 - \frac{\delta(\theta_{s+1})^{1-2\xi}}{\delta(\theta_s)^{1-2\xi}}\right] && \text{(D.16)}\\
&\geq \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t-1}\frac{1-2\xi}{\cancel{\delta(\theta_{s+1})^{1-2\xi}}} \cdot \frac{\cancel{\delta(\theta_{s+1})^{1-2\xi}}}{\delta(\theta_s)^{1-2\xi}} \cdot \left[1 - \frac{\delta(\theta_{s+1})}{\delta(\theta_s)}\right] && \text{(D.17)}\\
&\quad \text{(by Eq. (D.13))} && \text{(D.18)}\\
&= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t-1}\frac{1-2\xi}{\delta(\theta_s)^{2-2\xi}} \cdot [\delta(\theta_s) - \delta(\theta_{s+1})] && \text{(D.19)}\\
&\geq \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t-1}\frac{1-2\xi}{\cancel{\delta(\theta_s)^{2-2\xi}}} \cdot \frac{C^2}{2\beta} \cdot \cancel{\delta(\theta_s)^{2-2\xi}} \quad \text{(by Eq. (D.6))} && \text{(D.20)}\\
&= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \frac{(1-2\xi)\cdot C^2}{2\beta} \cdot (t-1), && \text{(D.21)}
\end{aligned}$$

which implies for all $t \geq 1$,

$$\begin{aligned}
f(\theta_t) - f(\theta^*) &= \delta(\theta_t) && \text{(D.22)}\\
&\leq \left[\frac{1}{(f(\theta_1) - f(\theta^*))^{1-2\xi}} + \frac{(1-2\xi)\cdot C^2}{2\beta} \cdot (t-1)\right]^{-\frac{1}{1-2\xi}} && \text{(D.23)}\\
&\in O\left(\frac{1}{t^{\frac{1}{1-2\xi}}}\right). && \text{(D.24)}
\end{aligned}$$

**(1a) Second part:** *$\Omega(1/t^{\frac{1}{1-2\xi}})$ lower bound for GD update $\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \nabla f(\theta_t)$ with $\eta_t \in (0, 1]$.*

According to the NS property of Definition 6, we have, for all $\theta$ and $\theta'$,

$$f(\theta') \leq f(\theta) + \nabla f(\theta)^\top (\theta' - \theta) + \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2. \tag{D.25}$$

Fix $\theta$ and take minimum over $\theta'$ on both sides of the above inequality. Then we have,

$$f(\theta^*) \leq f(\theta) + \min_{\theta'} \left\{ \nabla f(\theta)^\top (\theta' - \theta) + \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2 \right\} \tag{D.26}$$

$$= f(\theta) - \frac{1}{\beta(\theta)} \cdot \|\nabla f(\theta)\|_2^2 + \frac{1}{2 \cdot \beta(\theta)} \cdot \|\nabla f(\theta)\|_2^2 \tag{D.27}$$

$$\left( \theta' = \theta - \frac{1}{\beta(\theta)} \cdot \nabla f(\theta) \right) \tag{D.28}$$

$$= f(\theta) - \frac{1}{2 \cdot \beta(\theta)} \cdot \|\nabla f(\theta)\|_2^2, \tag{D.29}$$

which implies,

$$\|\nabla f(\theta)\|_2^2 \leq 2 \cdot \beta(\theta) \cdot \delta(\theta) \tag{D.30}$$

$$\leq 2 \cdot c \cdot \delta(\theta)^{2-2\xi}. \qquad \left(\beta(\theta) \leq c \cdot \delta(\theta)^{1-2\xi}\right) \tag{D.31}$$

Therefore, we have,

$$\delta(\theta_t) - \delta(\theta_{t+1}) \tag{D.32}$$

$$= f(\theta_t) - f(\theta_{t+1}) + \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) - \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) \tag{D.33}$$

$$\leq \frac{\beta}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 - \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) \tag{D.34}$$

$$\text{(by NS and } \beta(\theta_t) \leq \beta) \tag{D.35}$$

$$= \left( \frac{\beta}{2} \cdot \eta_t^2 + \eta_t \right) \cdot \|\nabla f(\theta_t)\|_2^2 \qquad (\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \nabla f(\theta_t)) \tag{D.36}$$

$$\leq \left( \frac{\beta}{2} \cdot \eta_t^2 + \eta_t \right) \cdot 2 \cdot c \cdot \delta(\theta_t)^{2-2\xi} \qquad \text{(by Eq. (D.30))} \tag{D.37}$$

$$\leq (\beta + 2) \cdot c \cdot \delta(\theta_t)^{2-2\xi}. \qquad (\eta_t \in (0, 1]) \tag{D.38}$$

Next, we show that $\frac{\delta(\theta_{t+1})}{\delta(\theta_t)} \geq \frac{3-4\xi}{4-4\xi}$ holds for all large enough $t \geq 1$ by contradiction. According to the upper bound results in the first part, we have $\delta(\theta_t) \to 0$ as $t \to \infty$. Suppose $\frac{\delta(\theta_{t+1})}{\delta(\theta_t)} < \frac{3-4\xi}{4-4\xi}$, where $t \geq 1$ is large enough and $\delta(\theta_t)$ is small enough. We have,

$$\delta(\theta_{t+1}) \geq \delta(\theta_t) - (\beta + 2) \cdot c \cdot \delta(\theta_t)^{2-2\xi} \qquad \text{(by Eq. (D.32))} \tag{D.39}$$

$$> \frac{4-4\xi}{3-4\xi} \cdot \delta(\theta_{t+1}) - (\beta + 2) \cdot c \cdot \left( \frac{4-4\xi}{3-4\xi} \right)^{2-2\xi} \cdot \delta(\theta_{t+1})^{2-2\xi}, \tag{D.40}$$

where the last inequality is because of the function $f : x \mapsto x - a \cdot x^{2-2\xi}$ with $a > 0$ is monotonically increasing for all $0 < x \leq \frac{1}{[(2-2\xi)a]^{1/(1-2\xi)}}$. Eq. (D.39) implies that,

$$\delta(\theta_{t+1})^{1-2\xi} > \frac{1}{3-4\xi} \cdot \frac{1}{(\beta+2)\cdot c} \cdot \left(\frac{3-4\xi}{4-4\xi}\right)^{2-2\xi}, \tag{D.41}$$

for large enough $t \geq 1$, which is a contradiction with $\delta(\theta_t) \to 0$ as $t \to \infty$. Thus we have $\frac{\delta(\theta_{t+1})}{\delta(\theta_t)} \geq \frac{3-4\xi}{4-4\xi}$ holds for all large enough $t \geq 1$. Denote

$$t_0 \coloneqq \min\left\{t \geq 1 : \frac{\delta(\theta_{s+1})}{\delta(\theta_s)} \geq \frac{3-4\xi}{4-4\xi}, \text{ for all } s \geq t\right\}. \tag{D.42}$$

According to Lemma 54, given any $\alpha > 0$, we have, for all $x \in \left[\frac{2\alpha+1}{2\alpha+2}, 1\right]$,

$$\frac{1}{2\alpha} \cdot (1 - x^{\alpha}) \leq x^{\alpha} \cdot (1 - x). \tag{D.43}$$

Let $\alpha = 1-2\xi > 0$, since $\xi < 1/2$. We have $\frac{2\alpha+1}{2\alpha+2} = \frac{3-4\xi}{4-4\xi}$. Also let $x = \frac{\delta(\theta_{t+1})}{\delta(\theta_t)} \in \left[\frac{3-4\xi}{4-4\xi}, 1\right]$. We have,

$$\frac{1}{2\cdot(1-2\xi)} \cdot \left[1 - \frac{\delta(\theta_{t+1})^{1-2\xi}}{\delta(\theta_t)^{1-2\xi}}\right] \leq \frac{\delta(\theta_{t+1})^{1-2\xi}}{\delta(\theta_t)^{1-2\xi}} \cdot \left[1 - \frac{\delta(\theta_{t+1})}{\delta(\theta_t)}\right], \tag{D.44}$$

for all $t \geq t_0$. On the other hand, since $t_0 \in O(1)$ and $1 - 2\xi > 0$, we have, for all $t < t_0$,

$$\delta(\theta_{t+1})^{1-2\xi} \geq c_0 > 0. \tag{D.45}$$

Next, we have, for all $t \geq t_0$,

$$\frac{1}{\delta(\theta_t)^{1-2\xi}} = \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t-1}\left[\frac{1}{\delta(\theta_{s+1})^{1-2\xi}} - \frac{1}{\delta(\theta_s)^{1-2\xi}}\right] \tag{D.46}$$

$$= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t_0-1}\frac{1}{\delta(\theta_{s+1})^{1-2\xi}} \cdot \left[1 - \frac{\delta(\theta_{s+1})^{1-2\xi}}{\delta(\theta_s)^{1-2\xi}}\right] \tag{D.47}$$

$$+ \sum_{s=t_0}^{t-1}\frac{2\cdot(1-2\xi)}{\delta(\theta_{s+1})^{1-2\xi}} \cdot \frac{1}{2\cdot(1-2\xi)} \cdot \left[1 - \frac{\delta(\theta_{s+1})^{1-2\xi}}{\delta(\theta_s)^{1-2\xi}}\right] \tag{D.48}$$

$$\leq \frac{1}{\delta(\theta_1)^{1-2\xi}} + \sum_{s=1}^{t_0-1}\frac{1}{c_0}\cdot 1 \tag{D.49}$$

$$+ \sum_{s=t_0}^{t-1}\frac{2\cdot(1-2\xi)}{\cancel{\delta(\theta_{s+1})^{1-2\xi}}} \cdot \frac{\cancel{\delta(\theta_{s+1})^{1-2\xi}}}{\delta(\theta_s)^{1-2\xi}} \cdot \left[1 - \frac{\delta(\theta_{s+1})}{\delta(\theta_s)}\right] \tag{D.50}$$

(by Eq. (D.44)) (D.51)

$$= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \frac{t_0-1}{c_0} + \sum_{s=t_0}^{t-1}\frac{2\cdot(1-2\xi)}{\delta(\theta_s)^{2-2\xi}} \cdot [\delta(\theta_s) - \delta(\theta_{s+1})] \tag{D.52}$$

$$\leq \frac{1}{\delta(\theta_1)^{1-2\xi}} + \frac{t_0-1}{c_0} + \sum_{s=t_0}^{t-1}\frac{2\cdot(1-2\xi)}{\cancel{\delta(\theta_s)^{2-2\xi}}} \cdot (\beta+2)\cdot c \cdot \cancel{\delta(\theta_s)^{2-2\xi}} \tag{D.53}$$

(by Eq. (D.32)) (D.54)

$$= \frac{1}{\delta(\theta_1)^{1-2\xi}} + \frac{t_0-1}{c_0} + 2\cdot(1-2\xi)\cdot(\beta+2)\cdot c\cdot(t-t_0), \tag{D.55}$$

which implies for all large enough $t \geq 1$,

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.56}$$

$$\geq \left[\frac{1}{(f(\theta_1) - f(\theta^*))^{1-2\xi}} + \frac{t_0-1}{c_0}\right. \tag{D.57}$$

$$\left. + 2\cdot(1-2\xi)\cdot(\beta+2)\cdot c\cdot(t-t_0)\right]^{-\frac{1}{1-2\xi}} \tag{D.58}$$

$$\in \Omega\left(\frac{1}{t^{\frac{1}{1-2\xi}}}\right). \tag{D.59}$$

**(1a) Third part:** *$O(e^{-c'\cdot t})$ upper bound for GNGD update $\theta_{t+1} \leftarrow \theta_t - \frac{\nabla f(\theta_t)}{\beta(\theta_t)}$.*

We have, for all $t \geq 1$ (or using Lemma 52),

$$\delta(\theta_{t+1}) - \delta(\theta_t) = f(\theta_{t+1}) - f(\theta_t) \tag{D.60}$$

$$\leq \nabla f(\theta_t)^\top (\theta_{t+1} - \theta_t) + \frac{\beta(\theta_t)}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \qquad \text{(NS)} \tag{D.61}$$

$$= -\frac{1}{\beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 + \frac{1}{2} \cdot \frac{1}{\beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.62}$$

$$\left(\theta_{t+1} \leftarrow \theta_t - \frac{\nabla f(\theta_t)}{\beta(\theta_t)}\right) \tag{D.63}$$

$$= -\frac{1}{2 \cdot \beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.64}$$

$$\leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot C(\theta_t)^2 \cdot \delta(\theta_t)^{2-2\xi} \qquad \text{(NŁ)} \tag{D.65}$$

$$\leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot C^2 \cdot \delta(\theta_t)^{2-2\xi} \qquad \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.66}$$

$$\leq -\frac{C^2}{2 \cdot c} \cdot \delta(\theta_t), \qquad \left(\beta(\theta_t) \leq c \cdot \delta(\theta_t)^{1-2\xi}\right) \tag{D.67}$$

which implies for all $t \geq 1$,

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \leq \left(1 - C^2/(2 \cdot c)\right) \cdot \delta(\theta_{t-1}) \tag{D.68}$$

$$\leq \exp\left\{-C^2/(2 \cdot c)\right\} \cdot \delta(\theta_{t-1}) \tag{D.69}$$

$$\leq \exp\left\{-(t-1) \cdot C^2/(2 \cdot c)\right\} \cdot \delta(\theta_1) \tag{D.70}$$

$$= \exp\left\{-(t-1) \cdot C^2/(2 \cdot c)\right\} \cdot \left(f(\theta_1) - f(\theta^*)\right). \tag{D.71}$$

**(1b) First part:** *$O(1/t^{\frac{1}{1-2\xi}})$ upper bound for GD update $\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla f(\theta_t)$ with $\eta \in O(1)$.*

Denote $\beta_1 \coloneqq c \cdot \|\nabla f(\theta_1)\|_2^{\frac{1-2\xi}{1-\xi}}$. We have $\beta_1 \in (0, \infty)$, since $f$ is differentiable (Definition 6). Using $\eta \leq \frac{1}{\beta_1}$ and according to Lemma 33, we have $\delta(\theta_2) \leq \delta(\theta_1)$. Denote $\beta_2 \coloneqq c \cdot \|\nabla f(\theta_2)\|_2^{\frac{1-2\xi}{1-\xi}}$. We also have $\beta_2 \in (0, \infty)$. Repeating the update, we generate $\{\theta_t\}_{t \geq 1}$ such that $\delta(\theta_{t+1}) \leq \delta(\theta_t)$. Denote

$$\beta \coloneqq \sup_{t \geq 1} \{\beta_t\} = \sup_{t \geq 1} \left\{c \cdot \|\nabla f(\theta_2)\|_2^{\frac{1-2\xi}{1-\xi}}\right\}. \tag{D.72}$$

Now we have $0 \leq \delta(\theta_{t+1}) \leq \delta(\theta_t) \leq \cdots \leq \delta(\theta_1)$. According to the monotone convergence theorem, $\delta(\theta_t)$ converges to some finite value. And the gradient $\|\nabla f(\theta_t)\|_2 \to 0$, otherwise a small gradient update can decrease the

sub-optimality, which is a contradiction with convergence. Thus we have $\beta \in (\beta_1, \infty)$, since $\beta_t \to 0$ as $t \to \infty$. Using $\eta = \frac{1}{\beta}$, we have $\eta \leq \frac{1}{\beta_t}$ holds for all $t \geq 1$, and,

$$\beta(\theta_t) \leq c \cdot \|\nabla f(\theta_t)\|_2^{\frac{1-2\xi}{1-\xi}} = \beta_t \leq \beta. \tag{D.73}$$

Using similar calculations in the first part of (1a), we have the $O(1/t^{\frac{1}{1-2\xi}})$ upper bound.

**(1b) Second part:** *$\Omega(1/t^{\frac{1}{1-2\xi}})$ lower bound for GD update $\theta_{t+1} \leftarrow \theta_t - \eta_t \cdot \nabla f(\theta_t)$ with $\eta_t \in (0, 1]$.*

According to Eq. (D.30), we have,

$$\|\nabla f(\theta)\|_2^2 \leq 2 \cdot \beta(\theta) \cdot \delta(\theta) \tag{D.74}$$

$$\leq 2 \cdot c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}} \cdot \delta(\theta), \qquad \left(\beta(\theta) \leq c \cdot \|\nabla f(\theta)\|_2^{\frac{1-2\xi}{1-\xi}}\right) \tag{D.75}$$

which is equivalent to,

$$\|\nabla f(\theta)\|_2^2 \leq 2 \cdot c_1 \cdot \delta(\theta)^{2-2\xi}, \tag{D.76}$$

where $c_1 \coloneqq \frac{1}{2} \cdot (2 \cdot c)^{2-2\xi}$. According to Eq. (D.32), we have,

$$\delta(\theta_t) - \delta(\theta_{t+1}) \leq \left(\frac{\beta}{2} \cdot \eta_t^2 + \eta_t\right) \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.77}$$

$$\leq (\beta + 2) \cdot c_1 \cdot \delta(\theta_t)^{2-2\xi}. \qquad \text{(by Eq. (D.76) and } \eta_t \in (0, 1]) \tag{D.78}$$

Using similar calculations in the second part of (1a), we have the $\Omega(1/t^{\frac{1}{1-2\xi}})$ lower bound.

**(1b) Third part:** *$O(e^{-c' \cdot t})$ upper bound for GNGD update $\theta_{t+1} \leftarrow \theta_t - \frac{\nabla f(\theta_t)}{\beta(\theta_t)}$.*

According to Lemma 52, we have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.79}$$

$$\leq -\frac{1}{2 \cdot c} \cdot \|\nabla f(\theta_t)\|_2^{\frac{1}{1-\xi}} \qquad \left(\beta(\theta_t) \leq c \cdot \|\nabla f(\theta_t)\|_2^{\frac{1-2\xi}{1-\xi}}\right) \tag{D.80}$$

$$\leq -\frac{1}{2 \cdot c} \cdot C(\theta_t)^{\frac{1}{1-\xi}} \cdot \delta(\theta_t) \qquad \text{(NŁ)} \tag{D.81}$$

$$\leq -\frac{1}{2 \cdot c} \cdot C^{\frac{1}{1-\xi}} \cdot \delta(\theta_t), \qquad \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.82}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.83}$$

$$\leq \exp\left\{-(t-1)\cdot C^{\frac{1}{1-\xi}}/(2\cdot c)\right\}\cdot\left(f(\theta_1) - f(\theta^*)\right). \tag{D.84}$$

**(2a) First part:** $O(1/t^{\frac{1}{1-2\xi}})$ upper bound for GD when $\xi < 1/2$.

Similar to the first part of (1b), we denote $\beta_t \coloneqq L_0 + L_1 \cdot \|\nabla f(\theta_t)\|_2$ and $\beta \coloneqq \sup_{t\geq 1}\{\beta_t\} \in (L_0, \infty)$ since $\|\nabla f(\theta_t)\|_2 \to 0$ as $t \to \infty$. Using $\eta = \frac{1}{\beta}$, we have $\eta \leq \frac{1}{\beta_t}$ holds for all $t \geq 1$ and $\beta(\theta_t) \leq L_0 + L_1 \cdot \|\nabla f(\theta_t)\|_2 \leq \beta$. According to Eq. (D.6) and the first part of (1a), we have the $O(1/t^{\frac{1}{1-2\xi}})$ upper bound.

**(2a) Second part:** $O(e^{-c'\cdot t})$ upper bound for GD when $\xi = 1/2$.

According to Lemma 33, we have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2\beta}\cdot\|\nabla f(\theta_t)\|_2^2 \tag{D.85}$$

$$\leq -\frac{1}{2\beta}\cdot C(\theta_t)^2\cdot\delta(\theta_t) \qquad (\text{NŁ with } \xi = 1/2) \tag{D.86}$$

$$\leq -\frac{1}{2\beta}\cdot C^2\cdot\delta(\theta_t), \qquad \left(C \coloneqq \inf_{t\geq 1} C(\theta_t) > 0\right) \tag{D.87}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.88}$$

$$\leq \exp\left\{-(t-1)\cdot C^2/(2\beta)\right\}\cdot\left(f(\theta_1) - f(\theta^*)\right). \tag{D.89}$$

**(2a) Third part:** $O(1/t^{\frac{1}{1-2\xi}})$ upper bound for GNGD when $\xi < 1/2$.

According to Lemma 52, we have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.90}$$

$$\leq -\frac{1}{2} \cdot \frac{\|\nabla f(\theta_t)\|_2^2}{L_0 + L_1 \cdot \|\nabla f(\theta_t)\|_2} \qquad (\beta(\theta_t) \leq L_0 + L_1 \cdot \|\nabla f(\theta_t)\|_2) \tag{D.91}$$

$$\leq -\frac{1}{2} \cdot \frac{\|\nabla f(\theta_t)\|_2^2}{L_0 + L_1 \cdot \beta} \tag{D.92}$$

$$\left(\beta \coloneqq \sup_{t \geq 1} \{\|\nabla f(\theta_t)\|_2\} \in (\|\nabla f(\theta_1)\|_2, \infty)\right) \tag{D.93}$$

$$\leq -\frac{1}{2} \cdot \frac{C^2}{L_0 + L_1 \cdot \beta} \cdot \delta(\theta_t)^{2-2\xi}, \tag{D.94}$$

$$\left(\text{NŁ and } C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.95}$$

which is similar to Eq. (D.6). Using similar calculations in the first part of (1a), we have the $O(1/t^{\frac{1}{1-2\xi}})$ upper bound.

**(2a) Fourth part:** $O(e^{-c' \cdot t})$ upper bound for GNGD when $\xi = 1/2$.

We have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2} \cdot \frac{\|\nabla f(\theta_t)\|_2^2}{L_0 + L_1 \cdot \beta} \qquad (\text{by Eq. (D.92)}) \tag{D.96}$$

$$\leq -\frac{1}{2} \cdot \frac{C^2}{L_0 + L_1 \cdot \beta} \cdot \delta(\theta_t), \tag{D.97}$$

$$\left(\text{NŁ with } \xi = 1/2 \text{ and } C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.98}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.99}$$

$$\leq \exp\left\{-(t-1) \cdot C^2/(2 \cdot (L_0 + L_1 \cdot \beta))\right\} \cdot (f(\theta_1) - f(\theta^*)). \tag{D.100}$$

**(2b) First part:** $O(1/t^{\frac{1}{1-2\xi}})$ upper bound for GD when $\xi < 1/2$.

Denote $\beta_t \coloneqq L_0 \cdot \frac{\|\nabla f(\theta_t)\|^2}{\delta(\theta_t)^{2-2\xi}} + L_1 \cdot \|\nabla f(\theta_t)\|_2$ and $\beta \coloneqq \sup_{t \geq 1} \{\beta_t\} \in (\beta_1, \infty)$. According to Eq. (D.6) and the first part of (1a), we have the $O(1/t^{\frac{1}{1-2\xi}})$ upper bound.

**(2b) Second part:** $O(e^{-c' \cdot t})$ upper bound for GD when $\xi = 1/2$.

According to Lemma 33, we have, for all $t \geq 1$ (same as the second part of (2a)),

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2\beta} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.101}$$

$$\leq -\frac{1}{2\beta} \cdot C(\theta_t)^2 \cdot \delta(\theta_t) \qquad (\text{NŁ with } \xi = 1/2) \tag{D.102}$$

$$\leq -\frac{1}{2\beta} \cdot C^2 \cdot \delta(\theta_t), \qquad \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.103}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.104}$$

$$\leq \exp\left\{-(t-1) \cdot C^2/(2\beta)\right\} \cdot (f(\theta_1) - f(\theta^*)) \,. \tag{D.105}$$

**(2b) Third part:** $O(1/t^{\frac{1}{1-2\xi}})$ upper bound for GNGD when $\xi < 1/2$.

According to Lemma 52, we have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.106}$$

$$\leq -\frac{1}{2} \cdot \frac{\|\nabla f(\theta_t)\|_2^2}{L_0 \cdot \frac{\|\nabla f(\theta_t)\|^2}{\delta(\theta_t)^{2-2\xi}} + L_1 \cdot \|\nabla f(\theta_t)\|_2} \tag{D.107}$$

$$\left(\beta(\theta_t) \leq L_0 \cdot \frac{\|\nabla f(\theta_t)\|^2}{\delta(\theta_t)^{2-2\xi}} + L_1 \cdot \|\nabla f(\theta_t)\|_2\right) \tag{D.108}$$

$$= -\frac{1}{2} \cdot \frac{\delta(\theta_t)^{2-2\xi}}{L_0 + L_1 \cdot \frac{\delta(\theta_t)^{2-2\xi}}{\|\nabla f(\theta_t)\|_2}} \tag{D.109}$$

$$\leq -\frac{1}{2} \cdot \frac{\delta(\theta_t)^{2-2\xi}}{L_0 + L_1 \cdot \frac{\delta(\theta_t)^{1-\xi}}{C(\theta_t)}} \qquad \left(\|\nabla f(\theta_t)\|_2 \geq C(\theta_t) \cdot \delta(\theta_t)^{1-\xi}\right) \tag{D.110}$$

$$\leq -\frac{1}{2} \cdot \frac{\delta(\theta_t)^{2-2\xi}}{L_0 + L_1 \cdot \frac{\delta(\theta_t)^{1-\xi}}{C}} \qquad \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.111}$$

$$\leq -\frac{1}{2} \cdot \frac{\delta(\theta_t)^{2-2\xi}}{L_0 + L_1 \cdot \frac{\delta(\theta_1)^{1-\xi}}{C}}, \qquad (\delta_{t+1} \leq \delta_t, \text{ by Eq. (D.64)}) \tag{D.112}$$

which is similar to Eq. (D.6). Using similar calculations in the first part of (1a), we have the $O(1/t^{\frac{1}{1-2\xi}})$ upper bound.

**(2b) Fourth part:** $O(e^{-c' \cdot t})$ upper bound for GNGD when $\xi = 1/2$.

We have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2} \cdot \frac{\delta(\theta_t)^{2-2\xi}}{L_0 + L_1 \cdot \frac{\delta(\theta_1)^{1-\xi}}{C}} \tag{D.113}$$

$$(\delta_{t+1} \leq \delta_t \text{ by Eq. (D.106)}) \tag{D.114}$$

$$= -\frac{1}{2} \cdot \frac{\delta(\theta_t)}{L_0 + L_1 \cdot \frac{\delta(\theta_1)^{1/2}}{C}}, \qquad (\xi = 1/2) \tag{D.115}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.116}$$

$$\leq \exp\left\{-\frac{C \cdot (t-1)}{2 \cdot (L_0 \cdot C + L_1 \cdot \delta(\theta_1)^{1/2})}\right\} \cdot (f(\theta_1) - f(\theta^*)) \tag{D.117}$$

$$\leq \exp\left\{-\frac{C \cdot (t-1)}{2 \cdot (L_0 + L_1 \cdot \delta(\theta_1)^{1/2})}\right\} \cdot (f(\theta_1) - f(\theta^*)) \,. \tag{D.118}$$

$$\text{(if } C \leq 1\text{)} \tag{D.119}$$

**(3a)** $O(e^{-c' \cdot t})$ upper bound for GNGD update when $\xi \in (1/2, 1)$.

According to Lemma 52, we have, for all $t \geq 1$ (same as the third part of (1b)),

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.120}$$

$$\leq -\frac{1}{2 \cdot c} \cdot \|\nabla f(\theta_t)\|_2^{\frac{1}{1-\xi}} \qquad \left(\beta(\theta_t) \leq c \cdot \|\nabla f(\theta_t)\|_2^{\frac{1-2\xi}{1-\xi}}\right) \tag{D.121}$$

$$\leq -\frac{1}{2 \cdot c} \cdot C(\theta_t)^{\frac{1}{1-\xi}} \cdot \delta(\theta_t) \qquad \text{(NŁ)} \tag{D.122}$$

$$\leq -\frac{1}{2 \cdot c} \cdot C^{\frac{1}{1-\xi}} \cdot \delta(\theta_t), \qquad \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.123}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.124}$$

$$\leq \exp\left\{-(t-1) \cdot C^{\frac{1}{1-\xi}}/(2 \cdot c)\right\} \cdot (f(\theta_1) - f(\theta^*)) \,. \tag{D.125}$$

**(3b)** $O(e^{-c' \cdot t})$ upper bound for GNGD update when $\xi \in (1/2, 1)$.

According to Lemma 52, we have, for all $t \geq 1$ (same as the third part of

(1a)),

$$\delta(\theta_{t+1}) - \delta(\theta_t) \leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot \|\nabla f(\theta_t)\|_2^2 \tag{D.126}$$

$$\leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot C(\theta_t)^2 \cdot \delta(\theta_t)^{2-2\xi} \qquad (\text{NŁ}) \tag{D.127}$$

$$\leq -\frac{1}{2 \cdot \beta(\theta_t)} \cdot C^2 \cdot \delta(\theta_t)^{2-2\xi} \qquad \left(C \coloneqq \inf_{t \geq 1} C(\theta_t) > 0\right) \tag{D.128}$$

$$\leq -\frac{C^2}{2 \cdot c} \cdot \delta(\theta_t), \qquad \left(\beta(\theta_t) \leq c \cdot \delta(\theta_t)^{1-2\xi}\right) \tag{D.129}$$

which implies (similar to Eq. (D.68)),

$$f(\theta_t) - f(\theta^*) = \delta(\theta_t) \tag{D.130}$$

$$\leq \exp\left\{-(t-1) \cdot C^2/(2 \cdot c)\right\} \cdot (f(\theta_1) - f(\theta^*)) \,. \qquad \square$$

### D.1.1 Function Classes in Fig. 4.2

**Proposition 7.** The following hold for an objective $f$:

**(1)** D $\subseteq$ C. If $f$ satisfies NŁ with degree $\xi$, it satisfies NŁ with degree $\xi' < \xi$;

**(2)** F $\subseteq$ D. A strongly convex $f$ satisfies NŁ with $\xi = 1/2$;

**(3)** F $\cap$ A $= \emptyset$. A strongly convex $f$ cannot satisfy NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$;

**(4)** E $\subseteq$ C. A (not strongly) convex $f$ satisfies NŁ with $\xi = 0$.

*Proof.* (1) D $\subseteq$ C. Suppose a function $f : \Theta \to \mathbb{R}$ satisfies NŁ with $\xi$, i.e.,

$$\left\| \frac{df(\theta)}{d\theta} \right\|_2 \geq C(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi} \,, \tag{D.131}$$

where $\xi \in (-\infty, 1]$, and $C(\theta) > 0$ holds for all $\theta \in \Theta$. Let $\xi' < \xi$. If $|f(\theta) - f(\theta^*)| > 0$, then we have,

$$|f(\theta) - f(\theta^*)|^{1-\xi} = \frac{|f(\theta) - f(\theta^*)|^{1-\xi'}}{|f(\theta) - f(\theta^*)|^{\xi-\xi'}} \tag{D.132}$$

$$\geq c(\theta) \cdot |f(\theta) - f(\theta^*)|^{1-\xi'} \,, \tag{D.133}$$

where $c(\theta) \coloneqq \frac{1}{|f(\theta)-f(\theta^*)|^{\xi-\xi'}} > 0$, and $c(\theta) \not\to 0$ as $\theta \to \theta^*$ (or $c(\theta) > c > 0$ for all $\theta$ within a finite distance of $\theta^*$). If $|f(\theta) - f(\theta^*)| = 0$, then it trivially holds that

$$|f(\theta) - f(\theta^*)|^{1-\xi} \geq |f(\theta) - f(\theta^*)|^{1-\xi'} \,. \tag{D.134}$$

(2) F $\subseteq$ D. Suppose a function $f : \Theta \to \mathbb{R}$ is strongly convex. We have, there exists $\mu > 0$, for all $\theta$, $\theta' \in \Theta$,

$$f(\theta') \geq f(\theta) + \nabla f(\theta)^\top (\theta' - \theta) + \frac{\mu}{2} \cdot \|\theta' - \theta\|_2^2 . \tag{D.135}$$

Fix $\theta$ and take minimum over $\theta'$ on both sides of the above inequality. Then we have,

$$f(\theta^*) \geq f(\theta) + \min_{\theta'} \left\{ \nabla f(\theta)^\top (\theta' - \theta) + \frac{\mu}{2} \cdot \|\theta' - \theta\|_2^2 \right\} \tag{D.136}$$

$$= f(\theta) - \frac{1}{\mu} \cdot \|\nabla f(\theta)\|_2^2 + \frac{1}{2\mu} \cdot \|\nabla f(\theta)\|_2^2 \tag{D.137}$$

$$\left( \theta' = \theta - \frac{1}{\mu} \cdot \nabla f(\theta) \right) \tag{D.138}$$

$$= f(\theta) - \frac{1}{2\mu} \cdot \|\nabla f(\theta)\|_2^2 , \tag{D.139}$$

which is equivalent to,

$$\|\nabla f(\theta)\|_2 \geq \sqrt{2\mu} \cdot (f(\theta) - f(\theta^*))^{\frac{1}{2}} , \tag{D.140}$$

which means $f$ satisfies NŁ inequality with $\xi = 1/2$.

(3) F $\cap$ A $= \emptyset$. Suppose a function $f : \Theta \to \mathbb{R}$ is strongly convex. There exists $\mu > 0$, for all $\theta \in \Theta$,

$$\left| z^\top \frac{\partial^2 f(\theta)}{\partial \theta^2} z \right| \geq \mu \cdot \|z\|_2^2 , \tag{D.141}$$

holds for all vector $z$ that has the same dimension as $\theta$. Next we show $f \notin$ A. Suppose $f \in$ A. We have,

$$\beta(\theta^*) = \sup_z \left| z^\top \frac{\partial^2 f(\theta^*)}{\partial (\theta^*)^2} z \right| = 0, \tag{D.142}$$

which is a contradiction with Eq. (D.141). Therefore $f \notin$ A, and F $\cap$ A $= \emptyset$.

(4) E $\subseteq$ C. Suppose a function $f : \Theta \to \mathbb{R}$ is convex. We have, for all $\theta$, $\theta' \in \Theta$,

$$f(\theta') \geq f(\theta) + \nabla f(\theta)^\top (\theta' - \theta). \tag{D.143}$$

Take $\theta' = \theta^*$. We have,

$$f(\theta^*) \geq f(\theta) + \nabla f(\theta)^\top (\theta^* - \theta), \tag{D.144}$$

which implies,

$$\|\nabla f(\theta)\|_2 = \frac{1}{\|\theta - \theta^*\|_2} \cdot \|\nabla f(\theta)\|_2 \cdot \|\theta - \theta^*\|_2 \tag{D.145}$$

$$\geq \frac{1}{\|\theta - \theta^*\|_2} \cdot \nabla f(\theta)^\top (\theta^* - \theta) \qquad \text{(by Cauchy-Schwarz)} \tag{D.146}$$

$$\geq \frac{1}{\|\theta - \theta^*\|_2} \cdot (f(\theta) - f(\theta^*)), \qquad \text{(by Eq. (D.144))} \tag{D.147}$$

and $C(\theta) = \frac{1}{\|\theta - \theta^*\|_2} \not\to 0$ as $\theta \to \theta^*$ (or $C(\theta) > c > 0$ for all $\|\theta - \theta^*\|_2$ smaller than a finite value, e.g., within a bounded constraint). Therefore $f$ satisfies NŁ inequality with $\xi = 0$. □

**Proposition 8.** The following results hold:

**(1)** ACE $\neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.

**(2)** ADE $\neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.

**(3)** BCE $\neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.

**(4)** BDE $\neq \emptyset$. There exists at least one (not strongly) convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.

**(5)** BF $\neq \emptyset$. There exists at least one strongly convex function which satisfies NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.

*Proof.* (1) ACE $\neq \emptyset$. Consider minimizing the following function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = x^4. \tag{D.148}$$

The second order derivative (Hessian) is $f''(x) = 12 \cdot x^2 \geq 0$, which means $f$ is (not strongly) convex. According to Taylor's theorem, we have, for all $x, x' \in \mathbb{R}$,

$$\left| f(x') - f(x) - \left\langle \frac{df(x)}{dx}, x' - x \right\rangle \right| \leq \frac{|f''(x_\zeta)|}{2} \cdot \|\theta' - \theta\|_2^2 \tag{D.149}$$

$$= \frac{12 \cdot x_\zeta^2}{2} \cdot \|\theta' - \theta\|_2^2, \tag{D.150}$$

where $x_\zeta \coloneqq x+\zeta\cdot(x'-x)$ with some $\zeta \in [0,1]$. Thus we have $\beta(x) = 12\cdot x_\zeta^2 \to 0$ as $x, x' \to 0$. Next, we have,

$$|f'(x)| = \left|4\cdot x^3\right| = 4\cdot\left(|x|^4\right)^{\frac{3}{4}} = 4\cdot\left(f(x)-f(0)\right)^{1-\frac{1}{4}}, \tag{D.151}$$

which means $f$ satisfies NŁ inequality with $\xi = 1/4 < 1/2$.

(2) ADE $\neq \emptyset$. Consider minimizing the following function $f : \mathbb{R}^K \to \mathbb{R}$,

$$f(\theta) = D_{\mathrm{KL}}(y\|\pi_\theta) = D_{\mathrm{KL}}(y\|\mathrm{softmax}(\theta)), \tag{D.152}$$

where $y \in \{0,1\}^K$ is a one-hot vector. We show that $f$ is a (not strongly) convex function. The gradient of $f$ is,

$$\frac{\partial f(\theta)}{\partial\theta} = \left(\frac{d\pi_\theta}{d\theta}\right)^\top \left(\frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\pi_\theta}\right) \tag{D.153}$$

$$= \left(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top\right)\mathrm{diag}\left(\frac{1}{\pi_\theta}\right)(-y) \tag{D.154}$$

$$= \pi_\theta - y. \tag{D.155}$$

Therefore the Hessian is,

$$\frac{\partial^2 f(\theta)}{\partial\theta^2} = \frac{d\pi_\theta}{d\theta} = \mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top. \tag{D.156}$$

According to Lemma 37, we have,

$$\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top \succeq \mathbf{0}, \tag{D.157}$$

and the minimum eigenvalue of $\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top$ is 0, which means $f$ is convex but not strongly convex. Next, according to Lemma 50, we have,

$$D_{\mathrm{KL}}(y\|\pi_\theta) = \sum_a y(a)\cdot\log\left(\frac{y(a)}{\pi_\theta(a)}\right) \tag{D.158}$$

$$\leq \sum_a y(a)\cdot\left(\frac{y(a)}{\pi_\theta(a)} - 1\right) \qquad (\log x \leq x-1) \tag{D.159}$$

$$= \sum_a \left(y(a) - \pi_\theta(a) + \pi_\theta(a)\right)\cdot\frac{y(a)-\pi_\theta(a)}{\pi_\theta(a)} \tag{D.160}$$

$$= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\pi_\theta(a)} \tag{D.161}$$

$$\leq \frac{1}{\min_a \pi_\theta(a)}\cdot\sum_a \left(y(a)-\pi_\theta(a)\right)^2, \tag{D.162}$$

which implies,

$$\left\|\frac{\partial f(\theta)}{\partial \theta}\right\|_2 = \|\pi_\theta - y\|_2 \qquad \text{(by Eq. (D.153))} \tag{D.163}$$

$$\geq \min_a \sqrt{\pi_\theta(a)} \cdot \left[D_{\mathrm{KL}}(y\|\pi_\theta) - D_{\mathrm{KL}}(y\|y)\right]^{\frac{1}{2}}, \tag{D.164}$$

$$\text{(by Eq. (D.158))} \tag{D.165}$$

which means $f$ satisfies NŁ with $\xi = 1/2$. Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. We have, as $\pi_\theta, \pi_{\theta'} \to y$,

$$\beta(\theta) = \sup_z \left| z^\top \frac{\partial^2 f(\theta_\zeta)}{\partial \theta_\zeta^2} z \right| \tag{D.166}$$

$$= \sup_z \left| z^\top \left(\mathrm{diag}(\pi_{\theta_\zeta}) - \pi_{\theta_\zeta} \pi_{\theta_\zeta}^\top\right) z \right| \qquad \text{(by Eq. (D.156))} \tag{D.167}$$

$$\to \sup_z \left| z^\top \left(\mathrm{diag}(y) - yy^\top\right) z \right| \tag{D.168}$$

$$= \sup_z \left| z^\top \mathbf{0} z \right| \qquad (y \text{ is one-hot}) \tag{D.169}$$

$$= 0. \tag{D.170}$$

(3) BCE $\neq \emptyset$. Consider the (modified) Huber loss function,

$$f(x) = \begin{cases} x^2, & \text{if } |x| \leq 1, \\ 2 \cdot |x| - 1, & \text{otherwise} \end{cases} \tag{D.171}$$

which is a (not strongly) convex function. According to (4) in Proposition 7, $f$ satisfies NŁ inequality with $\xi = 0$. Denote $x_\zeta \coloneqq x + \zeta \cdot (x' - x)$ with some $\zeta \in [0, 1]$. We have $\beta(x) = |f''(x_\zeta)| \to 2 > 0$, as $x, x' \to 0$.

(4) BDE $\neq \emptyset$. Consider minimizing the same function as in (2),

$$f(\theta) = D_{\mathrm{KL}}(y\|\pi_\theta) = D_{\mathrm{KL}}(y\|\mathrm{softmax}(\theta)), \tag{D.172}$$

where $y \in (0, 1)^K$ is a probability vector with $\min_a y(a) > 0$, i.e., $y$ is bounded away from the boundary of probability simplex. As shown in (2), $f$ is (not strongly) convex and $f$ satisfies NŁ with $\xi = 1/2$. Next, we have,

$$\beta(\theta) = \sup_z \left| z^\top \left(\mathrm{diag}(\pi_{\theta_\zeta}) - \pi_{\theta_\zeta} \pi_{\theta_\zeta}^\top\right) z \right| \qquad \text{(by Eq. (D.156))} \tag{D.173}$$

$$\to \sup_z \left| z^\top \left(\mathrm{diag}(y) - yy^\top\right) z \right| \tag{D.174}$$

$$= \sup_z \left| \mathop{\mathbb{E}}_{a\sim y}[z(a)^2] - \left(\mathop{\mathbb{E}}_{a\sim y}[z(a)]\right)^2 \right| \tag{D.175}$$

$$= \sup_z \left| \mathrm{Var}_{a\sim y}[z(a)] \right| > 0. \tag{D.176}$$

(5) BF $\neq \emptyset$. Consider minimizing the following function,

$$f(x) = x^2, \tag{D.177}$$

where $x \in \mathbb{R}$. $f$ is strongly convex, and $\beta(x) = \beta = 2$. Thus $\beta(x) \to 2 > 0$ as $x, x' \to 0$ in Definition 6. □

**Proposition 9.** The following results hold:

**(1)** $\mathcal{W} \coloneqq \text{AC} \setminus (\text{AD} \cup \text{ACE}) \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.

**(2)** $\mathcal{X} \coloneqq \text{AD} \setminus \text{ADE} \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$.

**(3)** $\mathcal{Y} \coloneqq \text{BC} \setminus (\text{BD} \cup \text{BCE}) \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi < 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.

**(4)** $\mathcal{Z} \coloneqq \text{BD} \setminus (\text{BDE} \cup \text{BF}) \neq \emptyset$. There exists at least one non-convex function which satisfies NŁ with $\xi \geq 1/2$ and NS with $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$.

*Proof.* (1) $\mathcal{W} \coloneqq \text{AC} \setminus (\text{AD} \cup \text{ACE}) \neq \emptyset$. Consider maximizing the expected reward,

$$f(\theta) = \pi_\theta^\top r, \tag{D.178}$$

where $\pi_\theta = \text{softmax}(\theta)$ and $\theta \in \mathbb{R}^K$. According to Proposition 1, $f$ is non-concave. According to Lemma 3, we have,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r, \tag{D.179}$$

which means $f$ satisfies NŁ inequality with $\xi = 0$. As shown in Lemma 21, we have $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$. Therefore, $\beta(\theta_\zeta) \to 0$ as $\pi_\theta, \pi_{\theta'} \to \pi^*$.
(2) $\mathcal{X} \coloneqq \text{AD} \setminus \text{ADE} \neq \emptyset$. Consider minimizing the function $f : \mathbb{R}^K \to \mathbb{R}$,

$$f(\theta) = \|\pi_\theta - y\|_2^2, \tag{D.180}$$

where $\pi_\theta = \text{softmax}(\theta)$, $\theta \in \mathbb{R}^K$, and $y \in \{0, 1\}$ is a one-hot vector. We show that $f$ is non-convex using one example. Let $y = (1, 0, 0)^\top$. Let $\theta_1 = (0, 0, 0)^\top$,

$\pi_{\theta_1} = \text{softmax}(\theta_1) = (1/3, 1/3, 1/3)^\top$, $\theta_2 = (\log 4, \log 36, \log 100)^\top$, and $\pi_{\theta_2} = \text{softmax}(\theta_2) = (4/140, 36/140, 100/140)^\top$. We have,

$$f(\theta_1) = \|\pi_{\theta_1} - y\|_2^2 = \frac{2}{3}, \text{ and } f(\theta_2) = \|\pi_{\theta_2} - y\|_2^2 = \frac{38}{25}. \tag{D.181}$$

Denote $\bar{\theta} = \frac{1}{2} \cdot (\theta_1 + \theta_2) = (\log 2, \log 6, \log 10)^\top$ we have $\pi_{\bar{\theta}} = \text{softmax}(\bar{\theta}) = (2/18, 6/18, 10/18)^\top$ and

$$f(\bar{\theta}) = \|\pi_{\bar{\theta}} - y\|_2^2 = \frac{98}{81}. \tag{D.182}$$

Therefore we have,

$$\frac{1}{2} \cdot (f(\theta_1) + f(\theta_2)) = \frac{82}{75} = \frac{2214}{2025} < \frac{2450}{2025} = \frac{98}{81} = f(\bar{\theta}), \tag{D.183}$$

which means $f$ is non-convex. Denote $H(\pi_\theta) \coloneqq \text{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top$ as the Jacobian of $\theta \mapsto \text{softmax}(\theta)$. We have,

$$\left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2 = \left\| \left( \frac{d\pi_\theta}{d\theta} \right)^\top \left( \frac{df(\theta)}{d\pi_\theta} \right) \right\|_2 \tag{D.184}$$

$$= 2 \cdot \|H(\pi_\theta)(\pi_\theta - y)\|_2 \tag{D.185}$$

$$\geq 2 \cdot \min_a \pi_\theta(a) \cdot \|\pi_\theta - y\|_2 \qquad \text{(by Lemma 38)} \tag{D.186}$$

$$= 2 \cdot \min_a \pi_\theta(a) \cdot [f(\theta) - f(y)]^{\frac{1}{2}}, \tag{D.187}$$

which means $f$ satisfies NŁ inequality with $\xi = 1/2$. Denote $S \coloneqq S(y, \theta) \in \mathbb{R}^{K \times K}$ as the second derivative (Hessian) of $f$. We have,

$$S = \frac{d}{d\theta} \left\{ \frac{df(\theta)}{d\theta} \right\} \tag{D.188}$$

$$= \frac{d}{d\theta} \{H(\pi_\theta)(\pi_\theta - y)\}. \tag{D.189}$$

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

$$S_{(i,j)} = \frac{d\{\pi_\theta(i) \cdot \left[\pi_\theta(i) - y(i) - \pi_\theta^\top (\pi_\theta - y)\right]\}}{d\theta(j)} \tag{D.190}$$

$$= \frac{d\pi_\theta(i)}{d\theta(j)} \cdot \left[\pi_\theta(i) - y(i) - \pi_\theta^\top (\pi_\theta - y)\right] \tag{D.191}$$

$$+ \pi_\theta(i) \cdot \frac{d\{\pi_\theta(i) - y(i) - \pi_\theta^\top (\pi_\theta - y)\}}{d\theta(j)} \tag{D.192}$$

$$= (\delta_{ij}\pi_\theta(j) - \pi_\theta(i)\pi_\theta(j)) \cdot \left[\pi_\theta(i) - y(i) - \pi_\theta^\top (\pi_\theta - y)\right] \tag{D.193}$$

$$+ \pi_\theta(i) \cdot \left[\delta_{ij}\pi_\theta(j) - \pi_\theta(i)\pi_\theta(j)\right. \tag{D.194}$$

$$- \pi_\theta(j) \cdot \left(\pi_\theta(j) - y(j) - \pi_\theta^\top (\pi_\theta - y)\right) \tag{D.195}$$

$$\left. - \pi_\theta(j) \cdot \left(\pi_\theta(j) - \pi_\theta^\top \pi_\theta\right) \right] \tag{D.196}$$

$$= \delta_{ij}\pi_\theta(j) \cdot \left[\pi_\theta(i) - y(i) - \pi_\theta^\top (\pi_\theta - y)\right] \tag{D.197}$$

$$- \pi_\theta(i)\pi_\theta(j) \cdot \left[\pi_\theta(i) - y(i) - \pi_\theta^\top (\pi_\theta - y)\right] \tag{D.198}$$

$$- \pi_\theta(i)\pi_\theta(j) \cdot \left[\pi_\theta(j) - y(j) - \pi_\theta^\top (\pi_\theta - y)\right] \tag{D.199}$$

$$+ \pi_\theta(i)\pi_\theta(j) \cdot \left[\delta_{ij} - \pi_\theta(i) - \pi_\theta(j) + \pi_\theta^\top \pi_\theta\right], \tag{D.200}$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases} \tag{D.201}$$

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

$$\left|z^\top S z\right| = \left|\sum_{i=1}^{K}\sum_{j=1}^{K} S_{(i,j)} \cdot z(i) \cdot z(j)\right| \tag{D.202}$$

$$= \Big| \left(H(\pi_\theta)\left(\pi_\theta - y\right)\right)^\top (z \odot z) - 2 \cdot \left(H(\pi_\theta)\left(\pi_\theta - y\right)\right)^\top z \cdot \left(\pi_\theta^\top z\right) \tag{D.203}$$

$$+ \left(\pi_\theta \odot \pi_\theta\right)^\top (z \odot z) - 2 \cdot \left(\pi_\theta \odot \pi_\theta\right)^\top z \cdot \left(\pi_\theta^\top z\right) \tag{D.204}$$

$$+ \left(\pi_\theta^\top z\right)^2 \cdot \left(\pi_\theta^\top \pi_\theta\right) \Big|, \tag{D.205}$$

where $\odot$ is Hadamard (component-wise) product. We have, as $\pi_\theta \to y$,

$$\left(H(\pi_\theta)\left(\pi_\theta - y\right)\right)^\top (z \odot z) - 2 \cdot \left(H(\pi_\theta)\left(\pi_\theta - y\right)\right)^\top z \cdot \left(\pi_\theta^\top z\right) \tag{D.206}$$

$$\to \left(H(y)\mathbf{0}\right)^\top (z \odot z) - 2 \cdot \left(H(y)\mathbf{0}\right)^\top z \cdot \left(y^\top z\right) \tag{D.207}$$

$$= 0. \tag{D.208}$$

Since $y$ is one-hot vector, we have, as $\pi_\theta \to y$,

$$(\pi_\theta \odot \pi_\theta)^\top (z \odot z) - 2 \cdot (\pi_\theta \odot \pi_\theta)^\top z \cdot \left(\pi_\theta^\top z\right) + \left(\pi_\theta^\top z\right)^2 \cdot \pi_\theta^\top \pi_\theta \tag{D.209}$$

$$\to y^\top (z \odot z) - 2 \cdot \left(y^\top z\right)^2 + \left(y^\top z\right)^2 \cdot y^\top y \tag{D.210}$$

$$= \left(y^\top z\right)^2 - 2 \cdot \left(y^\top z\right)^2 + \left(y^\top z\right)^2 = 0, \tag{D.211}$$

which means $\beta(\theta) \to 0$ as $\theta, \theta' \to \theta^*$ in Definition 6.

(3) $\mathcal{Y} \coloneqq \mathrm{BC} \setminus (\mathrm{BD} \cup \mathrm{BCE}) \neq \emptyset$. Consider minimizing the function $f : \mathbb{R} \to \mathbb{R}$,

$$f(\theta) = \begin{cases} 2 \cdot (\pi_\theta - \pi_{\theta^*})^2, & \text{if } |\pi_\theta - \pi_{\theta^*}| \leq 0.2, \\ 25 \cdot (\pi_\theta - \pi_{\theta^*})^4 + 0.04, & \text{otherwise} \end{cases} \tag{D.212}$$

where $\theta \in \mathbb{R}$, $\theta^* = 0$, and $\pi_\theta$ is defined as,

$$\pi_\theta = \sigma(\theta) = \frac{1}{1 + e^{-\theta}}, \tag{D.213}$$

where $\sigma : \mathbb{R} \to (0, 1)$ is the sigmoid activation. Fig. D.1 shows the image of $f$, indicating that $f$ is a non-convex function.

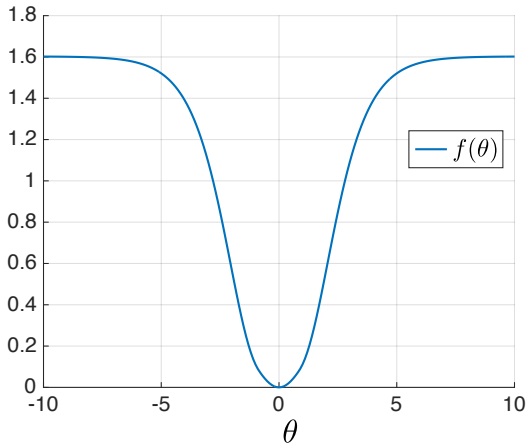



Figure D.1: The image of $f$.

Since $\theta^* = 0$, we have $\pi_{\theta^*} = 1/2$, and for all $|\pi_\theta - \pi_{\theta^*}| > 0.2$,

$$\left|\frac{df(\theta)}{d\theta}\right| = \left|\frac{d\pi_\theta}{d\theta} \cdot \frac{df(\theta)}{d\pi_\theta}\right| \tag{D.214}$$

$$= \left|\pi_\theta \cdot (1 - \pi_\theta) \cdot 100 \cdot (\pi_\theta - \pi_{\theta^*})^3\right| \tag{D.215}$$

$$= 100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot \left[(\pi_\theta - \pi_{\theta^*})^4\right]^{\frac{3}{4}} \tag{D.216}$$

$$= 100 \cdot \pi_\theta \cdot (1 - \pi_\theta) \cdot \left[f(\theta) - f(\theta^*)\right]^{1 - \frac{1}{4}}, \tag{D.217}$$

which means $f$ satisfies NŁ inequality with $\xi = 1/4 < 1/2$. For all $|\pi_\theta - \pi_{\theta^*}| \leq 1$, the Hessian of $f$ is,

$$\left|\frac{d^2 f(\theta)}{d\theta^2}\right| = \left|\frac{d}{d\theta}\left\{100 \cdot \pi_\theta \cdot (1-\pi_\theta) \cdot (\pi_\theta - \pi_{\theta^*})\right\}\right| \tag{D.218}$$

$$= \big|100 \cdot \pi_\theta \cdot (1-\pi_\theta) \cdot (\pi_\theta - \pi_{\theta^*}) \cdot (1 - 2\pi_\theta) \tag{D.219}$$

$$+ 100 \cdot \pi_\theta^2 \cdot (1-\pi_\theta)^2 \,\big|. \tag{D.220}$$

As $\pi_\theta \to \pi_{\theta^*} = 1/2$, we have

$$100 \cdot \pi_\theta \cdot (1-\pi_\theta) \cdot (\pi_\theta - \pi_{\theta^*}) \cdot (1 - 2\pi_\theta) \to 0, \tag{D.221}$$

and,

$$100 \cdot \pi_\theta^2 \cdot (1-\pi_\theta)^2 \to 100 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{25}{4} > 0, \tag{D.222}$$

which means $\beta(\theta) \to \beta > 0$ as $\theta, \theta' \to \theta^*$ in Definition 6.

(4) $\mathcal{Z} \coloneqq \mathrm{BD} \setminus (\mathrm{BDE} \cup \mathrm{BF}) \neq \emptyset$. Consider minimizing the same function as in (2),

$$f(\theta) = \|\pi_\theta - y\|_2^2, \tag{D.223}$$

where $\pi_\theta = \mathrm{softmax}(\theta)$, $\theta \in \mathbb{R}^K$, and $y \in (0,1)$ is a probability vector with $\min_a y(a) > 0$, i.e., $y$ is bounded away from the boundary of probability simplex. We show that $f$ is non-convex using one example. Let $y = (1/2, 1/4, 1/4)^\top$. Let $\theta_1 = (0,0,0)^\top$, $\pi_{\theta_1} = \mathrm{softmax}(\theta_1) = (1/3, 1/3, 1/3)^\top$, $\theta_2 = (\log 4, \log 36, \log 100)^\top$, and $\pi_{\theta_2} = \mathrm{softmax}(\theta_2) = (4/140, 36/140, 100/140)^\top$. We have,

$$f(\theta_1) = \|\pi_{\theta_1} - y\|_2^2 = \frac{1}{24}, \text{ and } f(\theta_2) = \|\pi_{\theta_2} - y\|_2^2 = \frac{613}{1400}. \tag{D.224}$$

Denote $\bar{\theta} = \frac{1}{2} \cdot (\theta_1 + \theta_2) = (\log 2, \log 6, \log 10)^\top$ we have $\pi_{\bar{\theta}} = \mathrm{softmax}(\bar{\theta}) = (2/18, 6/18, 10/18)^\top$ and

$$f(\bar{\theta}) = \|\pi_{\bar{\theta}} - y\|_2^2 = \frac{163}{648}. \tag{D.225}$$

Therefore we have,

$$\frac{1}{2} \cdot (f(\theta_1) + f(\theta_2)) = \frac{1007}{4200} = \frac{27189}{113400} < \frac{28525}{113400} = \frac{163}{648} = f(\bar{\theta}), \tag{D.226}$$

which means $f$ is non-convex. Similar as (2), we have the Hessian of $f$,

$$S_{(i,j)} = \underbrace{\delta_{ij}\pi_\theta(j)\cdot\left[\pi_\theta(i) - y(i) - \pi_\theta^\top(\pi_\theta - y)\right]}_{\text{(a)}} \tag{D.227}$$

$$- \underbrace{\pi_\theta(i)\pi_\theta(j)\cdot\left[\pi_\theta(i) - y(i) - \pi_\theta^\top(\pi_\theta - y)\right]}_{\text{(b)}} \tag{D.228}$$

$$- \underbrace{\pi_\theta(i)\pi_\theta(j)\cdot\left[\pi_\theta(j) - y(j) - \pi_\theta^\top(\pi_\theta - y)\right]}_{\text{(c)}} \tag{D.229}$$

$$+ \underbrace{\pi_\theta(i)\pi_\theta(j)\cdot\left[\delta_{ij} - \pi_\theta(i) - \pi_\theta(j) + \pi_\theta^\top\pi_\theta\right]}_{\text{(d)}}, \tag{D.230}$$

where $(a) = (b) = (c) = 0$ when $\pi_\theta = y$. Hence, at the optimal point $\theta^*$, we have,

$$S = \frac{1}{128}\cdot\begin{bmatrix}12 & -6 & -6\\ -6 & 7 & -1\\ -6 & -1 & 7\end{bmatrix}, \tag{D.231}$$

and the eigenvalues of $S$ are 0, $\frac{1}{16}$, and $\frac{9}{64}$. Thus as $\theta, \theta' \to \theta^*$, the Hessian spectral radius of $f$ satisfies $\beta(\theta) \to \beta = \frac{9}{64}$. □

**Proposition 10.** The convex function $f : x \mapsto |x|^p$ with $p > 1$ satisfies the NŁ inequality with $\xi = 1/p$ and the NS property with $\beta(x) \le c_1 \cdot \delta(x)^{1-2\xi}$.

*Proof.* For $p > 1$, $f$ is differentiable, and we have,

$$|f'(x)| = \left|p\cdot|x|^{p-1}\cdot\operatorname{sign}\{x\}\right| = p\cdot(|x|^p)^{\frac{p-1}{p}} = p\cdot(f(x) - f(0))^{1-\frac{1}{p}}, \tag{D.232}$$

which means $f$ satisfies NŁ inequality with $\xi = 1/p$. On the other hand, the Hessian of $f$ is,

$$|f''(x)| = \left|p\cdot(p-1)\cdot|x|^{p-2}\right| \tag{D.233}$$

$$= p\cdot(p-1)\cdot(|x|^p)^{\frac{p-2}{p}} \tag{D.234}$$

$$= p\cdot(p-1)\cdot(f(x) - f(0))^{1-\frac{2}{p}}.$$ □

## D.2 Proofs for Section 4.6: Geometry-aware Normalized Policy Gradient

### D.2.1 One-state MDPs

**Lemma 21** (NS) **.** Denote $\theta_\zeta := \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. For any $r \in [0,1]^K$, $\theta \mapsto \pi_\theta^\top r$ satisfies $\beta(\theta_\zeta)$ non-uniform smoothness with $\beta(\theta_\zeta) = 3 \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2$.

*Proof.* Let $S := S(r,\theta) \in \mathbb{R}^{K \times K}$ be the second derivative of the value map $\theta \mapsto \pi_\theta^\top r$. By Taylor's theorem, it suffices to show that the spectral radius of $S$ is upper bounded. Denote $H(\pi_\theta) := \mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top$ as the Jacobian of $\theta \mapsto \mathrm{softmax}(\theta)$. Now, by its definition we have

$$S = \frac{d}{d\theta} \left\{ \frac{d\pi_\theta^\top r}{d\theta} \right\} \tag{D.235}$$

$$= \frac{d}{d\theta} \left\{ H(\pi_\theta) r \right\} \tag{D.236}$$

$$= \frac{d}{d\theta} \left\{ (\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top) r \right\}. \tag{D.237}$$

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

$$S_{(i,j)} = \frac{d\{\pi_\theta(i) \cdot (r(i) - \pi_\theta^\top r)\}}{d\theta(j)} \tag{D.238}$$

$$= \frac{d\pi_\theta(i)}{d\theta(j)} \cdot (r(i) - \pi_\theta^\top r) + \pi_\theta(i) \cdot \frac{d\{r(i) - \pi_\theta^\top r\}}{d\theta(j)} \tag{D.239}$$

$$= (\delta_{ij} \pi_\theta(j) - \pi_\theta(i)\pi_\theta(j)) \cdot (r(i) - \pi_\theta^\top r) \tag{D.240}$$

$$- \pi_\theta(i) \cdot (\pi_\theta(j) r(j) - \pi_\theta(j) \pi_\theta^\top r) \tag{D.241}$$

$$= \delta_{ij} \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) \tag{D.242}$$

$$- \pi_\theta(i)\pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i)\pi_\theta(j) \cdot (r(j) - \pi_\theta^\top r), \tag{D.243}$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases} \tag{D.244}$$

is Kronecker's $\delta$-function as defined in Eq. (D.201). To show the bound on the

spectral radius of $S$, pick $y \in \mathbb{R}^K$. Then,

$$\left|y^\top S y\right| = \left|\sum_{i=1}^{K}\sum_{j=1}^{K} S_{(i,j)} \cdot y(i) \cdot y(j)\right| \tag{D.245}$$

$$= \Bigg|\sum_i \pi_\theta(i)(r(i) - \pi_\theta^\top r) y(i)^2 \tag{D.246}$$

$$- 2\sum_i \pi_\theta(i)(r(i) - \pi_\theta^\top r) y(i) \sum_j \pi_\theta(j) y(j)\Bigg| \tag{D.247}$$

$$= \left|(H(\pi_\theta) r)^\top (y \odot y) - 2 \cdot (H(\pi_\theta) r)^\top y \cdot \left(\pi_\theta^\top y\right)\right| \tag{D.248}$$

$$\leq \|H(\pi_\theta) r\|_\infty \cdot \|y \odot y\|_1 + 2 \cdot \|H(\pi_\theta) r\|_2 \cdot \|y\|_2 \cdot \|\pi_\theta\|_1 \cdot \|y\|_\infty \tag{D.249}$$

$$\leq 3 \cdot \|H(\pi_\theta) r\|_2 \cdot \|y\|_2^2 . \tag{D.250}$$

According to Taylor's theorem, $\forall \theta,\ \theta'$,

$$\left|(\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle\right| = \frac{1}{2} \cdot \left|(\theta' - \theta)^\top S(r, \theta_\zeta) (\theta' - \theta)\right| \tag{D.251}$$

$$\leq \frac{3}{2} \cdot \left\|H(\pi_{\theta_\zeta}) r\right\|_2 \cdot \|\theta' - \theta\|_2^2 \qquad \text{(by Eq. (D.245))} \tag{D.252}$$

$$= \frac{3}{2} \cdot \left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 \cdot \|\theta' - \theta\|_2^2 . \qquad \text{(by Lemma 55)} \qquad \square$$

**Lemma 22.** Let

$$\theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta} \Big/ \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 . \tag{D.253}$$

Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0, 1]$. We have, for all $\eta \in (0, 1/3)$,

$$\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 \leq \frac{1}{1 - 3\eta} \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 . \tag{D.254}$$

*Proof.* Denote $\zeta_1 \coloneqq \zeta$. Also denote $\theta_{\zeta_2} \coloneqq \theta + \zeta_2 \cdot (\theta_{\zeta_1} - \theta)$ with some $\zeta_2 \in [0, 1]$.

We have,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_\theta^\top r}{d\theta}\right\|_2 = \left\|\int_0^1 \left\langle \frac{d^2\{\pi_{\theta_{\zeta_2}}^\top r\}}{d\theta_{\zeta_2}^2}, \theta_{\zeta_1} - \theta \right\rangle d\zeta_2\right\|_2 \tag{D.255}$$

$$\leq \int_0^1 \left\|\frac{d^2\{\pi_{\theta_{\zeta_2}}^\top r\}}{d\theta_{\zeta_2}^2}\right\|_2 \cdot \|\theta_{\zeta_1} - \theta\|_2 \, d\zeta_2 \tag{D.256}$$

$$\leq \int_0^1 3 \cdot \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 \cdot \zeta_1 \cdot \|\theta' - \theta\|_2 \, d\zeta_2 \qquad \text{(by Eq. (D.245))} \tag{D.257}$$

$$\leq \int_0^1 3 \cdot \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 \cdot \eta \, d\zeta_2, \tag{D.258}$$

$$\left(\zeta_1 \in [0,1], \text{ using } \theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta} \Big/ \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\right) \tag{D.259}$$

where the second last inequality is because of the Hessian is symmetric, and its operator norm is equal to its spectral radius. Therefore we have,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}}\right\|_2 \tag{D.260}$$

$$\leq \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 + \left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \qquad \text{(by triangle inequality)} \tag{D.261}$$

$$\leq \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 + 3\eta \cdot \int_0^1 \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 d\zeta_2. \qquad \text{(by Eq. (D.255))} \tag{D.262}$$

Denote $\theta_{\zeta_3} \coloneqq \theta + \zeta_3 \cdot (\theta_{\zeta_2} - \theta)$ with some $\zeta_3 \in [0,1]$. Using similar calculation as in Eq. (D.255), we have,

$$\left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}}\right\|_2 \leq \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 + \left\|\frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} - \frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{D.263}$$

$$\leq \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 + 3\eta \cdot \int_0^1 \left\|\frac{d\pi_{\theta_{\zeta_3}}^\top r}{d\theta_{\zeta_3}}\right\|_2 d\zeta_3. \tag{D.264}$$

Combining Eqs. (D.260) and (D.263), we have,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}}\right\|_2 \leq (1+3\eta) \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{D.265}$$

$$+ (3\eta)^2 \cdot \int_0^1 \int_0^1 \left\|\frac{d\pi_{\theta_{\zeta_3}}^\top r}{d\theta_{\zeta_3}}\right\|_2 d\zeta_3 d\zeta_2, \tag{D.266}$$

which implies,

$$\left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \leq \left[ \sum_{i=0}^{\infty} (3\eta)^i \right] \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \tag{D.267}$$

$$= \frac{1}{1-3\eta} \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 . \qquad (\eta \in (0, 1/3)) \qquad \square$$

**Lemma 23** (Non-vanishing NŁ coefficient) **.** Using normalized policy gradient method, we have $\inf_{t\geq 1} \pi_{\theta_t}(a^*) > 0$.

*Proof.* The proof is similar to Mei et al. (2020b, Lemma 5). Let

$$c = \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right) \tag{D.268}$$

and

$$\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0 \tag{D.269}$$

denote the reward gap of $r$. We will prove that $\inf_{t\geq 1} \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(a^*)$, where $t_0 = \min\{t : \pi_{\theta_t}(a^*) \geq \frac{c}{c+1}\}$. Note that $t_0$ depends only on $\theta_1$ and $c$, and $c$ depends only on the problem. Define the following regions,

$$\mathcal{R}_1 = \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)},\ \forall a \neq a^* \right\}, \tag{D.270}$$

$$\mathcal{R}_2 = \left\{ \theta : \pi_\theta(a^*) \geq \pi_\theta(a),\ \forall a \neq a^* \right\}, \tag{D.271}$$

$$\mathcal{N}_c = \left\{ \theta : \pi_\theta(a^*) \geq \frac{c}{c+1} \right\}. \tag{D.272}$$

We make the following three-part claim.

**Claim 2.** *The following hold :*

**a)** *Following a normalized PG update* $\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$, *if* $\theta_t \in \mathcal{R}_1$, *then* (i) $\theta_{t+1} \in \mathcal{R}_1$ *and* (ii) $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$.

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

**c)** *For* $\eta = 1/6$, *there exists a finite time* $t_0 \geq 1$, *such that* $\theta_{t_0} \in \mathcal{N}_c$, *and thus* $\theta_{t_0} \in \mathcal{R}_1$, *which implies that* $\inf_{t\geq 1} \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(a^*)$..

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

$$\mathcal{R}_1(a) = \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)} \right\} . \tag{D.273}$$

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

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

$$\theta_{t+1}(a^*) = \theta_t(a^*) + \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \tag{D.274}$$

$$\geq \theta_t(a) + \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \tag{D.275}$$

$$= \theta_{t+1}(a), \tag{D.276}$$

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

$$\pi_{\theta_{t+1}}(a^*) \cdot \left( r(a^*) - \pi_{\theta_{t+1}}^\top r \right) \geq \pi_{\theta_{t+1}}(a) \cdot \left( r(a) - \pi_{\theta_{t+1}}^\top r \right) , \tag{D.277}$$

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

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

$$r(a^*) - r(a) \geq \left( 1 - \frac{\pi_\theta(a^*)}{\pi_\theta(a)} \right) \cdot \left( r(a^*) - \pi_\theta^\top r \right) . \tag{D.278}$$

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

$$\pi_\theta(a^*) \cdot \left( r(a^*) - \pi_\theta^\top r \right) \geq \pi_\theta(a) \cdot \left( r(a) - \pi_\theta^\top r \right) \tag{D.279}$$

$$= \pi_\theta(a) \cdot \left( r(a^*) - \pi_\theta^\top r \right) - \pi_\theta(a) \cdot \left( r(a^*) - r(a) \right) , \tag{D.280}$$

which, after rearranging, is equivalent to Eq. (D.278). Hence, it suffices to show that Eq. (D.278) holds for $\theta_{t+1}$ provided it holds for $\theta_t$. From the latter condition, we get

$$r(a^*) - r(a) \geq \left( 1 - \exp\left\{ \theta_t(a^*) - \theta_t(a) \right\} \right) \cdot \left( r(a^*) - \pi_{\theta_t}^\top r \right) . \tag{D.281}$$

After an update of the parameters, according to Lemma 52 (or Eq. (D.302) below), $\pi_{\theta_{t+1}}^\top r \geq \pi_{\theta_t}^\top r$, i.e.,

$$0 < r(a^*) - \pi_{\theta_{t+1}}^\top r \leq r(a^*) - \pi_{\theta_t}^\top r\,. \tag{D.282}$$

On the other hand,

$$\begin{aligned}
\theta_{t+1}(a^*) - \theta_{t+1}(a) &= \theta_t(a^*) + \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} && \text{(D.283)}\\
&\quad - \theta_t(a) - \frac{\eta}{\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2} \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} && \text{(D.284)}\\
&\geq \theta_t(a^*) - \theta_t(a), && \text{(D.285)}
\end{aligned}$$

which implies that

$$1 - \exp\left\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\right\} \leq 1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\}. \tag{D.286}$$

Furthermore, by our assumption that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$, we have

$$1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\} = 1 - \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} > 0. \tag{D.287}$$

Putting things together, we get

$$\begin{aligned}
&\left(1 - \exp\left\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\right\}\right) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) && \text{(D.288)}\\
&\quad \leq \left(1 - \exp\left\{\theta_t(a^*) - \theta_t(a)\right\}\right) \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right) && \text{(D.289)}\\
&\quad \leq r(a^*) - r(a), && \text{(D.290)}
\end{aligned}$$

which is equivalent to

$$\left(1 - \frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)}\right) \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right) \leq r(a^*) - r(a), \tag{D.291}$$

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

Part (ii): Assume again that $\theta_t \in \mathcal{R}_1$. We want to show that $\pi_{\theta_{t+1}}(a^*) \geq$

$\pi_{\theta_t}(a^*)$. Since $\theta_t \in \mathcal{R}_1$, we have $\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \geq \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}$, $\forall a \neq a^*$. Hence,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\{\theta_{t+1}(a^*)\}}{\sum_a \exp\{\theta_{t+1}(a)\}} \tag{D.292}$$

$$= \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}} \tag{D.293}$$

$$\geq \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\right\}} \tag{D.294}$$

$$\left(\text{using } \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \geq \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right) \tag{D.295}$$

$$= \frac{\exp\{\theta_t(a^*)\}}{\sum_a \exp\{\theta_t(a)\}} = \pi_{\theta_t}(a^*). \tag{D.296}$$

**Claim b); Claim c)** The proof of those claims are exactly the same as Lemma 5, since they do not involve the update rule. □

**Theorem 17.** Using normalized PG $\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2$, with $\eta = 1/6$, for all $t \geq 1$, we have,

$$(\pi^* - \pi_{\theta_t})^\top r \leq e^{-\frac{c\cdot(t-1)}{12}} \cdot (\pi^* - \pi_{\theta_1})^\top r, \tag{D.297}$$

where $c = \inf_{t\geq 1} \pi_{\theta_t}(a^*) > 0$ is from Lemma 23, and $c$ is a constant that depends on $r$ and $\theta_1$, but not on the time $t$.

*Proof.* Denote $\theta_{\zeta_t} \coloneqq \theta_t + \zeta_t \cdot (\theta_{t+1} - \theta_t)$ with some $\zeta_t \in [0, 1]$. According to Lemma 21,

$$\left|(\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle\right| \tag{D.298}$$

$$\leq \frac{3}{2} \cdot \left\|\frac{d\pi_{\theta_{\zeta_t}}^\top r}{d\theta_{\zeta_t}}\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{D.299}$$

$$\leq \frac{3}{2} \cdot \frac{1}{1-3\eta} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{D.300}$$

$$(\eta = 1/6, \text{ by Lemma 22}) \tag{D.301}$$

which implies,

$$\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r \tag{D.302}$$

$$\leq -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{3}{2\cdot(1-3\eta)} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{D.303}$$

$$= -\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + \frac{3\cdot\eta^2}{2\cdot(1-3\eta)} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \tag{D.304}$$

$$\left( \text{using } \theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \Big/ \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \right) \tag{D.305}$$

$$= -\frac{1}{12} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \qquad (\text{using } \eta = 1/6) \tag{D.306}$$

$$\leq -\frac{1}{12} \cdot \pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r \qquad (\text{by Lemma 3}) \tag{D.307}$$

$$\leq -\frac{1}{12} \cdot \inf_{t\geq 1} \pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r. \tag{D.308}$$

According to Eq. (D.302), we have,

$$(\pi^* - \pi_{\theta_t})^\top r \tag{D.309}$$

$$\leq \left(1 - \frac{c}{12}\right) \cdot \left(\pi^* - \pi_{\theta_{t-1}}\right)^\top r \qquad \left(c \coloneqq \inf_{t\geq 1} \pi_{\theta_t}(a^*) > 0\right) \tag{D.310}$$

$$\leq \exp\{-c/12\} \cdot \left(\pi^* - \pi_{\theta_{t-1}}\right)^\top r \tag{D.311}$$

$$\leq \exp\{-(t-1)\cdot c/12\} \cdot (\pi^* - \pi_{\theta_1})^\top r. \qquad \square$$

### D.2.2 General MDPs

**Lemma 24** (NS) **.** Let Assumption 2 hold and denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. $\theta \mapsto V^{\pi_\theta}(\mu)$ satisfies $\beta(\theta_\zeta)$ non-uniform smoothness with

$$\beta(\theta_\zeta) = \left[3 + \frac{2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma}\right] \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta} \right\|_2, \tag{D.312}$$

where $C_\infty \coloneqq \max_\pi \left\| \frac{d_\mu^\pi}{\mu} \right\|_\infty \leq \frac{1}{\min_s \mu(s)} < \infty$.

*Proof.* The main part is to prove that for all $y \in \mathbb{R}^{SA}$ and $\theta$,

$$\left| y^\top \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} y \right| \tag{D.313}$$

$$\leq \left[3 + \frac{2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma}\right] \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \cdot \|y\|_2^2. \tag{D.314}$$

We first calculate the second order derivative of $V^{\pi_\theta}(\mu)$ w.r.t. $\theta$.

Denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. For any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} = \Big\langle \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha}\Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \Big\rangle \tag{D.315}$$

$$= \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta}, u \Big\rangle \tag{D.316}$$

$$= \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \Big\rangle, \qquad \left( \frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = \mathbf{0},\ \forall s' \neq s \right) \tag{D.317}$$

Similarly, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} = \Big\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \Big\rangle \tag{D.318}$$

$$= \Big\langle \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha^2}\Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \Big\rangle \tag{D.319}$$

$$= \Big\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \Big\rangle. \tag{D.320}$$

Define $\Pi(\alpha) \in \mathbb{R}^{S \times SA}$ as follows,

$$\Pi(\alpha) \coloneqq \begin{bmatrix} \pi_{\theta_\alpha}(\cdot|1)^\top & \mathbf{0}^\top & \cdots & \mathbf{0}^\top \\ \mathbf{0}^\top & \pi_{\theta_\alpha}(\cdot|2)^\top & \cdots & \mathbf{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^\top & \mathbf{0}^\top & \cdots & \pi_{\theta_\alpha}(\cdot|S)^\top \end{bmatrix}. \tag{D.321}$$

Denote $\mathcal{P} \in \mathbb{R}^{SA \times S}$ such that,

$$\mathcal{P}_{(sa,s')} \coloneqq \mathcal{P}(s'|s,a). \tag{D.322}$$

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

$$[P(\alpha)]_{(s,s')} = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \mathcal{P}(s'|s,a). \tag{D.323}$$

The derivative w.r.t. $\alpha$ is

$$\frac{\partial P(\alpha)}{\partial \alpha} = \frac{\partial \Pi(\alpha)\mathcal{P}}{\partial \alpha} = \frac{\partial \Pi(\alpha)}{\partial \alpha}\mathcal{P}. \tag{D.324}$$

And $\forall (s,s')$, we have,

$$\left[ \frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{D.325}$$

Next, consider the state value function of $\pi_{\theta_\alpha}$,

$$V^{\pi_{\theta_\alpha}}(s) = \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a) \tag{D.326}$$

$$+ \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot V^{\pi_{\theta_\alpha}}(s'), \tag{D.327}$$

which implies,

$$V^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) r_{\theta_\alpha} \tag{D.328}$$

$$V^{\pi_{\theta_\alpha}}(\mu) = \mu^\top M(\alpha) r_{\theta_\alpha}, \tag{D.329}$$

where

$$M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1}, \tag{D.330}$$

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

$$r_{\theta_\alpha} = \Pi(\alpha) r, \tag{D.331}$$

where $r \in \mathbb{R}^{SA}$. Taking derivative w.r.t. $\alpha$ in Eq. (D.329),

$$\frac{\partial V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha} = \gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \mu^\top M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \tag{D.332}$$

$$= \mu^\top M(\alpha) \left[ \gamma \cdot \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \right] \tag{D.333}$$

$$= \mu^\top M(\alpha) \left[ \gamma \cdot \frac{\partial \Pi(\alpha)}{\partial \alpha} \mathcal{P} M(\alpha) r_{\theta_\alpha} + \frac{\partial \Pi(\alpha)}{\partial \alpha} r \right] \tag{D.334}$$

(by Eqs. (D.324) and (D.331)) (D.335)

$$= \mu^\top M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}, \tag{D.336}$$

where $Q^{\pi_{\theta_\alpha}} \in \mathbb{R}^{SA}$ is the state-action value and it satisfies,

$$Q^{\pi_{\theta_\alpha}} = r + \gamma \cdot \mathcal{P} M(\alpha) r_{\theta_\alpha} \tag{D.337}$$

$$= r + \gamma \cdot \mathcal{P} V^{\pi_{\theta_\alpha}}. \qquad \text{(by Eq. (D.328))} \tag{D.338}$$

Similarly, taking second derivative w.r.t. $\alpha$,

$$\frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha^2} = 2\gamma^2 \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \tag{D.339}$$

$$+ \gamma \cdot \mu^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \tag{D.340}$$

$$+ 2\gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + \mu^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \tag{D.341}$$

$$= 2\gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} \left(\gamma \cdot \mathcal{P} M(\alpha) r_{\theta_\alpha} + r\right) \tag{D.342}$$

$$+ \mu^\top M(\alpha) \frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} \left(\gamma \cdot \mathcal{P} M(\alpha) r_{\theta_\alpha} + r\right) \tag{D.343}$$

$$= 2\gamma \cdot \mu^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \tag{D.344}$$

$$+ \mu^\top M(\alpha) \frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}}. \tag{D.345}$$

For the last term, we have,

$$\left[\frac{\partial^2 \Pi(\alpha)}{\partial \alpha^2} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s)} = \sum_a \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \cdot Q^{\pi_\theta}(s,a) \tag{D.346}$$

$$= \sum_a \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \cdot Q^{\pi_\theta}(s,a) \tag{D.347}$$

$$\text{(by Eq. (D.318))} \tag{D.348}$$

$$= u(s,\cdot)^\top \left[\sum_a \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} \cdot Q^{\pi_\theta}(s,a)\right] u(s,\cdot). \tag{D.349}$$

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

$$S_{(i,j)} = \frac{\partial \{\delta_{ia}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(i|s)\}}{\partial \theta(s,j)} \tag{D.350}$$

$$= \delta_{ia} \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] \tag{D.351}$$

$$- \pi_\theta(a|s) \cdot [\delta_{ij}\pi_\theta(j|s) - \pi_\theta(i|s)\pi_\theta(j|s)] \tag{D.352}$$

$$- \pi_\theta(i|s) \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)], \tag{D.353}$$

where the $\delta$ notation is as defined in Eq. (D.201). Then we have,

$$\left[\sum_a \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} \cdot Q^{\pi_\theta}(s,a)\right]_{(i,j)} = \sum_a S_{(i,j)} \cdot Q^{\pi_\theta}(s,a) \tag{D.354}$$

$$= \delta_{ij} \cdot \pi_\theta(i|s) \cdot [Q^{\pi_\theta}(s,i) - V^{\pi_\theta}(s)] \tag{D.355}$$

$$- \pi_\theta(i|s) \cdot \pi_\theta(j|s) \cdot [Q^{\pi_\theta}(s,i) - V^{\pi_\theta}(s)] \tag{D.356}$$

$$- \pi_\theta(i|s) \cdot \pi_\theta(j|s) \cdot [Q^{\pi_\theta}(s,j) - V^{\pi_\theta}(s)]. \tag{D.357}$$

Therefore we have,

$$\left[\frac{\partial^2\Pi(\alpha)}{\partial\alpha^2}Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s)} \tag{D.358}$$

$$= \sum_{i=1}^{A}\sum_{j=1}^{A} u(s,i)\cdot u(s,j)\cdot\left[\sum_a \frac{\partial^2\pi_\theta(a|s)}{\partial\theta^2(s,\cdot)}\cdot Q^{\pi_\theta}(s,a)\right]_{(i,j)} \tag{D.359}$$

$$= \left(H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right)^\top \left(u(s,\cdot)\odot u(s,\cdot)\right) \tag{D.360}$$

$$- 2\cdot\left[\left(H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right)^\top u(s,\cdot)\right]\cdot\left(\pi_\theta(\cdot|s)^\top u(s,\cdot)\right), \tag{D.361}$$

where $H(\pi) \coloneqq \mathrm{diag}(\pi) - \pi\pi^\top$. Combining the above results with Eq. (D.339), we have,

$$\left|\mu^\top M(\alpha)\frac{\partial^2\Pi(\alpha)}{\partial\alpha^2}Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right| \tag{D.362}$$

$$\le \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\left|\left[\frac{\partial^2\Pi(\alpha)}{\partial\alpha^2}Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s)}\right| \tag{D.363}$$

(by triangle inequality) (D.364)

$$\le \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot 3\cdot\left\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right\|_2\cdot\|u\|_2^2 \tag{D.365}$$

(by Hölder's inequality) (D.366)

$$\le \frac{3\cdot\sqrt{S}}{1-\gamma}\cdot\left[\sum_s d_\mu^{\pi_\theta}(s)^2\cdot\left\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\right\|_2^2\right]^{\frac{1}{2}}\cdot\|u\|_2^2 \tag{D.367}$$

(by Cauchy-Schwarz) (D.368)

$$= 3\cdot\sqrt{S}\cdot\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2\cdot\|u\|_2^2. \qquad \text{(by Lemma 55)} \tag{D.369}$$

For the first term in Eq. (D.339), we have,

$$\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial\alpha}M(\alpha)\frac{\partial\Pi(\alpha)}{\partial\alpha}Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0} \tag{D.370}$$

$$= \sum_{s'}\left[\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial\alpha}\Big|_{\alpha=0}\right]_{(s')}\cdot\left[M(\alpha)\frac{\partial\Pi(\alpha)}{\partial\alpha}Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s')}, \tag{D.371}$$

since,

$$\left(\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial\alpha}\right)^\top \in \mathbb{R}^S, \quad \text{and} \quad M(\alpha)\frac{\partial\Pi(\alpha)}{\partial\alpha}Q^{\pi_{\theta_\alpha}} \in \mathbb{R}^S. \tag{D.372}$$

Next we have,

$$\left[M(\alpha)\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s')} \tag{D.373}$$

$$= \frac{1}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \left[\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s)} \tag{D.374}$$

$$\left(\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}} \in \mathbb{R}^S\right) \tag{D.375}$$

$$= \frac{1}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \sum_a \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \cdot Q^{\pi_\theta}(s,a) \tag{D.376}$$

$$= \frac{1}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \sum_a \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot)\right\rangle \cdot Q^{\pi_\theta}(s,a) \tag{D.377}$$

(by Eq. (D.315)) (D.378)

$$= \frac{1}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \left\langle \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a), u(s,\cdot)\right\rangle \tag{D.379}$$

$$= \frac{1}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \left(H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot)\right)^\top u(s,\cdot), \tag{D.380}$$

($H(\pi_\theta)$ is the Jacobian of $\theta \mapsto \text{softmax}(\theta)$) (D.381)

which implies,

$$\left|\left[M(\alpha)\frac{\partial \Pi(\alpha)}{\partial \alpha} Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right]_{(s')}\right| \tag{D.382}$$

$$\leq \frac{1}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \left\|H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot)\right\|_2 \cdot \|u(s,\cdot)\|_2 \tag{D.383}$$

$$\leq \frac{\|u\|_2}{1-\gamma} \cdot \sum_s d_{s'}^{\pi_\theta}(s) \cdot \left\|H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot)\right\|_2. \tag{D.384}$$

On the other hand,

$$\left[\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0}\right]_{(s')} = \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\left[\frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0}\right]_{(s,s')} \tag{D.385}$$

$$\left(\frac{\partial P(\alpha)}{\partial \alpha}\in\mathbb{R}^{S\times S}\right) \tag{D.386}$$

$$= \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\sum_a\left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0}\right]\cdot\mathcal{P}(s'|s,a) \tag{D.387}$$

(by Eq. (D.325)) (D.388)

$$= \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\sum_a\left\langle\frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot)\right\rangle\cdot\mathcal{P}(s'|s,a) \tag{D.389}$$

(by Eq. (D.315)) (D.390)

$$= \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\sum_a \pi_\theta(a|s)\cdot\mathcal{P}(s'|s,a) \tag{D.391}$$

$$\cdot\left[u(s,a)-\pi_\theta(\cdot|s)^\top u(s,\cdot)\right], \tag{D.392}$$

which implies,

$$\left|\left[\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0}\right]_{(s')}\right| \tag{D.393}$$

$$\leq \frac{1}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\sum_a \pi_\theta(a|s)\cdot\mathcal{P}(s'|s,a)\cdot 2\cdot\|u(s,\cdot)\|_\infty \tag{D.394}$$

$$\leq \frac{2\cdot\|u\|_2}{1-\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\sum_a \pi_\theta(a|s)\cdot\mathcal{P}(s'|s,a). \tag{D.395}$$

According to

$$d_\mu^{\pi_\theta}(s') = (1-\gamma)\cdot\mu(s') \tag{D.396}$$

$$+\gamma\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\sum_a \pi_\theta(a|s)\cdot\mathcal{P}(s'|s,a),\quad \forall s'\in\mathcal{S} \tag{D.397}$$

we have,

$$\left|\left[\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}\Big|_{\alpha=0}\right]_{(s')}\right| \tag{D.398}$$

$$\leq \frac{2\cdot\|u\|_2}{(1-\gamma)\cdot\gamma}\cdot\left[d_\mu^{\pi_\theta}(s')-(1-\gamma)\cdot\mu(s')\right] \tag{D.399}$$

$$= \frac{2\cdot\|u\|_2}{(1-\gamma)\cdot\gamma}\cdot\left[\frac{d_\mu^{\pi_\theta}(s')}{\mu(s')}\cdot\mu(s')-(1-\gamma)\cdot\mu(s')\right] \tag{D.400}$$

$$\leq \frac{2\cdot\|u\|_2}{(1-\gamma)\cdot\gamma}\cdot(C_\infty-(1-\gamma))\cdot\mu(s'). \tag{D.401}$$

$$\left(C_\infty \coloneqq \max_\pi\left\|\frac{d_\mu^\pi}{\mu}\right\|_\infty < \left\|\frac{1}{\mu}\right\|_\infty < \infty\right) \tag{D.402}$$

Combining Eqs. (D.370), (D.382) and (D.398), we have,

$$\left|\mu^\top M(\alpha)\frac{\partial P(\alpha)}{\partial \alpha}M(\alpha)\frac{\partial \Pi(\alpha)}{\partial \alpha}Q^{\pi_{\theta_\alpha}}\Big|_{\alpha=0}\right| \tag{D.403}$$

$$\leq \sum_{s'}\frac{2\cdot\|u\|_2}{(1-\gamma)\cdot\gamma}\cdot(C_\infty-(1-\gamma))\cdot\mu(s')\cdot\frac{\|u\|_2}{1-\gamma} \tag{D.404}$$

$$\cdot\sum_s d_{s'}^{\pi_\theta}(s)\cdot\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\|_2 \tag{D.405}$$

$$= \frac{2\cdot(C_\infty-(1-\gamma))}{(1-\gamma)^2\cdot\gamma}\cdot\sum_s d_\mu^{\pi_\theta}(s)\cdot\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\|_2\cdot\|u\|_2^2 \tag{D.406}$$

$$\leq \frac{2\cdot(C_\infty-(1-\gamma))\cdot\sqrt{S}}{(1-\gamma)^2\cdot\gamma} \tag{D.407}$$

$$\cdot\left[\sum_s d_\mu^{\pi_\theta}(s)^2\cdot\|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\|_2^2\right]^{\frac{1}{2}}\cdot\|u\|_2^2 \tag{D.408}$$

(by Cauchy-Schwarz) (D.409)

$$= \frac{2\cdot(C_\infty-(1-\gamma))\cdot\sqrt{S}}{(1-\gamma)\cdot\gamma}\cdot\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2\cdot\|u\|_2^2. \tag{D.410}$$

(by Lemma 55) (D.411)

Combining Eqs. (D.339), (D.362) and (D.403),

$$\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial\alpha^2}\Big|_{\alpha=0}\right| \tag{D.412}$$

$$\leq\left[3+\frac{2\cdot(C_\infty-(1-\gamma))}{(1-\gamma)\cdot\gamma}\right]\cdot\sqrt{S}\cdot\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2\cdot\|u\|_2^2, \tag{D.413}$$

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

$$\left| y^\top \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} y \right| = \left| \left( \frac{y}{\|y\|_2} \right)^\top \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} \left( \frac{y}{\|y\|_2} \right) \right| \cdot \|y\|_2^2 \tag{D.414}$$

$$\leq \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 V^{\pi_\theta}(\mu)}{\partial \theta^2} u, u \right\rangle \right| \cdot \|y\|_2^2 \tag{D.415}$$

$$= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \tag{D.416}$$

$$= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \tag{D.417}$$

$$= \max_{\|u\|_2=1} \left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(\mu)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \tag{D.418}$$

$$\leq \left[ 3 + \frac{2 \cdot (C_\infty - (1-\gamma))}{(1-\gamma) \cdot \gamma} \right] \cdot \sqrt{S} \cdot \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \cdot \|y\|_2^2. \tag{D.419}$$

(by Eq. (D.412)) (D.420)

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

$$\left| V^{\pi_{\theta'}}(\mu) - V^{\pi_\theta}(\mu) - \left\langle \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}, \theta' - \theta \right\rangle \right| \tag{D.421}$$

$$= \frac{1}{2} \cdot \left| (\theta' - \theta)^\top \frac{\partial^2 V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta^2} (\theta' - \theta) \right| \tag{D.422}$$

$$\leq \frac{3 \cdot (1-\gamma) \cdot \gamma + 2 \cdot (C_\infty - (1-\gamma))}{2 \cdot (1-\gamma) \cdot \gamma} \cdot \sqrt{S} \tag{D.423}$$

$$\cdot \left\| \frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta} \right\|_2 \cdot \|\theta' - \theta\|_2^2, \qquad \text{(by Eq. (D.414))} \tag{D.424}$$

thus finishing the proof. □

**Lemma 25.** Let $\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}}$ and

$$\theta' = \theta + \eta \cdot \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \Big/ \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2. \tag{D.425}$$

Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. We have,

$$\left\| \frac{\partial V^{\pi_{\theta_\zeta}}(\mu)}{\partial \theta_\zeta} \right\|_2 \leq 2 \cdot \left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2. \tag{D.426}$$

*Proof.* Using the similar arguments of Lemma 22 (replacing 3 in Lemma 21 with $\frac{3\cdot(1-\gamma)\cdot\gamma + 2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma} \cdot \sqrt{S}$ in Lemma 24), we have the results. □

**Lemma 26** (Non-vanishing NŁ coefficient) **.** Let Assumption 2 hold. We have, $c := \inf_{s\in\mathcal{S},t\geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$, where $\{\theta_t\}_{t\geq 1}$ is generated by Algorithm 2.

*Proof.* The proof is similar to Lemma 9 and is an extension of the proof for Lemma 23. Denote $\Delta^*(s) = Q^*(s,a^*(s)) - \max_{a\neq a^*(s)} Q^*(s,a) > 0$ as the optimal value gap of state $s$, where $a^*(s)$ is the action that the optimal policy selects under state $s$, and $\Delta^* = \min_{s\in\mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. For each state $s \in \mathcal{S}$, define the following sets:

$$\mathcal{R}_1(s) = \left\{\theta : \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a^*(s))} \geq \frac{\partial V^{\pi_\theta}(\mu)}{\partial\theta(s,a)},\ \forall a \neq a^*\right\}, \tag{D.427}$$

$$\mathcal{R}_2(s) = \left\{\theta : Q^{\pi_\theta}(s,a^*(s)) \geq Q^*(s,a^*(s)) - \Delta^*(s)/2\right\}, \tag{D.428}$$

$$\mathcal{R}_3(s) = \big\{\theta_t : V^{\pi_{\theta_t}}(s) \geq Q^{\pi_{\theta_t}}(s,a^*(s)) - \Delta^*(s)/2, \tag{D.429}$$

$$\text{for all } t \geq 1 \text{ large enough}\big\}, \tag{D.430}$$

$$\mathcal{N}_c(s) = \left\{\theta : \pi_\theta(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}\right\}, \tag{D.431}$$

$$\text{where } c(s) = \frac{A}{(1-\gamma)\cdot\Delta^*(s)} - 1. \tag{D.432}$$

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

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

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

**Claim III.** There exists a finite time $t_0(s) \geq 1$, such that $\theta_{t_0(s)} \in \mathcal{N}_c(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, and thus $\theta_{t_0(s)} \in \mathcal{R}_1(s)\cap\mathcal{R}_2(s)\cap\mathcal{R}_3(s)$, which implies $\inf_{t\geq 1}\pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0(s)}\pi_{\theta_t}(a^*(s)|s)$.

**Claim IV.** Define $t_0 = \max_s t_0(s)$. Then, we have

$$\inf_{s\in\mathcal{S},t\geq 1}\pi_{\theta_t}(a^*(s)|s) = \min_{1\leq t\leq t_0}\min_s \pi_{\theta_t}(a^*(s)|s). \tag{D.433}$$

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

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

$$Q^{\pi_{\theta_t}}(s, a^*(s)) \geq Q^*(s, a^*(s)) - \Delta^*(s)/2. \tag{D.434}$$

According to monotonic improvement of Eq. (D.500), we have $V^{\pi_{\theta_{t+1}}}(s') \geq V^{\pi_{\theta_t}}(s')$, and

$$\begin{aligned}
Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) &= Q^{\pi_{\theta_t}}(s, a^*(s)) + Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a^*(s)) && \text{(D.435)}\\
&= Q^{\pi_{\theta_t}}(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s)) \cdot [V^{\pi_{\theta_{t+1}}}(s') - V^{\pi_{\theta_t}}(s')] && \text{(D.436)}\\
&\geq Q^{\pi_{\theta_t}}(s, a^*(s)) + 0 && \text{(D.437)}\\
&\geq Q^*(s, a^*(s)) - \Delta^*(s)/2, && \text{(D.438)}
\end{aligned}$$

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

$$\begin{aligned}
&Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) && \text{(D.439)}\\
&= Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^*(s, a^*(s)) + Q^*(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) && \text{(D.440)}\\
&\geq -\Delta^*(s)/2 + Q^*(s, a^*(s)) - Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) && \text{(D.441)}\\
&\geq -\Delta^*(s)/2 + Q^*(s, a^*(s)) && \text{(D.442)}\\
&\quad - \max_{a \neq a^*(s)} Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) && \text{(D.443)}\\
&= -\Delta^*(s)/2 + \Delta^*(s) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \cdot [V^*(s') - V^{\pi_{\theta_t}}(s')] && \text{(D.444)}\\
&\geq -\Delta^*(s)/2 + \Delta^*(s) + 0 && \text{(D.445)}\\
&= \Delta^*(s)/2. && \text{(D.446)}
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot A^{\pi_{\theta_t}}(s, a) && \text{(D.447)}\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)]\,. && \text{(D.448)}
\end{aligned}$$

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

$$\begin{aligned}
&\pi_{\theta_t}(a^*(s)|s) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] && \text{(D.449)}\\
&\quad \geq \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)]\,. && \text{(D.450)}
\end{aligned}$$

Similarly to the first part in the proof for Lemma 23. There are two cases. Case (a): If $\pi_{\theta_t}(a^*(s)|s) \geq \pi_{\theta_t}(a|s)$, then $\theta_t(s, a^*(s)) \geq \theta_t(s, a)$. After an update of the parameters,

$$\theta_{t+1}(s, a^*(s)) = \theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \tag{D.451}$$

$$\geq \theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \tag{D.452}$$

$$= \theta_{t+1}(s, a), \tag{D.453}$$

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

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \tag{D.454}$$

$$\geq \pi_{\theta_{t+1}}(a|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)], \tag{D.455}$$

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

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

$$\pi_{\theta_t}(a^*(s)|s) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] \tag{D.456}$$

$$\geq \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)] \tag{D.457}$$

$$= \pi_{\theta_t}(a|s) \cdot \big[Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s) \tag{D.458}$$

$$+ Q^{\pi_{\theta_t}}(s, a) - Q^{\pi_{\theta_t}}(s, a^*(s))\big], \tag{D.459}$$

which, after rearranging, is equivalent to

$$Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) \tag{D.460}$$

$$\geq \left(1 - \frac{\pi_{\theta_t}(a^*(s)|s)}{\pi_{\theta_t}(a|s)}\right) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] \tag{D.461}$$

$$= (1 - \exp\{\theta_t(s, a^*(s)) - \theta_t(s, a)\}) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)]. \tag{D.462}$$

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

$$Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \leq \Delta^*(s)/2 \tag{D.463}$$

$$\leq Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a). \tag{D.464}$$

On the other hand,

$$\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a) \tag{D.465}$$

$$= \theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 - \theta_t(s, a) \tag{D.466}$$

$$- \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \tag{D.467}$$

$$\geq \theta_t(s, a^*(s)) - \theta_t(s, a), \tag{D.468}$$

which implies

$$1 - \exp\left\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\right\} \tag{D.469}$$

$$\leq 1 - \exp\left\{\theta_t(s, a^*(s)) - \theta_t(s, a)\right\}. \tag{D.470}$$

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

$$\left(1 - \exp\left\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\right\}\right) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)\right] \tag{D.471}$$

$$\leq Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a), \tag{D.472}$$

which after rearranging is equivalent to

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)\right] \tag{D.473}$$

$$\geq \pi_{\theta_{t+1}}(a|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)\right], \tag{D.474}$$

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

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

update of the parameters,

$$\pi_{\theta_{t+1}}(a^*(s)|s) = \frac{\exp\{\theta_{t+1}(s,a^*(s))\}}{\sum_a \exp\{\theta_{t+1}(s,a)\}} \tag{D.475}$$

$$= \frac{\exp\left\{\theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}}{\sum_a \exp\left\{\theta_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}} \tag{D.476}$$

$$\geq \frac{\exp\left\{\theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}}{\sum_a \exp\left\{\theta_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \Big/ \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \right\}} \tag{D.477}$$

$$\left(\text{because } \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}\right) \tag{D.478}$$

$$= \frac{\exp\{\theta_t(s,a^*(s))\}}{\sum_a \exp\{\theta_t(s,a)\}} = \pi_{\theta_t}(a^*(s)|s). \tag{D.479}$$

**Claim II, Claim III, Claim IV.** The proof of those claims are exactly the same as Lemma 9, since they do not involve the update rule. □

**Theorem 18.** Let Assumption 2 hold and let $\{\theta_t\}_{t\geq 1}$ be generated using Algorithm 2 with learning rate

$$\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}}, \tag{D.480}$$

where $C_\infty := \max_\pi \left\| \frac{d_\mu^\pi}{\mu} \right\|_\infty$. Denote $C'_\infty := \max_\pi \left\| \frac{d_\rho^\pi}{\mu} \right\|_\infty$. Let $c$ be the positive constant from Lemma 26. We have, for all $t \geq 1$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{(V^*(\mu) - V^{\pi_{\theta_1}}(\mu))\cdot C'_\infty}{1-\gamma} \cdot e^{-C\cdot(t-1)}, \tag{D.481}$$

where

$$C = \frac{(1-\gamma)^2\cdot\gamma\cdot c}{12\cdot(1-\gamma)\cdot\gamma + 8\cdot(C_\infty - (1-\gamma))} \cdot \frac{1}{S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1}. \tag{D.482}$$

*Proof.* First note that for any $\theta$ and $\mu$,

$$d_\mu^{\pi_\theta}(s) = \mathop{\mathbb{E}}_{s_0\sim\mu}\left[d_\mu^{\pi_\theta}(s)\right] \tag{D.483}$$

$$= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\cdot\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s|s_0,\pi_\theta,\mathcal{P})\right] \tag{D.484}$$

$$\geq \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\cdot\Pr(s_0 = s|s_0)\right] \tag{D.485}$$

$$= (1-\gamma)\cdot\mu(s). \tag{D.486}$$

Next, according to Lemma 36, we have,

$$V^*(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{D.487}$$

$$= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{D.488}$$

$$\leq \frac{1}{1-\gamma} \cdot \left\| \frac{d_\rho^{\pi_\theta}}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{D.489}$$

$$\left( \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \geq 0 \right) \tag{D.490}$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{d_\rho^{\pi_\theta}}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{D.491}$$

$$\left( \text{by Eq. (B.346) and } \min_s \mu(s) > 0 \right) \tag{D.492}$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot C'_\infty \cdot \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \tag{D.493}$$

$$= \frac{1}{1-\gamma} \cdot C'_\infty \cdot [V^*(\mu) - V^{\pi_\theta}(\mu)] . \qquad \text{(by Lemma 36)} \tag{D.494}$$

Denote $\theta_{\zeta_t} \coloneqq \theta_t + \zeta_t \cdot (\theta_{t+1} - \theta_t)$ with some $\zeta_t \in [0,1]$. And note $\eta = \frac{(1-\gamma)\cdot\gamma}{6\cdot(1-\gamma)\cdot\gamma + 4\cdot(C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}}$. According to Lemma 24, we have,

$$\left| V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \tag{D.495}$$

$$\leq \frac{3\cdot(1-\gamma)\cdot\gamma + 2\cdot(C_\infty - (1-\gamma))}{2\cdot(1-\gamma)\cdot\gamma} \cdot \sqrt{S} \tag{D.496}$$

$$\cdot \left\| \frac{\partial V^{\pi_{\theta_{\zeta_t}}}(\mu)}{\partial \theta_{\zeta_t}} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{D.497}$$

$$\leq \frac{3\cdot(1-\gamma)\cdot\gamma + 2\cdot(C_\infty - (1-\gamma))}{(1-\gamma)\cdot\gamma} \cdot \sqrt{S} \tag{D.498}$$

$$\cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2. \qquad \text{(by Lemma 25)} \tag{D.499}$$

Denote $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. We have,

$$\delta_{t+1} - \delta_t = V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) \tag{D.500}$$

$$\leq -\left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{D.501}$$

$$+ \frac{3 \cdot (1-\gamma) \cdot \gamma + 2 \cdot (C_\infty - (1-\gamma))}{(1-\gamma) \cdot \gamma} \cdot \sqrt{S} \tag{D.502}$$

$$\cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{D.503}$$

$$= -\frac{(1-\gamma) \cdot \gamma}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \tag{D.504}$$

(using the value of $\eta$) (D.505)

$$\leq -\frac{(1-\gamma) \cdot \gamma}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{\sqrt{S}} \tag{D.506}$$

$$\cdot \frac{\min_s \pi_{\theta_t}(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\mu^{\pi^*} / d_\mu^{\pi_{\theta_t}} \right\|_\infty} \cdot \delta_t \tag{D.507}$$

(by Lemma 8) (D.508)

$$\leq -\frac{(1-\gamma)^2 \cdot \gamma}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{S} \tag{D.509}$$

$$\cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot \inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) \cdot \delta_t, \tag{D.510}$$

where the last inequality is by $d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (D.483)). According to Lemma 26, $c = \inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$. Therefore we have,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \leq (V^*(\mu) - V^{\pi_{\theta_1}}(\mu)) \tag{D.511}$$

$$\cdot \exp\left\{ -\frac{(1-\gamma)^2 \cdot \gamma \cdot c \cdot (t-1)}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \right\}, \tag{D.512}$$

which leads to the final result,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{(V^*(\mu) - V^{\pi_{\theta_1}}(\mu)) \cdot C'_\infty}{1-\gamma} \tag{D.513}$$

$$\cdot \exp\left\{ -\frac{(1-\gamma)^2 \cdot \gamma \cdot c \cdot (t-1)}{12 \cdot (1-\gamma) \cdot \gamma + 8 \cdot (C_\infty - (1-\gamma))} \cdot \frac{1}{S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \right\}, \tag{D.514}$$

thus, finishing the proof. □

## D.3 Proofs for Section 4.7: Generalized Linear Models

**Lemma 27** (NŁ) **.** Denote

$$u(\theta) \coloneqq \min_{i\in[N]} \{\pi_i \cdot (1-\pi_i)\}, \text{ and} \tag{D.515}$$

$$v \coloneqq \min_{i\in[N]} \{\pi_i^* \cdot (1-\pi_i^*)\}. \tag{D.516}$$

We have,

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \geq C(\theta,\phi) \cdot \left[ \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}, \tag{D.517}$$

holds for all $\theta \in \mathbb{R}^d$, where

$$C(\theta,\phi) = 8 \cdot u(\theta) \cdot \min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi}, \tag{D.518}$$

and $\lambda_\phi$ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$.

*Proof.* Denote $\pi_i' \coloneqq \sigma(z_i')$, where $z_i' \coloneqq \phi_i^\top \theta + \zeta \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right)$ for some $\zeta \in [0,1]$. We have,

$$(\pi_i - \pi_i^*)^2 = (\pi_i - \pi_i^*) \cdot \frac{d\sigma(z_i')}{dz_i'} \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \tag{D.519}$$

$$\text{(by the mean value theorem)} \tag{D.520}$$

$$= \pi_i' \cdot (1-\pi_i') \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \tag{D.521}$$

$$\leq \frac{1}{4} \cdot (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right). \tag{D.522}$$

$$\left( x \cdot (1-x) \leq \frac{1}{4},\ \forall x \in [0,1];\ (\pi_i - \pi_i^*) \cdot \left(\phi_i^\top \theta - \phi_i^\top \theta^*\right) \geq 0 \right) \tag{D.523}$$

Therefore we have,

$$\frac{1}{N}\cdot\sum_{i=1}^{N}(\pi_i-\pi_i^*)^2\le\frac{1}{4N}\cdot\sum_{i=1}^{N}(\pi_i-\pi_i^*)\cdot\left(\phi_i^\top\theta-\phi_i^\top\theta^*\right) \tag{D.524}$$

$$\text{(by Eq. (D.519))} \tag{D.525}$$

$$=\frac{1}{4N}\cdot\sum_{i=1}^{N}\frac{1}{\pi_i\cdot(1-\pi_i)}\cdot\pi_i\cdot(1-\pi_i)\cdot(\pi_i-\pi_i^*) \tag{D.526}$$

$$\cdot\left(\phi_i^\top\theta-\phi_i^\top\theta^*\right) \tag{D.527}$$

$$\le\frac{1}{4N}\cdot\frac{1}{\min_i\pi_i\cdot(1-\pi_i)}\cdot\sum_{i=1}^{N}\pi_i\cdot(1-\pi_i)\cdot(\pi_i-\pi_i^*) \tag{D.528}$$

$$\cdot\left(\phi_i^\top\theta-\phi_i^\top\theta^*\right)\qquad\left((\pi_i-\pi_i^*)\cdot\left(\phi_i^\top\theta-\phi_i^\top\theta^*\right)\ge 0\right) \tag{D.529}$$

$$=\frac{1}{8}\cdot\frac{1}{\min_i\pi_i\cdot(1-\pi_i)} \tag{D.530}$$

$$\cdot\left(\frac{2}{N}\cdot\sum_{i=1}^{N}\pi_i\cdot(1-\pi_i)\cdot(\pi_i-\pi_i^*)\cdot\phi_i\right)^\top(\theta-\theta^*-c\cdot v_{\phi,\perp}) \tag{D.531}$$

$$=\frac{1}{8}\cdot\frac{1}{\min_i\pi_i\cdot(1-\pi_i)}\cdot\left(\frac{\partial\mathcal{L}(\theta)}{\partial\theta}\right)^\top(\theta-\theta^*-c\cdot v_{\phi,\perp}) \tag{D.532}$$

$$\left(\frac{\partial\mathcal{L}(\theta)}{\partial\theta}=\frac{2}{N}\cdot\sum_{i=1}^{N}\pi_i\cdot(1-\pi_i)\cdot(\pi_i-\pi_i^*)\cdot\phi_i\right) \tag{D.533}$$

$$\le\frac{1}{8}\cdot\frac{1}{\min_i\pi_i\cdot(1-\pi_i)}\cdot\left\|\frac{\partial\mathcal{L}(\theta)}{\partial\theta}\right\|_2\cdot\|\theta-\theta^*-c\cdot v_{\phi,\perp}\|_2 \tag{D.534}$$

$$\text{(by Cauchy-Schwarz)} \tag{D.535}$$

$$=\frac{1}{8}\cdot\frac{1}{u(\theta)}\cdot\left\|\frac{\partial\mathcal{L}(\theta)}{\partial\theta}\right\|_2\cdot\|\theta-\theta^*-c\cdot v_{\phi,\perp}\|_2\,, \tag{D.536}$$

$$\left(u(\theta)\coloneqq\min_i\{\pi_i\cdot(1-\pi_i)\}\right) \tag{D.537}$$

where $v_{\phi,\perp}$ is orthogonal to the space Span $\{\phi_1,\phi_2,\ldots,\phi_N\}$, and $\theta-\theta^*-c\cdot v_{\phi,\perp}$ refers to the vector after cutting off all the components $v_{\phi,\perp}$ from $\theta-\theta^*$, such

that $\theta - \theta^* - c \cdot v_{\phi,\perp} \in \text{Span}\{\phi_1, \phi_2, \ldots, \phi_N\}$. Next, we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left( \frac{d\sigma(z_i')}{dz_i'} \right)^2 \cdot \left( \phi_i^\top \theta - \phi_i^\top \theta^* \right)^2 \tag{D.538}$$

$$\text{(by the mean value theorem)} \tag{D.539}$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i')^2 \cdot (1 - \pi_i')^2 \cdot \left( \phi_i^\top \theta - \phi_i^\top \theta^* \right)^2 \tag{D.540}$$

$$\text{(by Eq. (D.519))} \tag{D.541}$$

$$\geq \min_i \left\{ (\pi_i')^2 \cdot (1 - \pi_i')^2 \right\} \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} \left( \phi_i^\top \theta - \phi_i^\top \theta^* \right)^2 \tag{D.542}$$

$$= \min_i \left\{ (\pi_i')^2 \cdot (1 - \pi_i')^2 \right\} \cdot (\theta - \theta^*)^\top \left( \frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top \right) (\theta - \theta^*) \tag{D.543}$$

$$= \min_i \left\{ (\pi_i')^2 \cdot (1 - \pi_i')^2 \right\} \tag{D.544}$$

$$\cdot (\theta - \theta^* - c \cdot v_{\phi,\perp})^\top \left( \frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top \right) (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{D.545}$$

$$\geq \min \left\{ u(\theta)^2, v^2 \right\} \tag{D.546}$$

$$\cdot (\theta - \theta^* - c \cdot v_{\phi,\perp})^\top \left( \frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top \right) (\theta - \theta^* - c \cdot v_{\phi,\perp}) \tag{D.547}$$

$$\left( v \coloneqq \min_i \{ \pi_i^* \cdot (1 - \pi_i^*) \} \right) \tag{D.548}$$

$$\geq \min \left\{ u(\theta)^2, v^2 \right\} \cdot \lambda_\phi \cdot \| \theta - \theta^* - c \cdot v_{\phi,\perp} \|_2^2, \tag{D.549}$$

where $\lambda_\phi$ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^{N} \phi_i \phi_i^\top$. Therefore, we have,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \leq \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \| \theta - \theta^* - c \cdot v_{\phi,\perp} \|_2 \tag{D.550}$$

$$\text{(by Eq. (D.524))} \tag{D.551}$$

$$\leq \frac{1}{8} \cdot \frac{1}{u(\theta)} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \cdot \frac{1}{\min \{ u(\theta), v \}} \cdot \frac{1}{\sqrt{\lambda_\phi}} \tag{D.552}$$

$$\cdot \left[ \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}, \qquad \text{(by Eq. (D.538))} \tag{D.553}$$

which implies,

$$\left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \geq 8 \cdot u(\theta) \cdot \min \{ u(\theta), v \} \cdot \sqrt{\lambda_\phi} \cdot \left[ \frac{1}{N} \cdot \sum_{i=1}^{N} (\pi_i - \pi_i^*)^2 \right]^{\frac{1}{2}}. \qquad \square$$

**Lemma 28.** Denote $u(\theta) \coloneqq \min_i \{\pi_i \cdot (1-\pi_i)\}$, $v \coloneqq \min_i \{\pi_i^* \cdot (1-\pi_i^*)\}$, and $\lambda_\phi$ is the smallest positive eigenvalue of $\frac{1}{N} \cdot \sum_{i=1}^N \phi_i \phi_i^\top$. We have, $\mathcal{L}(\theta)$ satisfies $\beta$ smoothness with

$$\beta = \frac{3}{8} \cdot \max_{i \in [N]} \|\phi_i\|_2^2, \tag{D.554}$$

and $\beta(\theta)$ NS with

$$\beta(\theta) = L_1 \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 + L_0 \cdot \left( \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta) \right), \tag{D.555}$$

where

$$L_1 = \frac{\max_i \|\phi_i\|_2^2}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}}, \text{ and} \tag{D.556}$$

$$L_0 = \frac{17 \cdot \max_i \|\phi_i\|_2^2}{512 \cdot u(\theta)^2 \cdot \min\{u(\theta)^2, v^2\} \cdot \lambda_\phi}. \tag{D.557}$$

*Proof.* Note that the gradient of $\mathcal{L}(\theta)$ is,

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{2}{N} \cdot \sum_{i=1}^N \pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i \in \mathbb{R}^d. \tag{D.558}$$

Denote the second order derivative (Hessian) of $\mathcal{L}(\theta)$ as,

$$S(\theta) \coloneqq \frac{\partial}{\partial \theta} \left\{ \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\} \in \mathbb{R}^{d \times d}. \tag{D.559}$$

For all $j, k \in [d]$, we calculate the corresponding component value of $S(\theta)$ matrix as follows,

$$S_{(j,k)} = \frac{d}{d\theta(k)} \left\{ \frac{2}{N} \cdot \sum_{i=1}^N \pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot \phi_i(j) \right\} \tag{D.560}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \frac{d\{\pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*)\}}{d\theta(k)} \cdot \phi_i(j) \tag{D.561}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \frac{d\{\pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*)\}}{d\{\phi_i^\top \theta\}} \cdot \frac{d\{\phi_i^\top \theta\}}{d\theta(k)} \cdot \phi_i(j) \tag{D.562}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \left[ \pi_i \cdot (1-\pi_i)^2 \cdot (\pi_i - \pi_i^*) \right. \tag{D.563}$$

$$\left. - \pi_i^2 \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) + \pi_i^2 \cdot (1-\pi_i)^2 \right] \cdot \phi_i(k) \cdot \phi_i(j) \tag{D.564}$$

$$= \frac{2}{N} \cdot \sum_{i=1}^N \left[ \pi_i \cdot (1-\pi_i) \cdot (1-2\pi_i) \cdot (\pi_i - \pi_i^*) \right. \tag{D.565}$$

$$\left. + \pi_i^2 \cdot (1-\pi_i)^2 \right] \cdot \phi_i(k) \cdot \phi_i(j). \tag{D.566}$$

To calculate the smoothness coefficient, take a vector $z \in \mathbb{R}^d$. We have,

$$\left|z^\top S(\theta) z\right| = \left|\sum_{j=1}^{d}\sum_{k=1}^{d} S_{(j,k)} \cdot z(j) \cdot z(k)\right| \tag{D.567}$$

$$= \Bigg| \frac{2}{N} \cdot \sum_{i=1}^{N} \big[\pi_i \cdot (1-\pi_i) \cdot (1-2\pi_i) \cdot (\pi_i - \pi_i^*) \tag{D.568}$$

$$+ \pi_i^2 \cdot (1-\pi_i)^2 \big] \cdot \left(\phi_i^\top z\right)^2 \Bigg| \qquad \text{(by Eq. (D.560))} \tag{D.569}$$

$$\leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z\right)^2 \tag{D.570}$$

$$\cdot \sum_{i=1}^{N} \left|\pi_i \cdot (1-\pi_i) \cdot (1-2\pi_i) \cdot (\pi_i - \pi_i^*) + \pi_i^2 \cdot (1-\pi_i)^2\right| \tag{D.571}$$

$$\text{(by Hölder's inequality)} \tag{D.572}$$

$$\leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z\right)^2 \tag{D.573}$$

$$\cdot \sum_{i=1}^{N} \left[\pi_i \cdot (1-\pi_i) \cdot |1-2\pi_i| \cdot |\pi_i - \pi_i^*| + \pi_i^2 \cdot (1-\pi_i)^2\right] \tag{D.574}$$

$$\text{(by triangle inequality)} \tag{D.575}$$

$$\leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z\right)^2 \cdot \sum_{i=1}^{N} \left[\frac{1}{8} + \frac{1}{16}\right] \tag{D.576}$$

$$(x \cdot (1-x) \leq 1/4, \text{ and } x \cdot (1-x) \cdot |1-2x| \leq 1/8, \ \forall x \in [0,1]) \tag{D.577}$$

$$= \frac{3}{8} \cdot \max_i \left[\phi_i^\top \left(\frac{z}{\|z\|_2}\right)\right]^2 \cdot \|z\|_2^2 \tag{D.578}$$

$$\leq \frac{3}{8} \cdot \max_i \|\phi_i\|_2^2 \cdot \|z\|_2^2 . \tag{D.579}$$

Therefore, $\mathcal{L}(\theta)$ satisfies $\beta$ (uniform) smoothness with $\beta = \frac{3}{8} \cdot \max_i \|\phi_i\|_2^2$. Next, we calculate the NS. We have,

$$\sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2 \cdot \mathcal{L}(\theta) = \sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2 \cdot \frac{1}{N} \cdot \sum_{j=1}^{N} (\pi_j - \pi_j^*)^2 \tag{D.580}$$

$$\leq \frac{N}{16} \cdot \frac{1}{N} \cdot \sum_{j=1}^{N} (\pi_j - \pi_j^*)^2 \tag{D.581}$$

$$\leq \frac{N}{16} \cdot \frac{1}{64 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 , \tag{D.582}$$

$$\text{(by Lemma 27)} \tag{D.583}$$

which implies,

$$\sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2 \tag{D.584}$$

$$\leq \frac{N}{2} \cdot \frac{1}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta). \tag{D.585}$$

According to Eq. (D.538), we have

$$\sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)} \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2^{3/2}} \tag{D.586}$$

$$\geq \sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)}} \cdot (\min\{u(\theta)^2, v^2\} \cdot \lambda_\phi)^{3/4} \cdot \frac{1}{\mathcal{L}(\theta)^{3/4}} \tag{D.587}$$

$$= (\min\{u(\theta)^2, v^2\} \cdot \lambda_\phi)^{3/4} \cdot \sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\mathcal{L}(\theta)^{5/4}} \tag{D.588}$$

$$= N \cdot (\min\{u(\theta)^2, v^2\} \cdot \lambda_\phi)^{3/4} \cdot \frac{\mathcal{L}(\theta)}{\mathcal{L}(\theta)^{5/4}} \tag{D.589}$$

$$\geq N \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}. \qquad (\mathcal{L}(\theta) \in (0,1]) \tag{D.590}$$

Therefore we have,

$$\sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |1-2\pi_i| \cdot |\pi_i - \pi_i^*| \tag{D.591}$$

$$\leq \sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*| \tag{D.592}$$

$$\leq \left( \sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*| \right) \tag{D.593}$$

$$\cdot \left( \sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)} \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2^{3/2}} \right) \tag{D.594}$$

$$\cdot \frac{1}{N \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \tag{D.595}$$

$$= \frac{1}{N \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left( \sum_{i=1}^{N} \frac{\pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*|}{\sqrt{\|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2}} \right) \tag{D.596}$$

$$\cdot \left( \sum_{i=1}^{N} \frac{(\pi_i - \pi_i^*)^2}{\sqrt{\mathcal{L}(\theta)} \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2} \right) \tag{D.597}$$

$$\leq \frac{1}{(\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left( \sum_{i=1}^{N} \frac{\pi_i^2 \cdot (1-\pi_i)^2 \cdot (\pi_i - \pi_i^*)^2}{2 \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2} \right. \tag{D.598}$$

$$\left. + \frac{(\pi_i - \pi_i^*)^4}{2 \cdot \mathcal{L}(\theta) \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2^2} \right) \tag{D.599}$$

$$\leq \frac{1}{(\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \tag{D.600}$$

$$\cdot \left( \frac{1}{32} \cdot \sum_{i=1}^{N} \frac{\pi_i \cdot (1-\pi_i) \cdot (\pi_i - \pi_i^*) \cdot (\phi_i^\top \theta - \phi_i^\top \theta^*)}{\|\theta - \theta^* - c \cdot v_{\phi,\perp}\|_2} \right) \tag{D.601}$$

$$+ \frac{1}{(\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \tag{D.602}$$

$$\cdot \left( \frac{1}{32 \cdot u(\theta)^2} \cdot \sum_{i=1}^{N} \frac{\pi_i^2 \cdot (1-\pi_i)^2 \cdot (\pi_i - \pi_i^*)^2 \cdot (\phi_i^\top \theta - \phi_i^\top \theta^*)^2}{\mathcal{L}(\theta) \cdot \|\theta - \theta^* - c \cdot v_{\phi,\perp}\|^2} \right) \tag{D.603}$$

$$\leq \frac{N}{64 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2 \tag{D.604}$$

$$+ \frac{N}{64 \cdot u(\theta)^2 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\| \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \right\|_2^2 \Big/ \mathcal{L}(\theta), \tag{D.605}$$

where the third inequality is according to,

$$\left(\sum_{i=1}^{N} a_i\right) \cdot \left(\sum_{i=1}^{N} b_i\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i \cdot b_j \tag{D.606}$$

$$\leq \frac{1}{2} \cdot \sum_{i=1}^{N}\sum_{j=1}^{N} \left(a_i^2 + b_j^2\right) = \frac{N}{2} \cdot \sum_{i=1}^{N} \left(a_i^2 + b_i^2\right), \tag{D.607}$$

and the last inequality is from the intermediate results in Eq. (D.524). Combining Eqs. (D.567), (D.584) and (D.591), we have

$$\left|z^\top S(\theta) z\right| \leq \frac{2}{N} \cdot \max_i \left(\phi_i^\top z\right)^2 \tag{D.608}$$

$$\cdot \left[\sum_{i=1}^{N} \pi_i \cdot (1-\pi_i) \cdot |\pi_i - \pi_i^*| + \sum_{i=1}^{N} \pi_i^2 \cdot (1-\pi_i)^2\right] \tag{D.609}$$

$$\leq \max_i \left(\phi_i^\top z\right)^2 \tag{D.610}$$

$$\cdot \left(\frac{1}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 \right. \tag{D.611}$$

$$\left. + \frac{17}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta)\right) \tag{D.612}$$

$$\leq \max_i \|\phi_i\|_2^2 \cdot \|z\|_2^2 \tag{D.613}$$

$$\cdot \left(\frac{1}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 \right. \tag{D.614}$$

$$\left. + \frac{17}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi} \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta)\right). \tag{D.615}$$

Therefore, $\mathcal{L}(\theta)$ satisfies $\beta(\theta)$ NS with

$$\beta(\theta) = L_1 \cdot \left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2 + L_0 \cdot \left(\left\|\frac{\partial \mathcal{L}(\theta)}{\partial \theta}\right\|_2^2 \Big/ \mathcal{L}(\theta)\right), \tag{D.616}$$

where

$$L_1 = \frac{\max_i \|\phi_i\|_2^2}{32 \cdot (\min\{u(\theta), v\} \cdot \sqrt{\lambda_\phi})^{3/2}}, \text{ and} \tag{D.617}$$

$$L_0 = \frac{17 \cdot \max_i \|\phi_i\|_2^2}{512 \cdot u(\theta)^2 \cdot \min\left\{u(\theta)^2, v^2\right\} \cdot \lambda_\phi}. \qquad \square$$

**Theorem 20.** With $\eta = 1/\beta$, GD update satisfies for all $t \geq 1$,

$$\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_1) \cdot e^{-C^2 \cdot (t-1)}. \tag{D.618}$$

With $\eta \in \Theta(1)$, GNGD update satisfies for all $t \geq 1$,

$$\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_1) \cdot e^{-C \cdot (t-1)}, \tag{D.619}$$

where $C \in (0, 1)$, i.e., GNGD is strictly faster than GD.

*Proof.* Combining Lemmas 27 and 28, and the second part of (2b) in Theorem 16, we have the results for GD. Using the fourth part of (2b) in Theorem 16, we have the results for GNGD. □

## D.4 Miscellaneous Extra Supporting Results

**Lemma 52** (Descent lemma for NS function)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function that satisfies NS with $\beta(\theta) > 0$, for all $\theta \in \mathbb{R}^d$ and $\theta' = \theta - \frac{1}{\beta(\theta)} \cdot \frac{\partial f(\theta)}{\partial \theta}$. We have,*

$$f(\theta') \leq f(\theta) - \frac{1}{2 \cdot \beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \tag{D.620}$$

*Proof.* According to Definition 6, we have,

$$f(\theta') - f(\theta) \leq \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle + \frac{\beta(\theta)}{2} \cdot \|\theta' - \theta\|_2^2 \tag{D.621}$$

$$= -\frac{1}{\beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 + \frac{1}{2 \cdot \beta(\theta)} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 \tag{D.622}$$

$$\left( \theta' = \theta - \frac{1}{\beta(\theta)} \cdot \frac{\partial f(\theta)}{\partial \theta} \right) \tag{D.623}$$

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

□

**Lemma 53.** *Given any $\alpha > 0$, we have, for all $x \in [0, 1]$,*

$$\frac{1}{\alpha} \cdot (1 - x^\alpha) \geq x^\alpha \cdot (1 - x). \tag{D.624}$$

*Proof.* Define $f : x \mapsto \frac{1}{\alpha} \cdot (1 - x^\alpha) - x^\alpha \cdot (1 - x)$. We show that $f(x) \geq 0$ for all $x \in [0, 1]$. Note that,

$$f(0) = \frac{1}{\alpha} > 0, \text{ and } f(1) = 0. \tag{D.625}$$

On the other hand,

$$f'(x) = -x^{\alpha-1} - \alpha \cdot x^{\alpha-1} \cdot (1-x) + x^{\alpha} \tag{D.626}$$

$$= -x^{\alpha-1} \cdot [1 + \alpha \cdot (1-x) - x] \tag{D.627}$$

$$= -x^{\alpha-1} \cdot (1+\alpha) \cdot (1-x) \tag{D.628}$$

$$\leq 0, \qquad (\alpha > 0, \text{ and } x \in [0,1]) \tag{D.629}$$

which means $f$ is monotonically decreasing over $[0,1]$. Therefore $f(x) \geq 0$ for all $x \in [0,1]$, finishing the proof. □

**Lemma 54.** *Given any $\alpha > 0$, we have, for all $x \in \left[\frac{2\alpha+1}{2\alpha+2}, 1\right]$,*

$$\frac{1}{2\alpha} \cdot (1 - x^{\alpha}) \leq x^{\alpha} \cdot (1-x)\,. \tag{D.630}$$

*Proof.* Define $g : x \mapsto x^{\alpha} \cdot (1-x) - \frac{1}{2\alpha} \cdot (1 - x^{\alpha})$. The derivative of $g$ is,

$$g'(x) = \alpha \cdot x^{\alpha-1} \cdot (1-x) - x^{\alpha} + (1/2) \cdot x^{\alpha-1} \tag{D.631}$$

$$= x^{\alpha-1} \cdot [\alpha \cdot (1-x) - x + 1/2] \tag{D.632}$$

$$= x^{\alpha-1} \cdot [(1+\alpha) \cdot (1-x) - 1/2]\,. \tag{D.633}$$

Then we have,

$$g'(x) > 0 \text{ for all } x \in [0, (2\alpha+1)/(2\alpha+2))\,, \text{ and} \tag{D.634}$$

$$g'(x) \leq 0 \text{ for all } x \in [(2\alpha+1)/(2\alpha+2), 1]\,, \tag{D.635}$$

which means $g$ is monotonically increasing over $[0, (2\alpha+1)/(2\alpha+2))$ and de-

creasing over $[(2\alpha+1)/(2\alpha+2), 1]$. On the other hand,

$$
\begin{aligned}
& g((2\alpha+1)/(2\alpha+2)) && \text{(D.636)}\\
&= \left(\frac{2\alpha+1}{2\alpha+2}\right)^{\alpha} \cdot \left(1 - \frac{2\alpha+1}{2\alpha+2}\right) - \frac{1}{2\alpha} \cdot \left[1 - \left(\frac{2\alpha+1}{2\alpha+2}\right)^{\alpha}\right] && \text{(D.637)}\\
&= \frac{1}{2\alpha} \cdot \left[\left(\frac{2\alpha+1}{2\alpha+2}\right)^{\alpha} \cdot \frac{2\alpha+1}{\alpha+1} - 1\right] && \text{(D.638)}\\
&= \frac{1}{2\alpha} \cdot \left[\exp\left\{\log\left(\frac{2\alpha+1}{\alpha+1}\right) - \alpha \cdot \log\left(1 + \frac{1}{2\alpha+1}\right)\right\} - 1\right] && \text{(D.639)}\\
&\geq \frac{1}{2\alpha} \cdot \left[\exp\left\{\log\left(\frac{2\alpha+1}{\alpha+1}\right) - \frac{\alpha}{2\alpha+1}\right\} - 1\right] \qquad (1 + x \leq e^x) && \text{(D.640)}\\
&\geq \frac{1}{2\alpha} \cdot \left[\exp\left\{\frac{\alpha}{2\alpha+1} - \frac{\alpha}{2\alpha+1}\right\} - 1\right] && \text{(D.641)}\\
&\qquad (\log(x) \geq 1 - 1/x \text{ for } x > 0) && \text{(D.642)}\\
&= 0. && \text{(D.643)}
\end{aligned}
$$

Also note that $g(1) = 0$. Therefore we have, for all $x \in [(2\alpha+1)/(2\alpha+2), 1]$, $g(x) \geq 0$, finishing the proof. □

**Lemma 55.** *Softmax policy gradient norm is*

$$\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 = \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\|_2^2\right]^{\frac{1}{2}}. \tag{D.644}$$

*Proof.* We have,

$$
\begin{aligned}
\left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 &= \left[\sum_{s,a}\left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)}\right)^2\right]^{\frac{1}{2}} && \text{(D.645)}\\
&= \left[\sum_s \left\|\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)}\right\|_2^2\right]^{\frac{1}{2}} && \text{(D.646)}\\
&= \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \|H(\pi_\theta(\cdot|s))Q^{\pi_\theta}(s,\cdot)\|_2^2\right]^{\frac{1}{2}}. && \text{(D.647)}\\
&\qquad \text{(by Lemma 1)} && \square
\end{aligned}
$$

# Appendix E

# Proofs for Chapter 5: Understanding Stochasticity in Policy Optimization

## E.1 Proofs for Section 5.2: Algorithm Preferability

**Lemma 29** (Natural NŁ inequality, continuous)**.** Let $r \in (0,1)^K$. Denote $\Delta(a) \coloneqq r(a^*) - r(a)$, and $\Delta \coloneqq r(a^*) - \max_{a \neq a^*} r(a)$ as the reward gap of $r$. We have, for any policy $\pi_\theta \coloneqq \text{softmax}(\theta)$,

$$\left\langle \frac{d\pi_\theta^\top r}{d\theta}, r \right\rangle \geq \pi_\theta(a^*) \cdot \Delta \cdot (\pi^* - \pi_\theta)^\top r. \tag{E.1}$$

*Proof.* Without loss of generality, let $r(1) > r(2) > \cdots > r(K)$. We have,

$$\left\langle \frac{d\pi_\theta^\top r}{d\theta}, r \right\rangle = r^\top \left( \mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) r \tag{E.2}$$

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

$$= \sum_{i=1}^{K} \pi_\theta(i) \cdot r(i)^2 - \sum_{i=1}^{K} \pi_\theta(i)^2 \cdot r(i)^2 \tag{E.4}$$

$$- 2 \cdot \sum_{i=1}^{K-1} \pi_\theta(i) \cdot r(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot r(j) \tag{E.5}$$

$$= \sum_{i=1}^{K} \pi_\theta(i) \cdot r(i)^2 \cdot [1 - \pi_\theta(i)] \tag{E.6}$$

$$- 2 \cdot \sum_{i=1}^{K-1} \pi_\theta(i) \cdot r(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot r(j) \tag{E.7}$$

$$= \sum_{i=1}^{K} \pi_\theta(i) \cdot r(i)^2 \cdot \sum_{j \neq i} \pi_\theta(j) \tag{E.8}$$

$$- 2 \cdot \sum_{i=1}^{K-1} \pi_\theta(i) \cdot r(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot r(j) \tag{E.9}$$

$$= \sum_{i=1}^{K-1} \pi_\theta(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot \left[ r(i)^2 + r(j)^2 \right] \tag{E.10}$$

$$- 2 \cdot \sum_{i=1}^{K-1} \pi_\theta(i) \cdot r(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot r(j) \tag{E.11}$$

$$= \sum_{i=1}^{K-1} \pi_\theta(i) \cdot \sum_{j=i+1}^{K} \pi_\theta(j) \cdot [r(i) - r(j)]^2, \tag{E.12}$$

which can be lower bounded as,

$$\left\langle \frac{d\pi_\theta^\top r}{d\theta}, r \right\rangle \geq \pi_\theta(1) \cdot \sum_{j=2}^{K} \pi_\theta(j) \cdot [r(1) - r(j)]^2 \qquad \text{(fewer terms)} \tag{E.13}$$

$$= \pi_\theta(a^*) \cdot \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta(a)^2 \qquad (a^* = 1) \tag{E.14}$$

$$\geq \pi_\theta(a^*) \cdot \Delta \cdot \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta(a) \qquad (\Delta(a) \geq \Delta) \tag{E.15}$$

$$= \pi_\theta(a^*) \cdot \Delta \cdot (\pi^* - \pi_\theta)^\top r. \qquad \square$$

**Remark 22.** *The natural NL inequality of Lemma 29 is tight. Consider* $K =$

2, *we have,*

$$
\begin{aligned}
& r^\top \left(diag(\pi_\theta) - \pi_\theta \pi_\theta^\top\right) r && \text{(E.16)}\\
&= \pi_\theta(1) \cdot r(1)^2 + \pi_\theta(2) \cdot r(2)^2 - \left[\pi_\theta(1) \cdot r(1) + \pi_\theta(2) \cdot r(2)\right]^2 && \text{(E.17)}\\
&= \pi_\theta(1) \cdot r(1)^2 \cdot [1 - \pi_\theta(1)] + \pi_\theta(2) \cdot r(2)^2 \cdot [1 - \pi_\theta(2)] && \text{(E.18)}\\
&\quad - 2 \cdot \pi_\theta(1) \cdot r(1) \cdot \pi_\theta(2) \cdot r(2) && \text{(E.19)}\\
&= \pi_\theta(1) \cdot r(1)^2 \cdot \pi_\theta(2) + \pi_\theta(2) \cdot r(2)^2 \cdot \pi_\theta(1) && \text{(E.20)}\\
&\quad - 2 \cdot \pi_\theta(1) \cdot r(1) \cdot \pi_\theta(2) \cdot r(2) \qquad (\pi_\theta(1) + \pi_\theta(2) = 1) && \text{(E.21)}\\
&= \pi_\theta(1) \cdot \pi_\theta(2) \cdot [r(1) - r(2)]^2 && \text{(E.22)}\\
&= \pi_\theta(a^*) \cdot \Delta \cdot (\pi^* - \pi_\theta)^\top r, && \text{(E.23)}\\
&\quad \left(a^* = 1,\ \Delta = r(1) - r(2),\ (\pi^* - \pi_\theta)^\top r = \pi_\theta(2) \cdot [r(1) - r(2)]\right) && \text{(E.24)}
\end{aligned}
$$

*which means the equality holds for the above problem.*

**Remark 23.** *For the continuous natural PG flow:* $\frac{d\theta_t}{dt} = \eta \cdot r$, *and* $\pi_{\theta_t} = \text{softmax}(\theta_t)$, *Lemma 29 can be used to characterize the progress at each time step. We have, for all* $t \geq 1$,

$$
\begin{aligned}
\frac{d\left(\pi^* - \pi_{\theta_t}\right)^\top r}{dt} &= -\frac{d\pi_{\theta_t}^\top r}{dt} && \text{(E.25)}\\
&= -\left(\frac{d\theta_t}{dt}\right)^\top \left(\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) && \text{(E.26)}\\
&= -\eta \cdot r^\top \left(diag(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right) r \qquad (\textit{NPG flow}) && \text{(E.27)}\\
&\leq -\eta \cdot \pi_{\theta_t}(a^*) \cdot \Delta \cdot \left(\pi^* - \pi_{\theta_t}\right)^\top r, \qquad (\textit{by Lemma 29}) && \text{(E.28)}
\end{aligned}
$$

*which means the progress at time* $t$ *is proportional to the sub-optimality gap* $\left(\pi^* - \pi_{\theta_t}\right)^\top r$, *leading to a linear convergence rate.*

**Lemma 30** (Natural NŁ, discrete)**.** Given any policy $\pi$, define $\pi'$ as

$$
\pi'(a) \coloneqq \frac{\pi(a) \cdot e^{\eta \cdot r(a)}}{\sum_{a'} \pi(a') \cdot e^{\eta \cdot r(a')}}, \text{ for all } a \in [K], \tag{E.29}
$$

where $\eta > 0$ is the learning rate. We have,

$$
\left(\pi' - \pi\right)^\top r \geq \left[1 - \frac{1}{\pi(a^*) \cdot (e^{\eta \cdot \Delta} - 1) + 1}\right] \cdot \left(\pi^* - \pi\right)^\top r. \tag{E.30}
$$

*Proof.* Without loss of generality, let $r(1) > r(2) > \cdots > r(K)$. We have,

$$(\pi' - \pi)^\top r = \sum_{i=1}^{K} [\pi'(i) \cdot r(i) - \pi(i) \cdot r(i)] \tag{E.31}$$

$$= \sum_{i=1}^{K} \left[ \frac{\pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i)}{\sum_{j=1}^{K} \pi(j) \cdot e^{\eta \cdot r(j)}} - \pi(i) \cdot r(i) \right] \qquad \text{(by definition of } \pi') \tag{E.32}$$

$$= \frac{1}{\sum_{j=1}^{K} \pi(j) \cdot e^{\eta \cdot r(j)}} \cdot \Bigg[ \sum_{i=1}^{K} \pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i) \tag{E.33}$$

$$- \sum_{i=1}^{K} \pi(i) \cdot r(i) \cdot \sum_{j=1}^{K} \pi(j) \cdot e^{\eta \cdot r(j)} \Bigg], \tag{E.34}$$

Next, we have,

$$\sum_{i=1}^{K} \pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i) - \sum_{i=1}^{K} \pi(i) \cdot r(i) \cdot \sum_{j=1}^{K} \pi(j) \cdot e^{\eta \cdot r(j)} \tag{E.35}$$

$$= \sum_{i=1}^{K} \pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i) - \sum_{i=1}^{K} \pi(i)^2 \cdot e^{\eta \cdot r(i)} \cdot r(i) \tag{E.36}$$

$$- \sum_{i=1}^{K} \pi(i) \cdot r(i) \cdot \sum_{j \neq i} \pi(j) \cdot e^{\eta \cdot r(j)} \tag{E.37}$$

$$= \sum_{i=1}^{K} \pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i) \cdot [1 - \pi(i)] \tag{E.38}$$

$$- \sum_{i=1}^{K} \pi(i) \cdot r(i) \cdot \sum_{j \neq i} \pi(j) \cdot e^{\eta \cdot r(j)} \tag{E.39}$$

$$= \sum_{i=1}^{K} \pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i) \cdot \sum_{j \neq i} \pi(j) \tag{E.40}$$

$$- \sum_{i=1}^{K} \pi(i) \cdot r(i) \cdot \sum_{j \neq i} \pi(j) \cdot e^{\eta \cdot r(j)} \tag{E.41}$$

$$= \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot \left[ e^{\eta \cdot r(i)} \cdot r(i) + e^{\eta \cdot r(j)} \cdot r(j) \right] \tag{E.42}$$

$$- \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot \left[ e^{\eta \cdot r(j)} \cdot r(i) + e^{\eta \cdot r(i)} \cdot r(j) \right] \tag{E.43}$$

$$= \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot \left[ e^{\eta \cdot r(i)} - e^{\eta \cdot r(j)} \right] \cdot [r(i) - r(j)], \tag{E.44}$$

which can be lower bounded as,

$$\sum_{i=1}^{K} \pi(i) \cdot e^{\eta \cdot r(i)} \cdot r(i) - \sum_{i=1}^{K} \pi(i) \cdot r(i) \cdot \sum_{j=1}^{K} \pi(j) \cdot e^{\eta \cdot r(j)} \tag{E.45}$$

$$\geq \pi(1) \cdot \sum_{j=2}^{K} \pi(j) \cdot \left[e^{\eta \cdot r(1)} - e^{\eta \cdot r(j)}\right] \cdot [r(1) - r(j)] \tag{E.46}$$

$$\text{(fewer terms)} \tag{E.47}$$

$$\geq \pi(1) \cdot \sum_{j=2}^{K} \pi(j) \cdot \left[e^{\eta \cdot r(1)} - e^{\eta \cdot r(2)}\right] \cdot [r(1) - r(j)] \tag{E.48}$$

$$(r(j) \leq r(2), \text{ for all } j \geq 2) \tag{E.49}$$

$$= \pi(1) \cdot e^{\eta \cdot r(2)} \cdot \left(e^{\eta \cdot \Delta} - 1\right) \cdot \sum_{a \neq a^*} \pi(a) \cdot \Delta(a) \tag{E.50}$$

$$(\Delta = r(1) - r(2)) \tag{E.51}$$

$$= \pi(1) \cdot e^{\eta \cdot r(2)} \cdot \left(e^{\eta \cdot \Delta} - 1\right) \cdot (\pi^* - \pi)^\top r. \tag{E.52}$$

Combining Eqs. (E.31) and (E.35), we have,

$$(\pi' - \pi)^\top r \geq \frac{\pi(1) \cdot e^{\eta \cdot r(2)} \cdot \left(e^{\eta \cdot \Delta} - 1\right)}{\pi(1) \cdot e^{\eta \cdot r(1)} + \sum_{j=2}^{K} \pi(j) \cdot e^{\eta \cdot r(j)}} \cdot (\pi^* - \pi)^\top r \tag{E.53}$$

$$= \frac{\pi(1) \cdot \left(e^{\eta \cdot \Delta} - 1\right)}{\pi(1) \cdot e^{\eta \cdot \Delta} + \sum_{j=2}^{K} \pi(j) \cdot e^{\eta \cdot [r(j) - r(2)]}} \cdot (\pi^* - \pi)^\top r \tag{E.54}$$

$$\geq \frac{\pi(1) \cdot \left(e^{\eta \cdot \Delta} - 1\right)}{\pi(1) \cdot e^{\eta \cdot \Delta} + \sum_{j=2}^{K} \pi(j)} \cdot (\pi^* - \pi)^\top r \tag{E.55}$$

$$(r(j) - r(2) \leq 0, \text{ for all } j \geq 2) \tag{E.56}$$

$$= \frac{\pi(1) \cdot \left(e^{\eta \cdot \Delta} - 1\right)}{\pi(1) \cdot e^{\eta \cdot \Delta} + 1 - \pi(1)} \cdot (\pi^* - \pi)^\top r \tag{E.57}$$

$$= \left[1 - \frac{1}{\pi(a^*) \cdot (e^{\eta \cdot \Delta} - 1) + 1}\right] \cdot (\pi^* - \pi)^\top r. \qquad (a^* = 1) \qquad \square$$

**Remark 24.** *The natural NL inequality of Lemma 30 is tight. Consider* $K =$

2, *we have,*

$$
\begin{aligned}
&(\pi' - \pi)^\top r && \text{(E.58)}\\
&= \frac{\pi(1)\cdot e^{\eta\cdot r(1)}\cdot r(1) + \pi(2)\cdot e^{\eta\cdot r(2)}\cdot r(2)}{\pi(1)\cdot e^{\eta\cdot r(1)} + \pi(2)\cdot e^{\eta\cdot r(2)}} && \text{(E.59)}\\
&\quad - [\pi(1)\cdot r(1) + \pi(2)\cdot r(2)] && \text{(E.60)}\\
&= \frac{\pi(1)\cdot\pi(2)\cdot[r(1)-r(2)]\cdot\left[e^{\eta\cdot r(1)} - e^{\eta\cdot r(2)}\right]}{\pi(1)\cdot e^{\eta\cdot r(1)} + \pi(2)\cdot e^{\eta\cdot r(2)}} && \text{(E.61)}\\
&= \frac{\pi(1)\cdot\left(e^{\eta\cdot[r(1)-r(2)]} - 1\right)}{\pi(1)\cdot e^{\eta\cdot[r(1)-r(2)]} + \pi(2)}\cdot\pi(2)\cdot[r(1)-r(2)] && \text{(E.62)}\\
&= \frac{\pi(a^*)\cdot\left(e^{\eta\cdot\Delta} - 1\right)}{\pi(a^*)\cdot e^{\eta\cdot\Delta} + 1 - \pi(a^*)}\cdot(\pi^* - \pi)^\top r, && \text{(E.63)}\\
&\quad (a^* = 1,\ \Delta = r(1) - r(2)) && \text{(E.64)}
\end{aligned}
$$

*which means the equality holds for the above problem.*

**Theorem 21.** Using Update 4 with any $\eta > 0$, i.e., $\forall t \geq 1$,

$$\theta_{t+1} \leftarrow \theta_t + \eta\cdot r,\ \text{and}\ \pi_{\theta_{t+1}} = \text{softmax}(\theta_{t+1}), \tag{E.65}$$

where $\eta > 0$ is the learning rate. We have, for all $t \geq 1$,

$$(\pi^* - \pi_{\theta_t})^\top r \leq (\pi^* - \pi_{\theta_1})^\top r\cdot e^{-c\cdot(t-1)}, \tag{E.66}$$

where $c \coloneqq \log\left(\pi_{\theta_1}(a^*)\cdot\left(e^{\eta\cdot\Delta} - 1\right) + 1\right) > 0$ for any $\eta > 0$, and $\Delta = r(a^*) - \max_{a\neq a^*} r(a) > 0$.

*Proof.* We have, for all $t \geq 1$,

$$
\begin{aligned}
&\left(\pi^* - \pi_{\theta_{t+1}}\right)^\top r = (\pi^* - \pi_{\theta_t})^\top r - \left(\pi_{\theta_{t+1}} - \pi_{\theta_t}\right)^\top r && \text{(E.67)}\\
&\leq \frac{1}{\pi_{\theta_t}(a^*)\cdot(e^{\eta\cdot\Delta} - 1) + 1}\cdot(\pi^* - \pi_{\theta_t})^\top r \qquad \text{(by Lemma 30)} && \text{(E.68)}\\
&\leq \frac{1}{\pi_{\theta_1}(a^*)\cdot(e^{\eta\cdot\Delta} - 1) + 1}\cdot(\pi^* - \pi_{\theta_t})^\top r \qquad \text{(see below)} && \text{(E.69)}\\
&\leq \frac{1}{[\pi_{\theta_1}(a^*)\cdot(e^{\eta\cdot\Delta} - 1) + 1]^t}\cdot(\pi^* - \pi_{\theta_1})^\top r && \text{(E.70)}\\
&= \frac{(\pi^* - \pi_{\theta_1})^\top r}{e^{c\cdot t}}, && \text{(E.71)}
\end{aligned}
$$

where the second inequality is because of for all $t \geq 1$,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\pi_{\theta_t}(a^*) \cdot e^{\eta \cdot r(a^*)}}{\sum_a \pi_{\theta_t}(a) \cdot e^{\eta \cdot r(a)}} \tag{E.72}$$

$$= \frac{\pi_{\theta_t}(a^*)}{\sum_a \pi_{\theta_t}(a) \cdot e^{-\eta \cdot \Delta(a)}} \tag{E.73}$$

$$\geq \pi_{\theta_t}(a^*). \qquad (\Delta(a) \geq 0) \qquad \square$$

**Lemma 31.** Let $\hat{r}$ be the IS estimator using on-policy sampling $a \sim \pi_\theta(\cdot)$. The stochastic softmax PG estimator is unbiased and bounded, i.e.,

$$\mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot)} \left[ \frac{d\pi_\theta^\top \hat{r}}{d\theta} \right] = \frac{d\pi_\theta^\top r}{d\theta}, \text{ and} \tag{E.74}$$

$$\mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot)} \left\| \frac{d\pi_\theta^\top \hat{r}}{d\theta} \right\|_2^2 \leq \frac{K}{2}. \tag{E.75}$$

*Proof.* **First part.** $\mathbb{E}_{a \sim \pi_\theta(\cdot)} \left[ \frac{d\pi_\theta^\top \hat{r}}{d\theta} \right] = \frac{d\pi_\theta^\top r}{d\theta}$.

We have, for all $i \in [K]$, the true softmax PG is,

$$\frac{d\pi_\theta^\top r}{d\theta(i)} = \pi_\theta(i) \cdot \left( r(i) - \pi_\theta^\top \right). \tag{E.76}$$

On the other hand, we have, for all $i \in [K]$,

$$\frac{d\pi_\theta^\top \hat{r}}{d\theta(i)} = \pi_\theta(i) \cdot \left( \hat{r}(i) - \pi_\theta^\top \hat{r} \right) \tag{E.77}$$

$$= \pi_\theta(i) \cdot \left( \frac{\mathbb{I}\{a = i\}}{\pi_\theta(i)} \cdot r(i) - \sum_j \mathbb{I}\{a = j\} \cdot r(j) \right) \tag{E.78}$$

$$\text{(by Definition 10)} \tag{E.79}$$

$$= \mathbb{I}\{a = i\} \cdot r(i) - \pi_\theta(i) \cdot r(a). \tag{E.80}$$

The expectation of stochastic softmax PG is,

$$\mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot)} \left[ \frac{d\pi_\theta^\top \hat{r}}{d\theta(i)} \right] = \sum_{a \in [K]} \pi_\theta(a) \cdot \left( \mathbb{I}\{a = i\} \cdot r(i) - \pi_\theta(i) \cdot r(a) \right) \tag{E.81}$$

$$= \pi_\theta(i) \cdot r(i) - \pi_\theta(i) \cdot \pi_\theta^\top r \tag{E.82}$$

$$= \frac{d\pi_\theta^\top r}{d\theta(i)}. \tag{E.83}$$

**Second part.** $\mathbb{E}_{a \sim \pi_\theta(\cdot)} \left\| \frac{d\pi_\theta^\top \hat{r}}{d\theta} \right\|_2^2 \leq \frac{K}{2}$.

The squared stochastic PG norm is,

$$\left\|\frac{d\pi_\theta^\top \hat{r}}{d\theta}\right\|_2^2 = \sum_{i=1}^K \left(\frac{d\pi_\theta^\top \hat{r}}{d\theta(i)}\right)^2 = \sum_{i=1}^K \pi_\theta(i)^2 \cdot \left(\hat{r}(i) - \pi_\theta^\top \hat{r}\right)^2 \tag{E.84}$$

$$= \sum_{i=1}^K \pi_\theta(i)^2 \cdot \Bigg[\frac{\left(\mathbb{I}\left\{a=i\right\}\right)^2}{\pi_\theta(i)^2} \cdot r(i)^2 \tag{E.85}$$

$$- 2 \cdot \frac{\mathbb{I}\left\{a=i\right\}}{\pi_\theta(i)} \cdot r(i) \cdot \sum_{j=1}^K \mathbb{I}\left\{a=j\right\} \cdot r(j) \tag{E.86}$$

$$+ \left(\sum_{j=1}^K \mathbb{I}\left\{a=j\right\} \cdot r(j)\right)^2\Bigg] \tag{E.87}$$

$$= \sum_{i=1}^K \Bigg[\mathbb{I}\left\{a=i\right\} \cdot r(i)^2 \tag{E.88}$$

$$- 2 \cdot \left(\sum_{i=1}^K \pi_\theta(i) \cdot \mathbb{I}\left\{a=i\right\} \cdot r(i)\right) \cdot \left(\sum_{j=1}^K \mathbb{I}\left\{a=j\right\} \cdot r(j)\right) \tag{E.89}$$

$$+ \left(\sum_{i=1}^K \mathbb{I}\left\{a=i\right\} \cdot r(i)\right) \cdot \left(\sum_{j=1}^K \mathbb{I}\left\{a=j\right\} \cdot r(j)\right)\Bigg]. \tag{E.90}$$

The expected squared stochastic PG norm is,

$$\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)} \left\|\frac{d\pi_\theta^\top \hat{r}}{d\theta}\right\|_2^2 = \sum_{a\in[K]} \pi_\theta(a) \cdot r(a)^2 \tag{E.91}$$

$$- 2 \cdot \sum_{a\in[K]} \pi_\theta(a)^2 \cdot r(a)^2 + \sum_{a\in[K]} \pi_\theta(a) \cdot r(a)^2 \tag{E.92}$$

$$= 2 \cdot r^\top \left(\mathrm{diag}(\pi_\theta) - \mathrm{diag}(\pi_\theta \odot \pi_\theta)\right) r \tag{E.93}$$

$$= 2 \cdot \sum_{a\in[K]} \pi_\theta(a) \cdot (1 - \pi_\theta(a)) \cdot r(a)^2 \tag{E.94}$$

$$\leq \frac{2}{4} \cdot \sum_{i=1}^K r(i)^2 \qquad (x \cdot (1-x) \leq 1/4, \text{ for all } x \in [0,1]) \tag{E.95}$$

$$\leq \frac{K}{2}. \qquad \left(r \in (0,1]^K\right) \qquad \square$$

**Lemma 56** (Non-uniform Smoothness (NS) between two iterations)**.** *Let* $\theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta}$*. We have, for* $\eta = \frac{1}{3\cdot K} \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2$*,*

$$\left|(\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle\right| \leq 3 \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \cdot \|\theta' - \theta\|_2^2. \tag{E.96}$$

*Proof.* Denote $\theta_\zeta \coloneqq \theta + \zeta \cdot (\theta' - \theta)$ with some $\zeta \in [0,1]$. According to Taylor's theorem, $\forall \theta,\ \theta'$,

$$\left|(\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle\right| = \frac{1}{2} \cdot \left|(\theta' - \theta)^\top \frac{d^2 \pi_{\theta_\zeta}^\top r}{d{\theta_\zeta}^2} (\theta' - \theta)\right| \tag{E.97}$$

$$= \frac{3}{2} \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2 \cdot \|\theta' - \theta\|_2^2. \qquad \text{(by Lemma 21)} \tag{E.98}$$

Denote $\zeta_1 \coloneqq \zeta$. Also denote $\theta_{\zeta_2} \coloneqq \theta + \zeta_2 \cdot (\theta_{\zeta_1} - \theta)$ with some $\zeta_2 \in [0,1]$. We have,

$$\left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left\| \int_0^1 \left\langle \frac{d^2\{\pi_{\theta_{\zeta_2}}^\top r\}}{d\theta_{\zeta_2}^2}, \theta_{\zeta_1} - \theta \right\rangle d\zeta_2 \right\|_2 \tag{E.99}$$

$$\leq \int_0^1 \left\| \frac{d^2\{\pi_{\theta_{\zeta_2}}^\top r\}}{d\theta_{\zeta_2}^2} \right\|_2 \cdot \|\theta_{\zeta_1} - \theta\|_2 \, d\zeta_2 \tag{E.100}$$

$$\leq \int_0^1 3 \cdot \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 \cdot \zeta_1 \cdot \|\theta' - \theta\|_2 \, d\zeta_2 \qquad \text{(by Lemma 21)} \tag{E.101}$$

$$\leq \int_0^1 3 \cdot \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 \cdot \eta \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 d\zeta_2, \tag{E.102}$$

$$\left(\zeta_1 \in [0,1],\ \text{using } \theta' = \theta + \eta \cdot \frac{d\pi_\theta^\top r}{d\theta}\right) \tag{E.103}$$

where the second inequality is because of the Hessian is symmetric, and its operator norm is equal to its spectral radius. Therefore we have,

$$\left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \leq \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 + \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \tag{E.104}$$

$$\text{(by triangle inequality)} \tag{E.105}$$

$$\leq \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 + 3\eta \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \cdot \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 d\zeta_2. \tag{E.106}$$

$$\text{(by Eq. (E.99))} \tag{E.107}$$

Denote $\theta_{\zeta_3} \coloneqq \theta + \zeta_3 \cdot (\theta_{\zeta_2} - \theta)$ with $\zeta_3 \in [0,1]$. Using similar calculation in Eq. (E.99), we have,

$$\left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 \leq \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 + \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} - \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \tag{E.108}$$

$$\leq \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 + 3\eta \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \cdot \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_3}}^\top r}{d\theta_{\zeta_3}} \right\|_2 d\zeta_3. \tag{E.109}$$

Combining Eqs. (E.104) and (E.108), we have,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}}\right\|_2 \leq \left(1+3\eta\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\right)\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{E.110}$$

$$+\left(3\eta\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\right)^2\cdot\int_0^1\int_0^1\left\|\frac{d\pi_{\theta_{\zeta_3}}^\top r}{d\theta_{\zeta_3}}\right\|_2 d\zeta_3 d\zeta_2, \tag{E.111}$$

which implies,

$$\left\|\frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}}\right\|_2 \leq \sum_{i=0}^{\infty}\left(3\eta\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\right)^i\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{E.112}$$

$$=\frac{1}{1-3\eta\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2}\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{E.113}$$

$$\left(3\eta\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\in(0,1),\ \text{see below}\right) \tag{E.114}$$

$$=\frac{1}{1-\frac{1}{K}\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2^2}\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{E.115}$$

$$\left(\eta=\frac{1}{3\cdot K}\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\right) \tag{E.116}$$

$$\leq\frac{1}{1-\frac{1}{2}\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2^2}\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \qquad (K\geq 2) \tag{E.117}$$

$$\leq 2\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2, \tag{E.118}$$

where the first equation and the last inequality are from,

$$\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2^2=\sum_{a\in[K]}\pi_\theta(a)^2\cdot(r(a)-\pi_\theta^\top r)^2 \tag{E.119}$$

$$\leq\sum_{a\in[K]}\pi_\theta(a)^2 \qquad \left(r\in(0,1]^K\right) \tag{E.120}$$

$$\leq 1. \qquad (\|x\|_2\leq\|x\|_1) \tag{E.121}$$

Combining Eqs. (E.97) and (E.112) finishes the proof. □

**Theorem 22.** Using Update 6, $(\pi^*-\pi_{\theta_t})^\top r\to 0$ as $t\to\infty$ with probability 1.

*Proof.* See (Chung et al., 2020, Proposition 4). We include a proof for completeness.

Denote $\delta(\theta_t) \coloneqq (\pi^* - \pi_{\theta_t})^\top r$. Let $\eta = \frac{1}{3 \cdot K} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$ for all $t \geq 1$. We have, for all $t \geq 1$,

$$\delta(\theta_{t+1}) - \delta(\theta_t) \tag{E.122}$$

$$= -\pi_{\theta_{t+1}}^\top r + \pi_{\theta_t}^\top r + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{E.123}$$

$$\leq 3 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{E.124}$$

(by Lemma 56) (E.125)

$$= 3 \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2 - \eta \cdot \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\rangle. \tag{E.126}$$

(using Update 6) (E.127)

Next, taking expectation over the random sampling on Eq. (E.122), we have,

$$\mathbb{E}\left[\delta(\theta_{t+1})\right] - \mathbb{E}\left[\delta(\theta_t)\right] \tag{E.128}$$

$$= 3 \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \mathbb{E}\left[ \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2 \right] - \eta \cdot \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \mathbb{E}\left[ \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right] \right\rangle \tag{E.129}$$

$$= 3 \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \mathbb{E}\left[ \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2 \right] - \eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \tag{E.130}$$

(unbiased PG, by Lemma 31) (E.131)

$$\leq \frac{3 \cdot K}{2} \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 - \eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2, \quad \text{(by Lemma 31)} \tag{E.132}$$

$$= -\frac{1}{6 \cdot K} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^3 \quad \left( \text{using } \eta = \frac{1}{3 \cdot K} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \right) \tag{E.133}$$

$$\leq -\frac{1}{6 \cdot K} \cdot \mathbb{E}\left[\pi_{\theta_t}(a^*)^3\right] \cdot \mathbb{E}\left[\delta(\theta_t)^3\right] \quad \text{(by Lemma 3)} \tag{E.134}$$

$$\leq -\frac{c}{6 \cdot K} \cdot \left(\mathbb{E}\left[\delta(\theta_t)\right]\right)^3, \quad \text{(by Jensen's inequality)} \tag{E.135}$$

where

$$c \coloneqq \inf_{t \geq 1} \mathbb{E}\left[\pi_{\theta_t}(a^*)^3\right] \tag{E.136}$$

$$\geq \inf_{t \geq 1} \left(\mathbb{E}\left[\pi_{\theta_t}(a^*)\right]\right)^3, \quad \text{(by Jensen's inequality)} \tag{E.137}$$

$$> 0, \tag{E.138}$$

and the last inequality is from Lemma 5, since the expected iteration equals the true gradient update, which converges to global optimal policy. Denote

$\tilde{\delta}(\theta_t) \coloneqq \mathbb{E}\left[\delta(\theta_t)\right]$. We have, for all $t \geq 1$,

$$\frac{1}{\tilde{\delta}(\theta_t)^2} = \frac{1}{\tilde{\delta}(\theta_1)^2} + \sum_{s=1}^{t-1}\left[\frac{1}{\tilde{\delta}(\theta_{s+1})^2} - \frac{1}{\tilde{\delta}(\theta_s)^2}\right] \tag{E.139}$$

$$= \frac{1}{\tilde{\delta}(\theta_1)^2} + \sum_{s=1}^{t-1}\frac{1}{\tilde{\delta}(\theta_{s+1})^2}\cdot\left[1 - \frac{\tilde{\delta}(\theta_{s+1})^2}{\tilde{\delta}(\theta_s)^2}\right] \tag{E.140}$$

$$\geq \frac{1}{\tilde{\delta}(\theta_1)^2} + \sum_{s=1}^{t-1}\frac{2}{\cancel{\tilde{\delta}(\theta_{s+1})^2}}\cdot\frac{\cancel{\tilde{\delta}(\theta_{s+1})^2}}{\tilde{\delta}(\theta_s)^2}\cdot\left[1 - \frac{\tilde{\delta}(\theta_{s+1})}{\tilde{\delta}(\theta_s)}\right] \tag{E.141}$$

$$\left(1 - x^2 \geq 2\cdot x^2\cdot(1-x) \text{ for all } x \in (0,1]\right) \tag{E.142}$$

$$= \frac{1}{\tilde{\delta}(\theta_1)^2} + 2\cdot\sum_{s=1}^{t-1}\frac{1}{\tilde{\delta}(\theta_s)^3}\cdot\left(\tilde{\delta}(\theta_s) - \tilde{\delta}(\theta_{s+1})\right) \tag{E.143}$$

$$\geq \frac{1}{\tilde{\delta}(\theta_1)^2} + 2\cdot\sum_{s=1}^{t-1}\frac{1}{\cancel{\tilde{\delta}(\theta_s)^3}}\cdot\frac{c}{6\cdot K}\cdot\cancel{\tilde{\delta}(\theta_s)^3} \text{ (by Eq. (E.128))} \tag{E.144}$$

$$= \frac{1}{\tilde{\delta}(\theta_1)^2} + \frac{c}{3\cdot K}\cdot(t-1) \tag{E.145}$$

$$\geq \frac{c\cdot t}{3\cdot K}, \qquad \left(\tilde{\delta}(\theta_1)^2 \leq 1 < \frac{3\cdot K}{c}\right) \tag{E.146}$$

which implies that,

$$\mathop{\mathbb{E}}_{a_t\sim\pi_{\theta_t}(\cdot)}\left[(\pi^* - \pi_{\theta_t})^\top r\right] \leq \frac{\sqrt{3\cdot K}}{\sqrt{c}}\cdot\frac{1}{\sqrt{t}}, \tag{E.147}$$

where $c$ is from Eq. (E.136). This implies $(\pi^* - \pi_{\theta_t})^\top r \to 0$ as $t \to \infty$ with probability 1. □

**Lemma 32.** For NPG, we have, $\mathbb{E}_{a\sim\pi_\theta(\cdot)}\left[\hat{r}\right] = r$, and $\mathbb{E}_{a\sim\pi_\theta(\cdot)}\left\|\hat{r}\right\|_2^2 = \sum_{a\in[K]}\frac{r(a)^2}{\pi_\theta(a)}$.

*Proof.* **First part.** $\mathbb{E}_{a\sim\pi_\theta(\cdot)}\left[\hat{r}\right] = r$.

We have, for all $i \in [K]$,

$$\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)}\left[\hat{r}(i)\right] = \sum_{a\in[K]}\pi_\theta(a)\cdot\frac{\mathbb{I}\left\{a=i\right\}}{\pi_\theta(i)}\cdot r(i) = r(i). \tag{E.148}$$

**Second part.** $\mathbb{E}_{a\sim\pi_\theta(\cdot)}\left\|\hat{r}\right\|_2^2 = \sum_{a\in[K]}\frac{r(a)^2}{\pi_\theta(a)}$.

The squared $\ell_2$ norm of natural policy gradient is,

$$\left\|\hat{r}\right\|_2^2 = \sum_i \hat{r}(i)^2 = \sum_i \frac{\left(\mathbb{I}\left\{a=i\right\}\right)^2}{\pi_\theta(i)^2}\cdot r(i)^2 \tag{E.149}$$

$$= \sum_i \frac{\mathbb{I}\left\{a=i\right\}}{\pi_\theta(i)^2}\cdot r(i)^2. \tag{E.150}$$

The expected squared norm is,

$$\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)} \|\hat{r}\|_2^2 = \sum_{a\in[K]} \pi_\theta(a) \cdot \sum_i \frac{\mathbb{I}\{a=i\}}{\pi_\theta(i)^2} \cdot r(i)^2 \tag{E.151}$$

$$= \sum_{a\in[K]} \pi_\theta(a) \cdot \frac{1}{\pi_\theta(a)^2} \cdot r(a)^2 \tag{E.152}$$

$$= \sum_{a\in[K]} \frac{r(a)^2}{\pi_\theta(a)}. \qquad \square$$

**Theorem 23.** Using Update 7, we have: **(i)** with positive probability, as $t\to\infty$, $\sum_{a\neq a^*}\pi_{\theta_t}(a)\to 1$; **(ii)** $\forall a\in[K]$, with positive probability, $\pi_{\theta_t}(a)\to 1$, as $t\to\infty$.

*Proof.* **First part.** With positive probability, $\sum_{a\neq a^*}\pi_{\theta_t}(a)\to 1$ as $t\to\infty$.

We show that $\prod_{t=1}^{\infty}\Big(\sum_{a\neq a^*}\pi_{\theta_t}(a)\Big) > 0$, which implies $\sum_{a\neq a^*}\pi_{\theta_t}(a)\to 1$ as $t\to\infty$ with positive probability. The meaning of $\prod_{t=1}^{\infty}\Big(\sum_{a\neq a^*}\pi_{\theta_t}(a)\Big)$ is "the probability of sampling sub-optimal actions forever using on-policy sampling $a_t\sim\pi_{\theta_t}(\cdot)$". Note that,

$$\prod_{t=1}^{\infty}\Big(\sum_{a\neq a^*}\pi_{\theta_t}(a)\Big) = \lim_{T\to\infty}\prod_{t=1}^{T}\Big(\sum_{a\neq a^*}\pi_{\theta_t}(a)\Big) \tag{E.153}$$

$$= \lim_{T\to\infty}\Pr\left(a_1\neq a^*, a_2\neq a^*, \cdots, a_T\neq a^* \mid a_t\sim\pi_{\theta_t}(\cdot),\ \forall t\geq 1\right) \tag{E.154}$$

$$= \lim_{T\to\infty}\prod_{t=1}^{T}\Pr\left(a_t\neq a^* \mid a_1\neq a^*, a_2\neq a^*, \cdots, a_{t-1}\neq a^*\right). \tag{E.155}$$

$$\text{(by chain rule)} \tag{E.156}$$

Next, we calculate $\Pr\left(a_t\neq a^* \mid a_1\neq a^*, a_2\neq a^*, \cdots, a_{t-1}\neq a^*\right)$, i.e., the sum of probabilities of all sub-optimal actions $\sum_{a\neq a^*}\pi_{\theta_t}(a)$ at $t$-th iteration, given that the optimal action $a^*$ has not been sampled for the first $t-1$ iterations. Now suppose $a_1\neq a^*, a_2\neq a^*, \cdots, a_{t-1}\neq a^*$. We have, for each sub-optimal

action $a \neq a^*$,

$$\theta_t(a) = \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \hat{r}_s(a) \qquad \text{(by Update 7)} \tag{E.157}$$

$$= \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \frac{\mathbb{I}\{a_s = a\}}{\pi_{\theta_s}(a)} \cdot r(a) \qquad \text{(by Definition 10)} \tag{E.158}$$

$$\geq \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} \cdot r(a) \qquad (\pi_{\theta_s}(a) \in (0,1),\ r(a) \in (0,1]) \tag{E.159}$$

$$\geq \theta_1(a) + \eta \cdot r_{\min} \cdot \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}, \qquad \left(r_{\min} \coloneqq \min_{a \neq a^*} r(a)\right) \tag{E.160}$$

where $r_{\min} \in (0,1]$ according to Assumption 3, i.e., $r(a) \in (0,1]$ for all $a \in [K]$. Then we have,

$$\sum_{a \neq a^*} \exp\{\theta_t(a)\} \geq (K-1) \cdot \exp\left\{\frac{\sum_{a \neq a^*} \theta_t(a)}{K-1}\right\} \tag{E.161}$$

$$\text{(by Jensen's inequality)} \tag{E.162}$$

$$\geq (K-1) \cdot \exp\left\{\frac{\sum_{a \neq a^*} \theta_1(a) + \eta \cdot r_{\min} \cdot \sum_{a \neq a^*} \sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}}{K-1}\right\} \tag{E.163}$$

$$\text{(by Eq. (E.157))} \tag{E.164}$$

$$= (K-1) \cdot \exp\left\{\frac{\sum_{a \neq a^*} \theta_1(a) + \eta \cdot r_{\min} \cdot (t-1)}{K-1}\right\}. \tag{E.165}$$

$$(a_1 \neq a^*, a_2 \neq a^*, \cdots, a_{t-1} \neq a^*) \tag{E.166}$$

On the other hand, we have,

$$\theta_t(a^*) = \theta_1(a^*) + \eta \cdot \sum_{s=1}^{t-1} \frac{\mathbb{I}\{a_s = a^*\}}{\pi_{\theta_s}(a^*)} \cdot r(a^*) \tag{E.167}$$

$$\text{(by Update 7 and Definition 10)} \tag{E.168}$$

$$= \theta_1(a^*). \qquad (a_s \neq a^* \text{ for all } s \in \{1, 2, \ldots t-1\}) \tag{E.169}$$

Next, we have,

$$\sum_{a\neq a^*}\pi_{\theta_t}(a)=1-\pi_{\theta_t}(a^*) \tag{E.170}$$

$$=1-\frac{\exp\{\theta_t(a^*)\}}{\sum_{a\neq a^*}\exp\{\theta_t(a)\}+\exp\{\theta_t(a^*)\}} \tag{E.171}$$

$$\geq 1-\frac{\exp\{\theta_1(a^*)\}}{(K-1)\cdot\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)+\eta\cdot r_{\min}\cdot(t-1)}{K-1}\right\}+\exp\{\theta_1(a^*)\}}. \tag{E.172}$$

$$\text{(by Eqs. (E.161) and (E.167))} \tag{E.173}$$

According to Lemma 58, for all $x\in(0,1)$,

$$1-x\geq\exp\left\{\frac{-1}{1/x-1}\right\}. \tag{E.174}$$

Let

$$x=\frac{\exp\{\theta_1(a^*)\}}{(K-1)\cdot\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)+\eta\cdot r_{\min}\cdot(t-1)}{K-1}\right\}+\exp\{\theta_1(a^*)\}}\in(0,1). \tag{E.175}$$

We have,

$$\sum_{a\neq a^*}\pi_{\theta_t}(a)\geq\exp\left\{\frac{-1}{\frac{(K-1)\cdot\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)+\eta\cdot r_{\min}\cdot(t-1)}{K-1}\right\}+\exp\{\theta_1(a^*)\}}{\exp\{\theta_1(a^*)\}}-1}\right\} \tag{E.176}$$

$$\text{(by Eqs. (E.170) and (E.174))} \tag{E.177}$$

$$=\exp\left\{\frac{-\exp\{\theta_1(a^*)\}}{(K-1)\cdot\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)+\eta\cdot r_{\min}\cdot(t-1)}{K-1}\right\}}\right\}. \tag{E.178}$$

Therefore we have,

$$\prod_{t=1}^{\infty}\left(\sum_{a\neq a^*}\pi_{\theta_t}(a)\right) \tag{E.179}$$

$$\geq \prod_{t=1}^{\infty}\exp\left\{\frac{-\exp\{\theta_1(a^*)\}}{(K-1)\cdot\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)+\eta\cdot r_{\min}\cdot(t-1)}{K-1}\right\}}\right\} \tag{E.180}$$

(by Eq. (E.176)) (E.181)

$$= \exp\left\{-\frac{\exp\{\theta_1(a^*)\}}{\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)}{K-1}\right\}}\cdot\frac{\exp\left\{\frac{\eta\cdot r_{\min}}{K-1}\right\}}{K-1}\cdot\sum_{t=1}^{\infty}\frac{1}{\exp\left\{\frac{\eta\cdot r_{\min}\cdot t}{K-1}\right\}}\right\} \tag{E.182}$$

$$\geq \exp\left\{-\frac{\exp\{\theta_1(a^*)\}}{\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)}{K-1}\right\}}\cdot\frac{\exp\left\{\frac{\eta\cdot r_{\min}}{K-1}\right\}}{K-1}\cdot\int_{t=0}^{\infty}\frac{1}{\exp\left\{\frac{\eta\cdot r_{\min}\cdot t}{K-1}\right\}}dt\right\} \tag{E.183}$$

$$= \exp\left\{-\frac{\exp\{\theta_1(a^*)\}}{\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)}{K-1}\right\}}\cdot\frac{\exp\left\{\frac{\eta\cdot r_{\min}}{K-1}\right\}}{K-1}\cdot\frac{K-1}{\eta\cdot r_{\min}}\right\} \tag{E.184}$$

$$= \exp\left\{-\frac{\exp\{\theta_1(a^*)\}}{\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)}{K-1}\right\}}\cdot\frac{\exp\left\{\frac{\eta\cdot r_{\min}}{K-1}\right\}}{\eta\cdot r_{\min}}\right\}. \tag{E.185}$$

Note that $r_{\min}\in\Theta(1)$, $\exp\{\theta_1(a^*)\}\in\Theta(1)$, $\eta\in\Theta(1)$, $\exp\left\{\frac{\eta\cdot r_{\min}}{K-1}\right\}\in\Theta(1)$ and,

$$\exp\left\{\frac{\sum_{a\neq a^*}\theta_1(a)}{K-1}\right\}\in\Theta(1). \tag{E.186}$$

Therefore, we have "the probability of sampling sub-optimal actions forever using on-policy sampling $a_t\sim\pi_{\theta_t}(\cdot)$" is lower bounded by a constant of $\frac{1}{\exp\{\Theta(1)\}}\in\Theta(1)$, which implies that with positive probability $\Theta(1)$, we have $\sum_{a\neq a^*}\pi_{\theta_t}(a)\to 1$ as $t\to\infty$.

**Second part.** $\forall a\in[K]$, with positive probability, $\pi_{\theta_t}(a)\to 1$, as $t\to\infty$.

For each action $a\in[K]$, we show that $\prod_{t=1}^{\infty}\pi_{\theta_t}(a)>0$, which implies $\pi_{\theta_t}(a)\to 1$ as $t\to\infty$. The meaning of $\prod_{t=1}^{\infty}\pi_{\theta_t}(a)$ is "the probability of

sampling action $a$ forever using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$". Note that,

$$\prod_{t=1}^{\infty} \pi_{\theta_t}(a) = \lim_{T\to\infty} \prod_{t=1}^{T} \pi_{\theta_t}(a) \tag{E.187}$$

$$= \lim_{T\to\infty} \Pr\left(a_1 = a, a_2 = a, \cdots, a_T = a \mid a_t \sim \pi_{\theta_t}(\cdot),\ \forall t \geq 1\right) \tag{E.188}$$

$$= \lim_{T\to\infty} \prod_{t=1}^{T} \Pr\left(a_t = a \mid a_1 = a, a_2 = a, \cdots, a_{t-1} = a\right). \tag{E.189}$$

$$\text{(by chain rule)} \tag{E.190}$$

Next, we calculate $\Pr\left(a_t = a \mid a_1 = a, a_2 = a, \cdots, a_{t-1} = a\right)$, i.e., the probability of sampling action $a$ at $t$-th iteration, given that the action $a$ has been sampled for the first $t-1$ iterations. Now suppose $a_1 = a, a_2 = a, \cdots, a_{t-1} = a$. We have,

$$\theta_t(a) = \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \hat{r}_s(a) \qquad \text{(by Update 7)} \tag{E.191}$$

$$= \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \frac{\mathbb{I}\left\{a_s = a\right\}}{\pi_{\theta_s}(a)} \cdot r(a) \qquad \text{(by Definition 10)} \tag{E.192}$$

$$= \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \frac{r(a)}{\pi_{\theta_s}(a)} \qquad (a_s = a \text{ for all } s \in \{1, 2, \ldots t-1\}) \tag{E.193}$$

$$\geq \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} r(a) \qquad (\pi_{\theta_s}(a) \in (0, 1)) \tag{E.194}$$

$$= \theta_1(a) + \eta \cdot r(a) \cdot (t-1)\,. \tag{E.195}$$

On the other hand, we have, for any other action $a' \neq a$,

$$\theta_t(a') = \theta_1(a') + \eta \cdot \sum_{s=1}^{t-1} \frac{\mathbb{I}\left\{a_s = a'\right\}}{\pi_{\theta_s}(a')} \cdot r(a') \tag{E.196}$$

$$\text{(by Update 7 and Definition 10)} \tag{E.197}$$

$$= \theta_1(a'). \qquad (a_s \neq a' \text{ for all } s \in \{1, 2, \ldots t\}) \tag{E.198}$$

Therefore, we have,

$$\pi_{\theta_t}(a) = 1 - \sum_{a' \neq a} \pi_{\theta_t}(a') \tag{E.199}$$

$$= 1 - \frac{\sum_{a' \neq a} \exp\{\theta_t(a')\}}{\exp\{\theta_t(a)\} + \sum_{a' \neq a} \exp\{\theta_t(a')\}} \tag{E.200}$$

$$\geq 1 - \frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a) + \eta \cdot r(a) \cdot (t-1)\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}}. \tag{E.201}$$

(by Eqs. (E.191) and (E.196)) (E.202)

Let

$$x = \frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a) + \eta \cdot r(a) \cdot (t-1)\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}} \in (0, 1). \tag{E.203}$$

We have,

$$\pi_{\theta_t}(a) \geq 1 - x \qquad \text{(by Eq. (E.199))} \tag{E.204}$$

$$\geq \exp\left\{ \frac{-1}{\frac{\exp\{\theta_1(a)+\eta \cdot r(a) \cdot (t-1)\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}}{\sum_{a' \neq a} \exp\{\theta_1(a')\}} - 1} \right\} \tag{E.205}$$

(by Eq. (E.174)) (E.206)

$$= \exp\left\{ \frac{-\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a) + \eta \cdot r(a) \cdot (t-1)\}} \right\}. \tag{E.207}$$

Therefore we have,

$$\prod_{t=1}^{\infty} \pi_{\theta_t}(a) \tag{E.208}$$

$$\geq \prod_{t=1}^{\infty} \exp\left\{ \frac{-\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a) + \eta \cdot r(a) \cdot (t-1)\}} \right\} \qquad \text{(by Eq. (E.204))} \tag{E.209}$$

$$= \exp\left\{ -\sum_{a' \neq a} \exp\{\theta_1(a')\} \cdot \frac{\exp\{\eta \cdot r(a)\}}{\exp\{\theta_1(a)\}} \cdot \sum_{t=1}^{\infty} \frac{1}{\exp\{\eta \cdot r(a) \cdot t\}} \right\} \tag{E.210}$$

$$\geq \exp\left\{ -\sum_{a' \neq a} \exp\{\theta_1(a')\} \cdot \frac{\exp\{\eta \cdot r(a)\}}{\exp\{\theta_1(a)\}} \cdot \int_{t=0}^{\infty} \frac{1}{\exp\{\eta \cdot r(a) \cdot t\}} dt \right\} \tag{E.211}$$

$$= \exp\left\{ -\frac{\exp\{\eta \cdot r(a)\}}{\eta \cdot r(a)} \cdot \frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a)\}} \right\} \tag{E.212}$$

$$\in \Omega(1), \tag{E.213}$$

where the last line is due to $r(a) \in \Theta(1)$, $\exp\{\theta_1(a)\} \in \Theta(1)$ for all $a \in [K]$, and $\eta \in \Theta(1)$. With Eq. (E.208), we have "the probability of sampling action $a$ forever using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$" is lower bounded by a constant of $\Omega(1)$. Therefore, for all $a \in [K]$, with positive probability $\Omega(1)$, $\pi_{\theta_t}(a) \to 1$, as $t \to \infty$. □

**Lemma 57.** *Using on-policy IS estimator of Definition 10, the stochastic GNPG is biased, i.e.,*

$$\mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)}\left[\frac{d\pi_\theta^\top \hat{r}}{d\theta}\Big/\left\|\frac{d\pi_\theta^\top \hat{r}}{d\theta}\right\|_2\right] \neq \frac{d\pi_\theta^\top r}{d\theta}\Big/\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2. \tag{E.214}$$

*Proof.* Consider a two-action example with $r(1) > r(2)$. The true normalized PG of $a^* = 1$ is,

$$g(1) \coloneqq \frac{d\pi_\theta^\top r}{d\theta(1)}\Big/\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \tag{E.215}$$

$$= \frac{\pi_\theta(1)\cdot\left(r(1)-\pi_\theta^\top r\right)}{\sqrt{\pi_\theta(1)^2\cdot\left(r(1)-\pi_\theta^\top r\right)^2+\pi_\theta(2)^2\cdot\left(r(2)-\pi_\theta^\top r\right)^2}} \tag{E.216}$$

$$= \frac{\pi_\theta(1)\cdot\pi_\theta(2)\cdot(r(1)-r(2))}{\sqrt{\pi_\theta(1)^2\cdot\pi_\theta(2)^2\cdot(r(1)-r(2))^2+\pi_\theta(1)^2\cdot\pi_\theta(2)^2\cdot(r(1)-r(2))^2}} \tag{E.217}$$

$$= \frac{1}{\sqrt{2}}. \tag{E.218}$$

On the other hand, the stochastic normalized PG of $a^* = 1$ is,

$$\hat{g}(1) \coloneqq \mathop{\mathbb{E}}_{a\sim\pi_\theta(\cdot)}\left[\frac{d\pi_\theta^\top \hat{r}}{d\theta(1)} \Big/ \left\|\frac{d\pi_\theta^\top \hat{r}}{d\theta}\right\|_2\right] \tag{E.219}$$

$$= \pi_\theta(1) \cdot \frac{\pi_\theta(1)\cdot\left(\frac{r(1)}{\pi_\theta(1)} - \pi_\theta(1)\cdot\frac{r(1)}{\pi_\theta(1)}\right)}{\sqrt{\pi_\theta(1)^2\cdot\left(\frac{r(1)}{\pi_\theta(1)} - \pi_\theta(1)\cdot\frac{r(1)}{\pi_\theta(1)}\right)^2 + \pi_\theta(2)^2\cdot\left(0 - \pi_\theta(1)\cdot\frac{r(1)}{\pi_\theta(1)}\right)^2}} \tag{E.220}$$

$$+ \pi_\theta(2) \cdot \frac{\pi_\theta(1)\cdot\left(0 - \pi_\theta(2)\cdot\frac{r(2)}{\pi_\theta(2)}\right)}{\sqrt{\pi_\theta(1)^2\cdot\left(0 - \pi_\theta(2)\cdot\frac{r(2)}{\pi_\theta(2)}\right)^2 + \pi_\theta(2)^2\cdot\left(\frac{r(2)}{\pi_\theta(2)} - \pi_\theta(2)\cdot\frac{r(2)}{\pi_\theta(2)}\right)^2}} \tag{E.221}$$

$$= \pi_\theta(1) \cdot \frac{\pi_\theta(2)\cdot r(1)}{\sqrt{\pi_\theta(2)^2\cdot r(1)^2 + \pi_\theta(2)^2\cdot r(1)^2}} \tag{E.222}$$

$$- \pi_\theta(2) \cdot \frac{\pi_\theta(1)\cdot r(2)}{\sqrt{\pi_\theta(1)^2\cdot r(2)^2 + \pi_\theta(1)^2\cdot r(2)^2}} \tag{E.223}$$

$$= \frac{1}{\sqrt{2}}\cdot(\pi_\theta(1) - \pi_\theta(2))\,. \tag{E.224}$$

It is clear that the true normalized PG of $a^* = 1$ is always positive $g(1) > 0$, while the expectation of the stochastic normalized PG estimator of $a^* = 1$ is negative when $\pi_\theta(1) < \pi_\theta(2)$. □

**Theorem 24.** Using Update 8, we have, $\forall a \in [K]$, with positive probability, $\pi_{\theta_t}(a) \to 1$, as $t \to \infty$.

*Proof.* The proof is similar to the second part of Theorem 23. We first calculate the stochastic normalized PG in each iteration. Denote $a_t$ as the action sampled at $t$-th iteration. We have,

$$\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a_t)} = \pi_{\theta_t}(a_t)\cdot(\hat{r}_t(a_t) - \pi_{\theta_t}^\top \hat{r}_t) \tag{E.225}$$

$$= \pi_{\theta_t}(a_t)\cdot\left(\frac{r(a_t)}{\pi_{\theta_t}(a_t)} - \pi_{\theta_t}(a_t)\cdot\frac{r(a_t)}{\pi_{\theta_t}(a_t)}\right) \qquad \text{(by Definition 10)} \tag{E.226}$$

$$= (1 - \pi_{\theta_t}(a_t))\cdot r(a_t). \tag{E.227}$$

On the other hand, for all $a' \neq a_t$,

$$\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a')} = \pi_{\theta_t}(a') \cdot (\hat{r}_s(a') - \pi_{\theta_t}^\top \hat{r}_t) \tag{E.228}$$

$$= \pi_{\theta_t}(a') \cdot \left(0 - \pi_{\theta_t}(a_t) \cdot \frac{r(a_t)}{\pi_{\theta_t}(a_t)}\right) \qquad \text{(by Definition 10)} \tag{E.229}$$

$$= -\pi_{\theta_t}(a') \cdot r(a_t). \tag{E.230}$$

Therefore, the stochastic PG norm is,

$$\left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2 = \left[ \left( \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a_t)} \right)^2 + \sum_{a' \neq a_t} \left( \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a')} \right)^2 \right]^{\frac{1}{2}} \tag{E.231}$$

$$= \left[ (1 - \pi_{\theta_t}(a_t))^2 \cdot r(a_t)^2 + \sum_{a' \neq a_t} \pi_{\theta_t}(a')^2 \cdot r(a_t)^2 \right]^{\frac{1}{2}} \tag{E.232}$$

$$\text{(by Eqs. (E.225) and (E.228))} \tag{E.233}$$

$$\leq \left[ (1 - \pi_{\theta_t}(a_t))^2 \cdot r(a_t)^2 + \left( \sum_{a' \neq a_t} \pi_{\theta_t}(a') \right)^2 \cdot r(a_t)^2 \right]^{\frac{1}{2}} \tag{E.234}$$

$$(\|x\|_2 \leq \|x\|_1) \tag{E.235}$$

$$= \sqrt{2} \cdot (1 - \pi_{\theta_t}(a_t)) \cdot r(a_t). \tag{E.236}$$

Similar to the second part of Theorem 23, we show that $\prod_{t=1}^{\infty} \pi_{\theta_t}(a) > 0$, which implies $\pi_{\theta_t}(a) \to 1$ as $t \to \infty$. The meaning of $\prod_{t=1}^{\infty} \pi_{\theta_t}(a)$ is "the probability of sampling action $a$ forever using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$". Note that,

$$\prod_{t=1}^{\infty} \pi_{\theta_t}(a) = \lim_{T \to \infty} \prod_{t=1}^{T} \pi_{\theta_t}(a) \tag{E.237}$$

$$= \lim_{T \to \infty} \Pr\left(a_1 = a, a_2 = a, \cdots, a_T = a \mid a_t \sim \pi_{\theta_t}(\cdot),\ \forall t \geq 1\right) \tag{E.238}$$

$$= \lim_{T \to \infty} \prod_{t=1}^{T} \Pr\left(a_t = a \mid a_1 = a, a_2 = a, \cdots, a_{t-1} = a\right). \tag{E.239}$$

$$\text{(by chain rule)} \tag{E.240}$$

Next, we calculate $\Pr\left(a_t = a \mid a_1 = a, a_2 = a, \cdots, a_{t-1} = a\right)$, i.e., the probability of sampling action $a$ at $t$-th iteration, given that the action $a$ has been sampled for the first $t-1$ iterations. Now suppose $a_1 = a, a_2 = a, \cdots, a_{t-1} = a$.

We have,

$$\theta_t(a) = \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \frac{d\pi_{\theta_s}^\top \hat{r}_s}{d\theta_s(a)} \Big/ \left\| \frac{d\pi_{\theta_s}^\top \hat{r}_s}{d\theta_s} \right\|_2 \qquad \text{(by Update 8)} \tag{E.241}$$

$$\geq \theta_1(a) + \eta \cdot \sum_{s=1}^{t-1} \frac{(1 - \pi_{\theta_s}(a)) \cdot r(a)}{\sqrt{2} \cdot (1 - \pi_{\theta_s}(a)) \cdot r(a)} \tag{E.242}$$

$$\text{(by Eqs. (E.225) and (E.231))} \tag{E.243}$$

$$= \theta_1(a) + \frac{\eta}{\sqrt{2}} \cdot (t-1)\,. \tag{E.244}$$

On the other hand, for all $a' \neq a$, we have,

$$\theta_t(a') = \theta_1(a') - \eta \cdot \sum_{s=1}^{t-1} (\pi_{\theta_s}(a') \cdot r(a)) \Big/ \left\| \frac{d\pi_{\theta_s}^\top \hat{r}_s}{d\theta_s} \right\|_2 \tag{E.245}$$

$$\text{(by Update 8 and Eq. (E.228))} \tag{E.246}$$

$$\leq \theta_1(a'). \tag{E.247}$$

Then we have,

$$\pi_{\theta_t}(a) = 1 - \frac{\sum_{a' \neq a} \exp\{\theta_t(a')\}}{\exp\{\theta_t(a)\} + \sum_{a' \neq a} \exp\{\theta_t(a')\}} \tag{E.248}$$

$$\geq 1 - \frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\left\{\theta_1(a) + \frac{\eta}{\sqrt{2}} \cdot (t-1)\right\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}} \tag{E.249}$$

$$\text{(by Eqs. (E.241) and (E.245))} \tag{E.250}$$

$$\geq \exp\left\{ \frac{-\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\left\{\theta_1(a) + \frac{\eta}{\sqrt{2}} \cdot (t-1)\right\}} \right\}. \qquad \text{(by Eq. (E.174))} \tag{E.251}$$

Using similar calculation to Eq. (E.176), we have,

$$\prod_{t=1}^{\infty} \pi_{\theta_t}(a) \tag{E.252}$$

$$\geq \prod_{t=1}^{\infty} \exp\left\{ \frac{-\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\left\{\theta_1(a) + \frac{\eta}{\sqrt{2}} \cdot (t-1)\right\}} \right\} \qquad \text{(by Eq. (E.248))} \tag{E.253}$$

$$= \exp\left\{ -\frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a)\}} \cdot \exp\left\{\frac{\eta}{\sqrt{2}}\right\} \cdot \sum_{t=1}^{\infty} \frac{1}{\exp\left\{\frac{\eta}{\sqrt{2}} \cdot t\right\}} \right\} \tag{E.254}$$

$$\geq \exp\left\{ -\frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a)\}} \cdot \exp\left\{\frac{\eta}{\sqrt{2}}\right\} \cdot \int_{t=0}^{\infty} \frac{1}{\exp\left\{\frac{\eta}{\sqrt{2}} \cdot t\right\}} dt \right\} \tag{E.255}$$

$$= \exp\left\{ -\frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a)\}} \cdot \frac{\sqrt{2} \cdot \exp\left\{\frac{\eta}{\sqrt{2}}\right\}}{\eta} \right\} \tag{E.256}$$

$$\in \Omega(1), \tag{E.257}$$

where the last line is due to, $\exp\{\theta_1(a)\} \in \Theta(1)$ for all $a \in [K]$, and $\eta \in \Theta(1)$. With Eq. (E.252), we have "the probability of sampling action $a$ forever using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$" is lower bounded by a constant of $\Omega(1)$. Therefore, for all $a \in [K]$, with positive probability $\Omega(1)$, $\pi_{\theta_t}(a) \to 1$, as $t \to \infty$. □

## E.2 Proofs for Section 5.3: Committal Rate

**Theorem 25.** Consider a policy optimization method $\mathcal{A}$. Fix $r \in (0,1]^K$ an action $a \in [K]$ which is sub-optimal under $r$ so that $\kappa(\mathcal{A}, a) > 1$. Fix $\theta_1 \in \mathbb{R}^K$ so that $\pi_{\theta_1}(a) > 0$ and let $\{\theta_t\}_{t\geq 1}$ be the parameter sequence obtained by using $\mathcal{A}$ with online sampling, i.e., when $a_t \sim \pi_{\theta_t}(\cdot)$. Then, the event $\mathcal{E} = \{a_t = a \text{ holds for all } t \geq 1\}$ happens with positive probability, and it also holds that $\pi_{\theta_t}$ converges to a sub-optimal deterministic policy with positive probability.

*Proof.* Suppose $\kappa(\mathcal{A}, a) > 1$ for action $a \in [K]$. For convenience, denote $\alpha \coloneqq \kappa(\mathcal{A}, a)$. Define the history of actions for the first $t$ iterations,

$$\mathcal{H}_t \coloneqq (a_1, a_2, \cdots, a_t)\,. \tag{E.258}$$

Given the historical iterations, sampled actions and rewards, the next iteration is a deterministic result of the algorithm,

$$\theta_t = \mathcal{A}\left(\theta_1, a_1, r(a_1), \theta_2, a_2, r(a_2), \cdots, \theta_{t-1}, a_{t-1}, r(a_{t-1})\right). \tag{E.259}$$

We have, almost surely for all $a$ and $t \geq 1$,

$$\Pr\left(a_t = a \mid \mathcal{H}_{t-1}\right) = \pi_{\theta_t}(a). \tag{E.260}$$

Define the following event, for all $t \geq 1$,

$$\mathcal{E}_t \coloneqq \{a_s = a, \text{ for all } 1 \leq s \leq t\}\,. \tag{E.261}$$

We have $\mathcal{E}_t \supseteq \mathcal{E}_{t+1}$, and $\mathcal{E}_t$ approaches the limit event,

$$\mathcal{E} \coloneqq \{a_t = a, \text{ for all } t \geq 1\}\,. \tag{E.262}$$

We have $\Pr(\mathcal{E}_t)$ is monotonically decreasing and lower bounded by zero. According to monotone convergence theorem,

$$\Pr(\mathcal{E}) = \lim_{t\to\infty} \Pr(\mathcal{E}_t). \tag{E.263}$$

Next, we prove by induction on $t$ the following holds

$$\Pr(\mathcal{E}_t) = \Pr(a_t = a \mid \mathcal{E}_{t-1}) \cdot \Pr(\mathcal{E}_{t-1}) \tag{E.264}$$

$$= \prod_{s=1}^{t} \pi_{\tilde{\theta}_s}(a), \tag{E.265}$$

where $\tilde{\theta}_1 = \theta_1$, and,

$$\tilde{\theta}_t = \mathcal{A}\big(\theta_1, \underbrace{a, r(a)}_{s=1}, \cdots, \underbrace{a, r(a)}_{s=t-1}\big), \tag{E.266}$$

which means $a$ is used for the first $t-1$ iterations.

First, by definition of $\tilde{\theta}_1$, we have,

$$\Pr(\mathcal{E}_1) = \pi_{\theta_1}(a) = \pi_{\tilde{\theta}_1}(a), \tag{E.267}$$

where the first equation is from Eq. (E.260). Suppose the equation holds up to $t-1$. We have,

$$\Pr(\mathcal{E}_t) = \mathbb{E}\left[\Pr(a_t = a, \cdots, a_1 = a \mid \mathcal{H}_{t-1})\right] \quad \text{(by the tower rule)} \tag{E.268}$$

$$= \mathbb{E}\left[\mathbb{I}\left\{a_{t-1} = a, \cdots, a_1 = a\right\} \cdot \Pr(a_t = a \mid \mathcal{H}_{t-1})\right] \tag{E.269}$$

$$\text{(determined by } \mathcal{H}_{t-1}) \tag{E.270}$$

$$= \mathbb{E}\left[\mathbb{I}\left\{a_{t-1} = a, \cdots, a_1 = a\right\} \cdot \pi_{\theta_t}(a)\right] \quad \text{(by Eq. (E.260))} \tag{E.271}$$

$$= \mathbb{E}\left[\mathbb{I}\left\{a_{t-1} = a, \cdots, a_1 = a\right\} \cdot \pi_{\tilde{\theta}_t}(a)\right] \tag{E.272}$$

$$= \pi_{\tilde{\theta}_t}(a) \cdot \Pr(\mathcal{E}_{t-1}) \tag{E.273}$$

$$= \prod_{s=1}^{t} \pi_{\tilde{\theta}_s}(a). \tag{E.274}$$

Next, we show that $\prod_{t=1}^{\infty} \pi_{\tilde{\theta}_t}(a) > 0$. Note that,

$$\prod_{t=1}^{\infty} \pi_{\tilde{\theta}_t}(a) = \lim_{T\to\infty} \prod_{t=1}^{T} \pi_{\tilde{\theta}_t}(a) \tag{E.275}$$

$$= \lim_{T\to\infty} \prod_{t=1}^{T} \Pr\left(a_t = a \mid a_1 = a, a_2 = a, \cdots, a_{t-1} = a\right) \tag{E.276}$$

$$\text{(by chain rule)} \tag{E.277}$$

$$= \lim_{T\to\infty} \prod_{t=1}^{T} \Pr\left(a_t = a \mid \mathcal{E}_{t-1}\right). \tag{E.278}$$

In Eq. (E.278), $\Pr\left(a_t = a \mid \mathcal{E}_{t-1}\right)$ is the value of $\pi_{\tilde{\theta}_t}(a)$ given $\mathcal{A}$ is used when in the first $t-1$ iterations action $a$ is used. This is the sequence used in the definition of committal rate $\kappa$. Further, for simplicity, assume that in the definition of $\kappa$, the supremum is achieved. (To deal with the general case, one can redefine $\alpha$ to be $\alpha = \frac{1+\kappa(\mathcal{A},a)}{2} > 1$). It follows that there exists a universal constant $C > 0$ such that on $\mathcal{E}$, for all $t \geq 1$,

$$1 - \pi_{\tilde{\theta}_t}(a) = t^{\alpha} \cdot \left[1 - \pi_{\tilde{\theta}_t}(a)\right] \cdot \frac{1}{t^{\alpha}} \tag{E.279}$$

$$\leq \frac{C}{t^{\alpha}}. \qquad \text{(by Definition 11)} \tag{E.280}$$

Let $u_t \coloneqq 1 - \pi_{\tilde{\theta}_t}(a) \in (0,1)$ for all $t \geq 1$. We have,

$$\sum_{t=1}^{\infty} u_t \leq \sum_{t=1}^{\infty} \frac{C}{t^{\alpha}} \qquad \text{(by Eq. (E.279))} \tag{E.281}$$

$$< \infty. \qquad \text{(by Lemma 59, } \alpha \coloneqq \kappa(\mathcal{A}, a) > 1) \tag{E.282}$$

Therefore we have,

$$\prod_{t=1}^{\infty} \pi_{\tilde{\theta}_t}(a) = \prod_{t=1}^{\infty} (1 - u_t) \tag{E.283}$$

$$> 0. \qquad \text{(by Lemma 60 and Eq. (E.281))} \tag{E.284}$$

Hence, we have,

$$\Pr(\mathcal{E}) = \lim_{T\to\infty} \Pr(\mathcal{E}_T) = \lim_{T\to\infty} \prod_{t=1}^{T} \pi_{\tilde{\theta}_t}(a) = \prod_{t=1}^{\infty} \pi_{\tilde{\theta}_t}(a) > 0, \tag{E.285}$$

and thus $\pi_{\theta_t}(a) \to 1$ as $t \to \infty$. □

**Theorem 26.** Let Assumption 3 holds. For the stochastic updates NPG and GNPG from Updates 7 and 8 we obtain $\kappa(\text{NPG}, a) = \infty$ and $\kappa(\text{GNPG}, a) = \infty$ for all $a \in [K]$ respectively.

*Proof.* **First part (NPG).** We first show that $\kappa(\text{NPG}, a) = \infty$ for all $a \in [K]$. According to Definition 11, let action $a$ be sampled forever after initialization. We have, for stochastic NPG update,

$$1 - \pi_{\theta_t}(a) = \sum_{a' \neq a} \pi_{\theta_t}(a') \tag{E.286}$$

$$\leq \frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a) + \eta \cdot r(a) \cdot (t-1)\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}}. \tag{E.287}$$

$$\text{(by Eq. (E.199))} \tag{E.288}$$

Since $\exp\{\theta_1(i)\} \in \Theta(1)$ for all $i \in [K]$, we have, for any finite $\alpha \in (0, \infty)$,

$$\lim_{t \to \infty} t^\alpha \cdot [1 - \pi_{\theta_t}(a)] \tag{E.289}$$

$$\leq \lim_{t \to \infty} \frac{t^\alpha \cdot \sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\{\theta_1(a) + \eta \cdot r(a) \cdot (t-1)\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}} \tag{E.290}$$

$$\text{(by Eq. (E.286))} \tag{E.291}$$

$$= \lim_{t \to \infty} \frac{\Theta(t^\alpha)}{\Theta(\exp\{\eta \cdot r(a) \cdot (t-1)\})} = 0, \tag{E.292}$$

which means $\kappa(\text{NPG}, a) = \infty$ for all $a \in [K]$.

**Second part (GNPG).** We next show that $\kappa(\text{GNPG}, a) = \infty$ for all $a \in [K]$. Let action $a$ be sampled forever after initialization. We have, for stochastic GNPG update,

$$1 - \pi_{\theta_t}(a) = \sum_{a' \neq a} \pi_{\theta_t}(a') \tag{E.293}$$

$$\leq \frac{\sum_{a' \neq a} \exp\{\theta_1(a')\}}{\exp\left\{\theta_1(a) + \frac{\eta}{\sqrt{2}} \cdot (t-1)\right\} + \sum_{a' \neq a} \exp\{\theta_1(a')\}}. \tag{E.294}$$

$$\text{(by Eq. (E.248))} \tag{E.295}$$

Using similar arguments to Eq. (E.289), we have $\kappa(\text{GNPG}, a) = \infty$ for all $a \in [K]$. □

**Theorem 27.** Softmax PG obtains $\kappa(\text{PG}, a) = 1$ for all $a \in [K]$.

*Proof.* **First part.** $\kappa(\mathrm{PG}, a) \geq 1$.

According to Definition 11, let action $a$ be sampled forever after initialization. We have, for stochastic PG update,

$$
\begin{aligned}
&\big(1-\pi_{\theta_{t+1}}(a)\big)-\big(1-\pi_{\theta_t}(a)\big) && \text{(E.296)}\\
&\quad=\pi_{\theta_t}(a)-\pi_{\theta_{t+1}}(a)+\left\langle\frac{d\pi_{\theta_t}(a)}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle && \text{(E.297)}\\
&\qquad-\left\langle\frac{d\pi_{\theta_t}(a)}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle && \text{(E.298)}\\
&\quad\leq\frac{5}{4}\cdot\|\theta_{t+1}-\theta_t\|_2^2-\left\langle\frac{d\pi_{\theta_t}(a)}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle \qquad \text{(by Lemma 2)} && \text{(E.299)}\\
&\quad=\frac{5\cdot\eta^2}{4}\cdot\left\|\frac{d\pi_{\theta_t}^\top\hat{r}_t}{d\theta_t}\right\|_2^2-\eta\cdot\left\langle\frac{d\pi_{\theta_t}(a)}{d\theta_t},\frac{d\pi_{\theta_t}^\top\hat{r}_t}{d\theta_t}\right\rangle && \text{(E.300)}\\
&\qquad\text{(using Update 6)} && \text{(E.301)}\\
&\quad=\frac{5\cdot\eta^2}{4}\cdot\left(\sum_{a'\neq a}\pi_{\theta_t}(a')^2\cdot r(a)^2+\big(1-\pi_{\theta_t}(a)\big)^2\cdot r(a)^2\right) && \text{(E.302)}\\
&\qquad-\eta\cdot\left\langle\frac{d\pi_{\theta_t}(a)}{d\theta_t},\frac{d\pi_{\theta_t}^\top\hat{r}_t}{d\theta_t}\right\rangle \qquad \text{(by Eqs. (E.225) and (E.228))} && \text{(E.303)}\\
&\quad\leq\frac{5\cdot\eta^2}{2}\cdot\big(1-\pi_{\theta_t}(a)\big)^2\cdot r(a)^2-\eta\cdot\left\langle\frac{d\pi_{\theta_t}(a)}{d\theta_t},\frac{d\pi_{\theta_t}^\top\hat{r}_t}{d\theta_t}\right\rangle && \text{(E.304)}\\
&\qquad\left(\|x\|_2\leq\|x\|_1\right) && \text{(E.305)}\\
&\quad=\frac{5\cdot\eta^2}{2}\cdot\big(1-\pi_{\theta_t}(a)\big)^2\cdot r(a)^2 && \text{(E.306)}\\
&\qquad-\eta\cdot\pi_{\theta_t}(a)\cdot r(a)\cdot\left(\sum_{a'\neq a}\pi_{\theta_t}(a')^2+\big(1-\pi_{\theta_t}(a)\big)^2\right) && \text{(E.307)}\\
&\qquad\text{(see below)} && \text{(E.308)}\\
&\quad\leq\frac{5\cdot\eta^2}{2}\cdot\big(1-\pi_{\theta_t}(a)\big)^2\cdot r(a)^2-\eta\cdot\pi_{\theta_t}(a)\cdot r(a)\cdot\big(1-\pi_{\theta_t}(a)\big)^2, && \text{(E.309)}
\end{aligned}
$$

where the first inequality is because $\pi_\theta(a)=\pi_\theta^\top e_a$, where $e_a\in\{0,1\}^K$ with $e_a(a)=1$ and $e_a(a')=0$ for all $a'\neq a$, and the second last equality is because of

$$
\frac{d\pi_{\theta_t}(a)}{d\theta_t(i)}=\begin{cases}\pi_{\theta_t}(i)\cdot\big(1-\pi_{\theta_t}(i)\big), & \text{if } i=a,\\ -\pi_{\theta_t}(i)\cdot\pi_{\theta_t}(a). & \text{otherwise}\end{cases} \tag{E.310}
$$

Using $\eta=\frac{\pi_{\theta_t}(a)}{5\cdot r(a)}$, for all $t\geq 1$, we have,

$$
\big(1-\pi_{\theta_{t+1}}(a)\big)-\big(1-\pi_{\theta_t}(a)\big)\leq-\frac{1}{10}\cdot\pi_{\theta_t}(a)^2\cdot\big(1-\pi_{\theta_t}(a)\big)^2, \tag{E.311}
$$

which means $\pi_{\theta_{t+1}}(a) \geq \pi_{\theta_t}(a)$ for all $t \geq 1$. Therefore, we have $\eta \geq \frac{\pi_{\theta_1}(a)}{5 \cdot r(a)} \in \Theta(1)$ and,

$$\big(1 - \pi_{\theta_{t+1}}(a)\big) - \big(1 - \pi_{\theta_t}(a)\big) \leq -\frac{1}{10} \cdot \pi_{\theta_1}(a)^2 \cdot \big(1 - \pi_{\theta_t}(a)\big)^2 . \tag{E.312}$$

Then we have,

$$\begin{aligned}
\frac{1}{1 - \pi_{\theta_t}(a)} &= \frac{1}{1 - \pi_{\theta_1}(a)} + \sum_{s=1}^{t-1} \left[ \frac{1}{1 - \pi_{\theta_{s+1}}(a)} - \frac{1}{1 - \pi_{\theta_s}(a)} \right] && \text{(E.313)} \\
&= \frac{1}{1 - \pi_{\theta_1}(a)} + \sum_{s=1}^{t-1} \frac{\big(1 - \pi_{\theta_s}(a)\big) - \big(1 - \pi_{\theta_{s+1}}(a)\big)}{\big(1 - \pi_{\theta_{s+1}}(a)\big) \cdot \big(1 - \pi_{\theta_s}(a)\big)} && \text{(E.314)} \\
&\geq \frac{1}{1 - \pi_{\theta_1}(a)} + \sum_{s=1}^{t-1} \frac{\big(1 - \pi_{\theta_s}(a)\big)^2}{\big(1 - \pi_{\theta_{s+1}}(a)\big) \cdot \big(1 - \pi_{\theta_s}(a)\big)} \cdot \frac{\pi_{\theta_1}(a)^2}{10} && \text{(E.315)} \\
&\quad \text{(by Eq. (E.312))} && \text{(E.316)} \\
&\geq \frac{1}{1 - \pi_{\theta_1}(a)} + \frac{\pi_{\theta_1}(a)^2}{10} \cdot (t-1) \qquad \big(\pi_{\theta_{t+1}}(a) \geq \pi_{\theta_t}(a)\big) && \text{(E.317)} \\
&\geq \frac{\pi_{\theta_1}(a)^2}{10} \cdot t, \qquad \left( \frac{1}{1 - \pi_{\theta_1}(a)} \geq 1 \geq \frac{\pi_{\theta_1}(a)^2}{10} \right) && \text{(E.318)}
\end{aligned}$$

which implies for all $t \geq 1$,

$$\begin{aligned}
t \cdot \big[1 - \pi_{\theta_t}(a)\big] &\leq t \cdot \left[ \frac{10}{\pi_{\theta_1}(a)^2} \cdot \frac{1}{t} \right] \qquad \text{(by Eq. (E.313))} && \text{(E.319)} \\
&= \frac{10}{\pi_{\theta_1}(a)^2}, && \text{(E.320)}
\end{aligned}$$

which means $\kappa(\mathrm{PG}, a) \geq 1$ for all $a \in [K]$ according to Definition 11.

**Second part.** $\kappa(\mathrm{PG}, a) \leq 1$.

Let action $a$ be sampled forever after initialization. We show that $1 - \pi_{\theta_t}(a)$

cannot decrease faster than $O(1/t)$. Similar to Eq. (E.296), we have,

$$
\begin{aligned}
&(1-\pi_{\theta_t}(a)) - \big(1-\pi_{\theta_{t+1}}(a)\big) && \text{(E.321)}\\
&\quad = \pi_{\theta_{t+1}}(a) - \pi_{\theta_t}(a) - \left\langle \frac{d\pi_{\theta_t}(a)}{d\theta_t}, \theta_{t+1}-\theta_t \right\rangle && \text{(E.322)}\\
&\qquad + \left\langle \frac{d\pi_{\theta_t}(a)}{d\theta_t}, \theta_{t+1}-\theta_t \right\rangle && \text{(E.323)}\\
&\quad \le \frac{5}{4}\cdot \|\theta_{t+1}-\theta_t\|_2^2 + \left\langle \frac{d\pi_{\theta_t}(a)}{d\theta_t}, \theta_{t+1}-\theta_t \right\rangle \qquad \text{(by Lemma 2)} && \text{(E.324)}\\
&\quad = \frac{5\cdot\eta^2}{4}\cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2 + \eta \cdot \left\langle \frac{d\pi_{\theta_t}(a)}{d\theta_t}, \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\rangle && \text{(E.325)}\\
&\qquad \text{(using Update 6)} && \text{(E.326)}\\
&\quad = \frac{5\cdot\eta^2}{2}\cdot (1-\pi_{\theta_t}(a))^2 \cdot r(a)^2 + \eta \cdot \left\langle \frac{d\pi_{\theta_t}(a)}{d\theta_t}, \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\rangle && \text{(E.327)}\\
&\qquad \text{(by Eqs. (E.225) and (E.228))} && \text{(E.328)}\\
&\quad = \frac{5\cdot\eta^2}{2}\cdot (1-\pi_{\theta_t}(a))^2 \cdot r(a)^2 && \text{(E.329)}\\
&\qquad + \eta\cdot\pi_{\theta_t}(a)\cdot r(a)\cdot \left( \sum_{a'\neq a} \pi_{\theta_t}(a')^2 + (1-\pi_{\theta_t}(a))^2 \right) && \text{(E.330)}\\
&\quad \le \frac{5\cdot\eta^2}{2}\cdot (1-\pi_{\theta_t}(a))^2 \cdot r(a)^2 && \text{(E.331)}\\
&\qquad + 2\cdot\eta\cdot\pi_{\theta_t}(a)\cdot r(a)\cdot (1-\pi_{\theta_t}(a))^2 \qquad (\|x\|_2 \le \|x\|_1) && \text{(E.332)}\\
&\quad \le \frac{9}{2}\cdot (1-\pi_{\theta_t}(a))^2 \cdot r(a), && \text{(E.333)}
\end{aligned}
$$

where the last inequality is due to $\pi_{\theta_t}(a) \in (0,1)$, $r(a) \in (0,1]$, and $\eta \in (0,1]$. Denote $\delta(\theta_t) \coloneqq 1-\pi_{\theta_t}(a)$. We have, for all $t \ge 1$,

$$
\delta(\theta_t) - \delta(\theta_{t+1}) \le \frac{9}{2}\cdot r(a)\cdot \delta(\theta_t)^2, \tag{E.334}
$$

which is similar to Eq. (B.730). Therefore, using similar calculations in the proofs for Theorem 11, we have, for all large enough $t \ge 1$,

$$
\begin{aligned}
t\cdot[1-\pi_{\theta_t}(a)] &\ge t\cdot\left[\frac{1}{6\cdot r(a)}\cdot\frac{1}{t}\right] && \text{(E.335)}\\
&= \frac{1}{6\cdot r(a)}, && \text{(E.336)}
\end{aligned}
$$

which means $\kappa(\mathrm{PG}, a) \le 1$ for all $a \in [K]$ according to Definition 11. □

**Theorem 28.** Using Update 9, $(\pi^* - \pi_{\theta_t})^\top r \to 0$ as $t \to \infty$ with probability 1.

*Proof.* Consider the sequence $\{\pi_{\theta_t}(a^*)\}_{t\geq 1}$ produced by Update 9 using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$. We show that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$ with probability 1.

First, for convenience, we duplicate Update 9 here.

**Update 9** (NPG with oracle baseline)**.** $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \big(\hat{r}_t - \hat{b}_t\big)$, where $\hat{b}_t(a) = \left(\frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)} - 1\right) \cdot b$ for all $a \in [K]$, and $b \in (r(a^*) - \Delta, r(a^*))$.

Note that Update 9 is equivalent to the following update,

$$\theta_{t+1}(a) = \begin{cases} \theta_t(a) + \frac{\eta}{\pi_{\theta_t}(a)} \cdot \big(r(a) - b\big), & \text{if } a = a_t, \\ \theta_t(a), & \text{otherwise} \end{cases} \tag{E.337}$$

Next, we show that $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$ using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$. There are two cases.

Case (a): If $a_t = a^*$, then we have,

$$\begin{aligned} \theta_{t+1}(a^*) &= \theta_t(a^*) + \frac{\eta}{\pi_{\theta_t}(a^*)} \cdot (r(a^*) - b) \qquad \text{(by Eq. (E.337))} && \text{(E.338)} \\ &> \theta_t(a^*), \qquad (r(a^*) > b) && \text{(E.339)} \end{aligned}$$

while $\theta_{t+1}(a) = \theta_t(a)$ for all sub-optimal actions $a \neq a^*$. Then we have,

$$\begin{aligned} \pi_{\theta_{t+1}}(a^*) &= \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_{t+1}(a^*)\} + \sum_{a\neq a^*} \exp\{\theta_{t+1}(a)\}} && \text{(E.340)} \\ &> \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a^*)\} + \sum_{a\neq a^*} \exp\{\theta_{t+1}(a)\}} \qquad \text{(by Eq. (E.338))} && \text{(E.341)} \\ &= \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a^*)\} + \sum_{a\neq a^*} \exp\{\theta_t(a)\}} && \text{(E.342)} \\ &\qquad (\theta_{t+1}(a) = \theta_t(a), \text{ for all } a \neq a^*) && \text{(E.343)} \\ &= \pi_{\theta_t}(a^*). && \text{(E.344)} \end{aligned}$$

Case (b): If $a_t = a \neq a^*$, then we have,

$$\begin{aligned} \theta_{t+1}(a) &= \theta_t(a) + \frac{\eta}{\pi_{\theta_t}(a)} \cdot (r(a) - b) \qquad \text{(by Eq. (E.337))} && \text{(E.345)} \\ &< \theta_t(a), \qquad (r(a) \leq r(a^*) - \Delta < b) && \text{(E.346)} \end{aligned}$$

where $\Delta = r(a^*) - \max_{a\neq a^*} r(a) > 0$ is the reward gap. Also $\theta_{t+1}(a') = \theta_t(a')$

for all the other actions $a' \neq a$. Then we have,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_{t+1}(a)\} + \sum_{a' \neq a} \exp\{\theta_{t+1}(a')\}} \tag{E.347}$$

$$> \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_t(a)\} + \sum_{a' \neq a} \exp\{\theta_{t+1}(a')\}} \qquad \text{(by Eq. (E.345))} \tag{E.348}$$

$$= \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a)\} + \sum_{a' \neq a} \exp\{\theta_t(a')\}} \tag{E.349}$$

$$(\theta_{t+1}(a') = \theta_t(a'), \text{ for all } a' \neq a) \tag{E.350}$$

$$= \pi_{\theta_t}(a^*). \tag{E.351}$$

Therefore, we have $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$, for all $t \geq 1$. Note that $\pi_{\theta_t}(a^*) \leq 1$. According to monotone convergence theorem, we have $\pi_{\theta_{t+1}}(a^*)$ approaches to some finite value as $t \to \infty$.

Suppose $\pi_{\theta_t}(a^*) \to \pi_{\theta_\infty}(a^*)$ as $t \to \infty$. We show that $\pi_{\theta_\infty}(a^*) = 1$ by contradiction. Suppose $\pi_{\theta_\infty}(a^*) < 1$. Then at the convergent point, according to Eqs. (E.340) and (E.347), we can further improve the probability of $a^*$ by online sampling and updating once, which is a contradiction with convergence.

Thus we have $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$ with probability 1, which implies that $(\pi^* - \pi_{\theta_t})^\top r \to 0$ as $t \to \infty$ with probability 1. □

The Stochastic Approximation Markov Bandit Algorithm (SAMBA) (Denisov and Walton, 2020) algorithm is mentioned in Section 5.4 and Fig. 5.1.

**Update 10** (SAMBA)**.** *At iteration $t \geq 1$, denote the greedy action $\bar{a}_t \coloneqq \arg\max_{a \in [K]} \pi_t(a)$. Sample action $a_t \sim \pi_t(\cdot)$. (i) If $a_t = \bar{a}_t$, then perform update $\pi_{t+1}(a') \leftarrow \pi_t(a') - \eta \cdot \pi_t(a')^2 \cdot \frac{r(a_t)}{\pi_t(a_t)}$ for all non-greedy action $a' \neq a_t$; (ii) If $a_t \neq \bar{a}_t$, then perform update $\pi_{t+1}(a_t) \leftarrow \pi_t(a_t) + \eta \cdot \pi_t(a_t)^2 \cdot \frac{r(a_t)}{\pi_t(a_t)}$. After doing (i) or (ii), calculate $\pi_{t+1}(\bar{a}_t) = 1 - \sum_{a' \neq \bar{a}_t} \pi_{t+1}(a')$.*

The SAMBA algorithm does not maintain parameters $\theta$, and the last step $\pi_{t+1}(\bar{a}_t) = 1 - \sum_{a' \neq \bar{a}_t} \pi_{t+1}(a')$ in Update 10 is a necessary projection to the probability simplex, such that $\pi_t$ is a valid probability distribution over $[K]$. As shown in (Denisov and Walton, 2020), if the learning rate has the knowledge of the optimal action's reward and reward gap, i.e.,

$$\eta < \frac{\Delta}{r(a^*) - \Delta}, \tag{E.352}$$

then Update 10 converges to $\pi^*$ almost surely with a $O(1/t)$ rate, i.e.,

$$(\pi^* - \pi_t)^\top r \leq C/t. \tag{E.353}$$

We calculate the committal rate of SAMBA.

**Proposition 14.** *For SAMBA from Update 10, we have $\kappa(SAMBA, a) = 1$ for all $a \in [K]$.*

*Proof.* **First part.** $\kappa(\text{SAMBA}, a) \geq 1$.

According to Definition 11, let action $a$ be the greedy action and be sampled forever. According to *(i)* in Update 10, we have, for all $a' \neq a$,

$$\pi_{t+1}(a') = \pi_t(a') - \eta \cdot \pi_t(a')^2 \cdot \frac{r(a_t)}{\pi_t(a_t)} \tag{E.354}$$

$$= \pi_t(a') - \eta \cdot \pi_t(a')^2 \cdot \frac{r(a)}{\pi_t(a)} \qquad (a_t = a \text{ by fixed sampling}) \tag{E.355}$$

$$\leq \pi_t(a') - \eta \cdot \pi_t(a')^2 \cdot r(a). \qquad (\pi_t(a) \in (0,1)) \tag{E.356}$$

Using similar calculations in Eq. (B.46), we have, for all $a' \neq a$,

$$\frac{1}{\pi_t(a')} = \frac{1}{\pi_1(a')} + \sum_{s=1}^{t-1} \left[ \frac{1}{\pi_{s+1}(a')} - \frac{1}{\pi_s(a')} \right] \tag{E.357}$$

$$= \frac{1}{\pi_1(a')} + \sum_{s=1}^{t-1} \frac{1}{\pi_{s+1}(a') \cdot \pi_s(a')} \cdot (\pi_s(a')) - \pi_{s+1}(a')) \tag{E.358}$$

$$\geq \frac{1}{\pi_1(a')} + \sum_{s=1}^{t-1} \frac{1}{\pi_{s+1}(a') \cdot \pi_s(a')} \cdot \eta \cdot \pi_s(a')^2 \cdot r(a) \tag{E.359}$$

$$\text{(by Eq. (E.354))} \tag{E.360}$$

$$\geq \frac{1}{\pi_1(a')} + \eta \cdot r(a) \cdot (t-1) \tag{E.361}$$

$$(\pi_{t+1}(a') \leq \pi_t(a'), \text{ by Eq. (E.354)}) \tag{E.362}$$

$$\geq \eta \cdot r(a) \cdot t, \qquad \left( \frac{1}{\pi_1(a')} \geq 1 \geq \eta \cdot r(a) \right) \tag{E.363}$$

which implies, for all large enough $t \geq 1$,

$$t \cdot [1 - \pi_t(a)] = t \cdot \sum_{a' \neq a} \pi_t(a') \tag{E.364}$$

$$\leq t \cdot \sum_{a' \neq a} \frac{1}{\eta \cdot r(a) \cdot t} \qquad \text{(by Eq. (E.357))} \tag{E.365}$$

$$= \sum_{a' \neq a} \frac{1}{\eta \cdot r(a)}, \tag{E.366}$$

which means $\kappa(\text{SAMBA}, a) \geq 1$ for all $a \in [K]$ according to Definition 11.

**Second part.** $\kappa(\text{SAMBA}, a) \leq 1$.

Let action $a$ be the greedy action and be sampled forever. According to *(i)* in Update 10, we have, for all $a' \neq a$,

$$\pi_{t+1}(a') = \pi_t(a') - \eta \cdot \pi_t(a')^2 \cdot \frac{r(a_t)}{\pi_t(a_t)} \tag{E.367}$$

$$= \pi_t(a') - \eta \cdot \pi_t(a')^2 \cdot \frac{r(a)}{\pi_t(a)} \qquad (a_t = a \text{ by fixed sampling}) \tag{E.368}$$

$$\geq \pi_t(a') - \eta \cdot K \cdot \pi_t(a')^2 \cdot r(a), \tag{E.369}$$

$$(\pi_t(a) \geq 1/K, \ a \text{ is greedy action}) \tag{E.370}$$

which is similar to Eq. (B.730). Therefore, using similar calculations in the proofs for Theorem 9, we have, for all large enough $t \geq 1$, we have,

$$\frac{\pi_{t+1}(a')}{\pi_t(a')} \geq \frac{1}{2}. \tag{E.371}$$

Denote

$$t_0 \coloneqq \min\left\{t \geq 1 : \frac{\pi_{t+1}(a')}{\pi_t(a')} \geq \frac{1}{2}, \text{ for all } s \geq t\right\}. \tag{E.372}$$

On the other hand, since $t_0 \in O(1)$, we have, for all $t < t_0$,

$$\pi_{t+1}(a') \geq c_0 > 0. \tag{E.373}$$

Next, we have, for all $t \geq t_0$,

$$\frac{1}{\pi_t(a')} = \frac{1}{\pi_1(a')} + \sum_{s=1}^{t_0-1} \frac{1}{\pi_{s+1}(a')} \cdot \left(1 - \frac{\pi_{s+1}(a')}{\pi_s(a')}\right) \tag{E.374}$$

$$+ \sum_{s=t_0}^{t-1} \frac{1}{\pi_{s+1}(a') \cdot \pi_s(a')} \cdot (\pi_s(a')) - \pi_{s+1}(a')) \tag{E.375}$$

$$\leq \frac{1}{c_0} + \sum_{s=1}^{t_0-1} \frac{1}{c_0} \cdot 1 + \sum_{s=t_0}^{t-1} \frac{1}{\pi_{s+1}(a') \cdot \pi_s(a')} \cdot \eta \cdot K \cdot \pi_s(a')^2 \cdot r(a) \tag{E.376}$$

$$\text{(by Eqs. (E.367) and (E.373))} \tag{E.377}$$

$$\leq \frac{t_0}{c_0} + 2 \cdot \eta \cdot K \cdot r(a) \cdot (t - t_0), \qquad \text{(by Eq. (E.371))} \tag{E.378}$$

which implies, for all large enough $t \geq 1$,

$$t \cdot [1 - \pi_t(a)] = t \cdot \sum_{a' \neq a} \pi_t(a') \tag{E.379}$$

$$\geq t \cdot \sum_{a' \neq a} \frac{1}{t_0/c_0 + 2 \cdot \eta \cdot K \cdot r(a) \cdot (t - t_0)} \qquad \text{(by Eq. (E.374))} \tag{E.380}$$

$$\geq \sum_{a' \neq a} \frac{1}{3 \cdot \eta \cdot K \cdot r(a)}, \qquad (t_0/c_0 \leq \eta \cdot K \cdot r(a) \cdot t) \tag{E.381}$$

which means $\kappa(\text{SAMBA}, a) \leq 1$ for all $a \in [K]$ according to Definition 11. □

## E.3 Proofs for Section 5.4: Geometry-Convergence Trade-off

First, we show that the algorithms we study in this paper, i.e., softmax PG, NPG, and GNPG, are optimality-smart. Recall from the main paper that, a policy optimization method is said to be *optimality-smart* if for any $t \geq 1$, $\pi_{\tilde{\theta}_t}(a^*) \geq \pi_{\theta_t}(a^*)$ holds where $\tilde{\theta}_t$ is the parameter vector obtained when $a^*$ is chosen in every time step, starting at $\theta_1$, while $\theta_t$ is *any* parameter vector that can be obtained with $t$ updates (regardless of the action sequence chosen), but also starting from $\theta_1$.

**Proposition 15.** *Softmax PG, NPG, and GNPG are optimality-smart.*

*Proof.* We show that for softmax PG, NPG, and GNPG, if $a_t = a^*$, then $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$; if $a_t = a \neq a^*$, then $\pi_{\theta_{t+1}}(a^*) \leq \pi_{\theta_t}(a^*)$ (for softmax PG and GNPG the later claim holds when $\pi_{\theta_t}(a^*)$ is the dominating action, i.e., $\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a')$ for all $a' \neq a^*$).

**First part.** Softmax PG and GNPG are optimality-smart.

If $a_t = a^*$, then we have,

$$\theta_{t+1}(a^*) = \theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a^*)} \tag{E.382}$$

$$= \theta_t(a^*) + \eta \cdot (1 - \pi_{\theta_t}(a^*)) \cdot r(a^*) \qquad \text{(by Eq. (E.225))} \tag{E.383}$$

$$\geq \theta_t(a^*). \qquad \left(r \in (0, 1]^K\right) \tag{E.384}$$

And for any $a \neq a^*$, we have,

$$\theta_{t+1}(a) = \theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)} \tag{E.385}$$

$$= \theta_t(a) - \eta \cdot \pi_{\theta_t}(a) \cdot r(a^*) \qquad \text{(by Eq. (E.228))} \tag{E.386}$$

$$\leq \theta_t(a). \qquad \left(r \in (0,1]^K\right) \tag{E.387}$$

Therefore, we have,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_{t+1}(a^*)\} + \sum_{a \neq a^*} \exp\{\theta_{t+1}(a)\}} \tag{E.388}$$

$$\geq \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a^*)\} + \sum_{a \neq a^*} \exp\{\theta_t(a)\}} \tag{E.389}$$

$$\text{(by Eqs. (E.382) and (E.385))} \tag{E.390}$$

$$= \pi_{\theta_t}(a^*). \tag{E.391}$$

On the other hand, given $a_t = a \neq a^*$, we show that if $\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a')$ for all $a' \neq a^*$, then $\pi_{\theta_{t+1}}(a^*) \leq \pi_{\theta_t}(a^*)$. We have,

$$\theta_{t+1}(a) = \theta_t(a) + \eta \cdot (1 - \pi_{\theta_t}(a)) \cdot r(a) \qquad \text{(by Eq. (E.225))} \tag{E.392}$$

$$\geq \theta_t(a) - \eta \cdot \pi_{\theta_t}(a^*) \cdot r(a). \tag{E.393}$$

And for any $a' \neq a$, we have,

$$\theta_{t+1}(a') = \theta_t(a') - \eta \cdot \pi_{\theta_t}(a') \cdot r(a) \qquad \text{(by Eq. (E.228))} \tag{E.394}$$

$$\geq \theta_t(a') - \eta \cdot \pi_{\theta_t}(a^*) \cdot r(a). \qquad (\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a')) \tag{E.395}$$

Therefore, we have,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_{t+1}(a)\} + \sum_{a' \neq a} \exp\{\theta_{t+1}(a')\}} \tag{E.396}$$

$$\leq \frac{\exp\{\theta_t(a^*) - \eta \cdot \pi_{\theta_t}(a^*) \cdot r(a)\}}{\exp\{\theta_t(a) - \eta \cdot \pi_{\theta_t}(a^*) \cdot r(a)\} + \sum_{a' \neq a} \exp\{\theta_t(a') - \eta \cdot \pi_{\theta_t}(a^*) \cdot r(a)\}} \tag{E.397}$$

$$\text{(by Eqs. (E.392) and (E.394))} \tag{E.398}$$

$$= \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a)\} + \sum_{a' \neq a} \exp\{\theta_t(a')\}} \tag{E.399}$$

$$= \pi_{\theta_t}(a^*). \tag{E.400}$$

**Second part.** NPG is optimality-smart.

If $a_t = a^*$, then we have,

$$\theta_{t+1}(a^*) = \theta_t(a^*) + \eta \cdot \frac{r(a^*)}{\pi_{\theta_t}(a^*)} \tag{E.401}$$

$$> \theta_t(a^*). \tag{E.402}$$

while $\theta_{t+1}(a) = \theta_t(a)$ for all sub-optimal actions $a \neq a^*$. Then we have,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_{t+1}(a^*)\} + \sum_{a \neq a^*} \exp\{\theta_{t+1}(a)\}} \tag{E.403}$$

$$\geq \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a^*)\} + \sum_{a \neq a^*} \exp\{\theta_t(a)\}} \tag{E.404}$$

$$= \pi_{\theta_t}(a^*). \tag{E.405}$$

If $a_t = a \neq a^*$, then we have,

$$\theta_{t+1}(a) = \theta_t(a) + \eta \cdot \frac{r(a)}{\pi_{\theta_t}(a)} \tag{E.406}$$

$$\geq \theta_t(a), \tag{E.407}$$

while $\theta_{t+1}(a') = \theta_t(a')$ for all the other actions $a' \neq a$. Then we have,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\{\theta_{t+1}(a^*)\}}{\exp\{\theta_{t+1}(a)\} + \sum_{a' \neq a} \exp\{\theta_{t+1}(a')\}} \tag{E.408}$$

$$\leq \frac{\exp\{\theta_t(a^*)\}}{\exp\{\theta_t(a)\} + \sum_{a' \neq a} \exp\{\theta_t(a')\}} \tag{E.409}$$

$$= \pi_{\theta_t}(a^*). \qquad \square$$

**Theorem 29.** Let $\mathcal{A}$ be optimality-smart and pick a bandit instance. If $\mathcal{A}$ together with on-policy sampling leads to $\{\theta_t\}_{t \geq 1}$ such that $\{\pi_{\theta_t}\}_{t \geq 1}$ converges to a globally optimal policy at a rate $O(1/t^\alpha)$ with positive probability, for $\alpha > 0$, then $\kappa(\mathcal{A}, a^*) \geq \alpha$.

*Proof.* Fix an instance $r \in (0, 1]^K$ with a unique optimal action $a^*$. For any $\theta \in \mathbb{R}^K$, we have,

$$(\pi^* - \pi_\theta)^\top r = \sum_{a \neq a^*} \pi_\theta(a) \cdot (r(a^*) - r(a)) \tag{E.410}$$

$$\geq (1 - \pi_\theta(a^*)) \cdot \Delta, \tag{E.411}$$

where $\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0$ is the reward gap. Let $\{\theta_t\}_{t \geq 1}$ be the sequence obtained by using $\mathcal{A}$ together with online sampling on $r$. For $\alpha > 0$ let $\mathcal{E}_\alpha$ be the event when for all $t \geq 1$,

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{C}{t^\alpha}, \tag{E.412}$$

By our assumption, there exists $\alpha > 0$ such that $\Pr(\mathcal{E}_\alpha) > 0$. On this event, for any $t \geq 1$,

$$t^\alpha \cdot (1 - \pi_{\theta_t}(a^*)) \leq \frac{1}{\Delta} \cdot t^\alpha \cdot (\pi^* - \pi_{\theta_t})^\top r \qquad \text{(by Eq. (E.410))} \tag{E.413}$$

$$\leq \frac{C}{\Delta}. \qquad \text{(by Eq. (E.412))} \tag{E.414}$$

Let $\left\{\tilde{\theta}_t\right\}_{t \geq 1}$ with $\tilde{\theta}_1 = \theta_1$ be the sequence obtained by using $\mathcal{A}$ with fixed sampling on $r$, such that $a_t = a^*$ for all $t \geq 1$. Since, by the assumption, $\mathcal{A}$ is optimality-smart, we have $\pi_{\tilde{\theta}_t}(a^*) \geq \pi_{\theta_t}(a^*)$. Then, on $\mathcal{E}_\alpha$, for any $t \geq 1$

$$t^\alpha \cdot \left(1 - \pi_{\tilde{\theta}_t}(a^*)\right) \leq t^\alpha \cdot (1 - \pi_{\theta_t}(a^*)) \tag{E.415}$$

$$\leq \frac{C}{\Delta}, \qquad \text{(by Eq. (E.413))} \ . \tag{E.416}$$

Since $\mathbb{P}(\mathcal{E}_\alpha) > 0$ and $t^\alpha \cdot \left(1 - \pi_{\tilde{\theta}_t}(a^*)\right)$ is non-random, it follows that for any $t \geq 1$, $t^\alpha \cdot \left(1 - \pi_{\tilde{\theta}_t}(a^*)\right) \leq C/\Delta$, which, by Definition 11, means that $\kappa(\mathcal{A}, a^*) \geq \alpha$. □

**Theorem 30** (Geometry-Convergence trade-off)**.** If an algorithm $\mathcal{A}$ is optimality-smart, and $\kappa(\mathcal{A}, a^*) = \kappa(\mathcal{A}, a)$ for at least one $a \neq a^*$, then $\mathcal{A}$ with on-policy sampling can only exhibit at most one of the following two behaviors: **(i)** $\mathcal{A}$ converges to a globally optimal policy almost surely; **(ii)** $\mathcal{A}$ converges to a deterministic policy at a rate faster than $O(1/t)$ with positive probability.

*Proof.* We prove that $\mathcal{A}$ cannot achieve both of the two behaviors at the same time by contradiction. Suppose an algorithm $\mathcal{A}$ can **(i)** converge to a globally optimal policy almost surely; and **(ii)** converges at a rate $O(1/t^\alpha)$ with positive probability, where $\alpha > 1$.

Since **(ii)** holds, according to Theorem 29, we have $\kappa(\mathcal{A}, a^*) \geq \alpha > 1$. By condition, there exists at least one sub-optimal action $a \neq a^*$, such that

$\kappa(\mathcal{A}, a) = \kappa(\mathcal{A}, a^*) > 1$. According to Theorem 25, we have $\pi_{\theta_t}(a) \to 1$ as $t \to \infty$ with positive probability, which contradicts **(i)**. Therefore, **(i)** and **(ii)** cannot hold simultaneously. □

## E.4 Proofs for Section 5.5: Ensemble Methods

**Theorem 31.** With probability $1-\delta$, the best single run among $O(\log(1/\delta))$ independent runs of NPG (GNPG) converges to a globally optimal policy at an $O(e^{-c\cdot t})$ rate.

*Proof.* According to Theorem 23, stochastic NPG of Update 7 will sample the optimal action $a^*$ forever (thus converge to the optimal policy) with probability at least

$$
\begin{aligned}
& p(\mathrm{NPG}, a^*) && \text{(E.417)}\\
& \coloneqq \exp\left\{ -\frac{\exp\{\eta \cdot r(a^*)\}}{\eta \cdot r(a^*)} \cdot \frac{\sum_{a\neq a^*} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a^*)\}} \right\} && \text{(E.418)}\\
& \quad \text{(by Eq. (E.208))} && \text{(E.419)}\\
& \in \Omega(1). && \text{(E.420)}
\end{aligned}
$$

Moreover, with probability at least $p(\mathrm{NPG}, a^*)$, the convergence rate is,

$$
\begin{aligned}
& (\pi^* - \pi_{\theta_t})^\top r = \sum_{a\neq a^*} \pi_{\theta_t}(a) \cdot (r(a^*) - r(a)) && \text{(E.421)}\\
& \leq 1 - \pi_{\theta_t}(a^*) \qquad \left(r \in (0,1]^K\right) && \text{(E.422)}\\
& \leq \frac{\sum_{a\neq a^*} \exp\{\theta_1(a)\}}{\exp\{\theta_1(a^*) + \eta \cdot r(a^*) \cdot (t-1)\} + \sum_{a\neq a^*} \exp\{\theta_1(a)\}} && \text{(E.423)}\\
& \quad \text{(by Eq. (E.286))} && \text{(E.424)}\\
& \in O(e^{-c\cdot t}). && \text{(E.425)}
\end{aligned}
$$

Consider $n(\mathrm{NPG}) \in O(\log(1/\delta))$ independent runs of NPG, where

$$
n(\mathrm{NPG}) \coloneqq \frac{1}{\log\left(\frac{1}{1-p(\mathrm{NPG},a^*)}\right)} \cdot \log(1/\delta). \tag{E.426}
$$

The probability that all the $n(\text{NPG})$ runs do not converge to global optimal policy is at most

$$[1-p(\text{NPG},a^*)]^{n(\text{NPG})} = \left[\exp\left\{\log\left(1-p(\text{NPG},a^*)\right)\right\}\right]^{n(\text{NPG})} \tag{E.427}$$

$$= \exp\left\{-\log\left(\frac{1}{1-p(\text{NPG},a^*)}\right)\cdot\frac{1}{\log\left(\frac{1}{1-p(\text{NPG},a^*)}\right)}\cdot\log\left(1/\delta\right)\right\} \tag{E.428}$$

$$\text{(by Eq. (E.426))} \tag{E.429}$$

$$= e^{-\log(1/\delta)} = \delta, \tag{E.430}$$

which means with probability at least $1-\delta$, the best single run converges to a globally optimal policy at an $O(e^{-c\cdot t})$ rate.

For stochastic GNPG of Update 8, similar calculations show that with probability at least $1-\delta$, the best single run among $n(\text{GNPG}) \in O(\log(1/\delta))$ independent runs of GNPG converges to a globally optimal policy at an $O(e^{-c\cdot t})$ rate, where

$$n(\text{GNPG}) \coloneqq \frac{1}{\log\left(\frac{1}{1-p(\text{GNPG},a^*)}\right)}\cdot\log\left(1/\delta\right), \tag{E.431}$$

and

$$p(\text{GNPG},a^*) \tag{E.432}$$

$$\coloneqq \exp\left\{-\frac{\sum_{a\neq a^*}\exp\{\theta_1(a)\}}{\exp\{\theta_1(a^*)\}}\cdot\frac{\sqrt{2}\cdot\exp\left\{\frac{\eta}{\sqrt{2}}\right\}}{\eta}\right\} \tag{E.433}$$

$$\text{(by Eq. (E.252))} \tag{E.434}$$

$$\in \Omega(1), \tag{E.435}$$

thus finishing the proof. □

## E.5 Miscellaneous Extra Supporting Results

**Lemma 58.** *We have, for all $x \in (0,1)$,*

$$1-x \geq e^{-1/(1/x-1)}. \tag{E.436}$$

*Proof.* See the proof in (Chung et al., 2020, Proposition 1). We include a proof for completeness.

We have, for all $x \in (0,1)$,

$$1 - x = \exp\{\log(1-x)\} \tag{E.437}$$

$$\geq \exp\left\{1 - e^{-\log(1-x)}\right\} \qquad \left(y \geq 1 - e^{-y}\right) \tag{E.438}$$

$$= \exp\left\{\frac{-1}{1/x - 1}\right\}. \tag{E.439}$$

□

**Lemma 59.** *Let $\alpha > 0$. We have,*

**(i)** *if $\alpha \in (1, \infty)$, then for all $C > 0$,*

$$\sum_{t=1}^{\infty} \frac{C}{t^{\alpha}} < \infty, \tag{E.440}$$

*which means the series $\sum_{t=1}^{\infty} \frac{C}{t^{\alpha}}$ converges to a finite value.*

**(ii)** *if $\alpha \in (0, 1]$, then for all $C > 0$,*

$$\sum_{t=1}^{\infty} \frac{C}{t^{\alpha}} = \infty, \tag{E.441}$$

*which means the series $\sum_{t=1}^{\infty} \frac{C}{t^{\alpha}}$ diverges to positive infinity.*

**(iii)** *for all $C > 0$, $C' > 0$,*

$$\sum_{t=1}^{\infty} \frac{C}{\exp\{C' \cdot t\}} < \infty, \tag{E.442}$$

*which means the series $\sum_{t=1}^{\infty} \frac{C}{\exp\{C' \cdot t\}}$ converges to a finite value.*

*Proof.* It is easy to verify the results by calculating integrals. We include a proof for completeness.

**First part.** We have, for all $\alpha \in (1, \infty)$ and $C > 0$,

$$\sum_{t=1}^{\infty} \frac{C}{t^{\alpha}} \leq C \cdot \left(1 + \int_{t=1}^{\infty} \frac{1}{t^{\alpha}} dt\right) \tag{E.443}$$

$$= \frac{C \cdot \alpha}{\alpha - 1}. \tag{E.444}$$

**Second part.** We have, for all $\alpha \in (0, 1)$, $C > 0$, and $T \geq 1$,

$$\sum_{t=1}^{T} \frac{C}{t^{\alpha}} \geq \int_{t=1}^{T+1} \frac{C}{t^{\alpha}} dt \tag{E.445}$$

$$= \frac{C \cdot ((T+1)^{1-\alpha} - 1)}{1 - \alpha}. \tag{E.446}$$

Similarly, for $\alpha = 1$,

$$\sum_{t=1}^{T} \frac{C}{t} \geq \int_{t=1}^{T+1} \frac{C}{t} dt \tag{E.446}$$

$$= C \cdot \log(T+1). \tag{E.447}$$

Therefore, the partial sum approaches to positive infinity as $T \to \infty$.

**Third part.** We have, for all $C > 0$ and $C' > 0$,

$$\sum_{t=1}^{\infty} \frac{C}{\exp\{C' \cdot t\}} \leq \int_{t=0}^{\infty} \frac{C}{\exp\{C' \cdot t\}} \tag{E.448}$$

$$= \frac{C}{C'}. \qquad \square$$

**Lemma 60.** *Let $u_t \in (0,1)$ for all $t \geq 1$. The infinite product $\prod_{t=1}^{\infty}(1-u_t)$ converges to a positive value if and only if the series $\sum_{t=1}^{\infty} u_t$ converges to a finite value.*

*Proof.* See Knopp (1947, p. 220). We include a proof for completeness.

Define the following partial products and partial sums,

$$p_T \coloneqq \prod_{t=1}^{T}(1-u_t), \tag{E.449}$$

$$s_T \coloneqq \sum_{t=1}^{T} u_t. \tag{E.450}$$

Since $p_T$ is monotonically decreasing and non-negative, the infinite product converges to positive values, i.e.,

$$\prod_{t=1}^{\infty}(1-u_t) = \lim_{T\to\infty} \prod_{t=1}^{T}(1-u_t) = \lim_{T\to\infty} p_T > 0, \tag{E.451}$$

if and only if $p_T$ is lower bounded away from zero (boundedness convergence criterion for monotone sequence) (Knopp, 1947, p. 80).

Similarly, since $s_T$ is monotonically increasing, the series converges to finite values, i.e.,

$$\sum_{t=1}^{\infty} u_t = \lim_{T\to\infty} \sum_{t=1}^{T} u_t = \lim_{T\to\infty} s_T < \infty, \tag{E.452}$$

if and only if $s_T$ is upper bounded.

**First part.** $\prod_{t=1}^{\infty}(1-u_t)$ converges to a positive value only if $\sum_{t=1}^{\infty} u_t$ converges to a finite value.

Suppose $\prod_{t=1}^{\infty}(1-u_t)$ converges to a positive value. We have, for all $T \geq 1$,

$$q_T \geq q > 0. \tag{E.453}$$

Then we have,

$$\begin{aligned}
q &\leq q_T && \text{(E.454)}\\
&= \exp\left\{ \log\left( \prod_{t=1}^{T}(1-u_t) \right) \right\} && \text{(E.455)}\\
&= \exp\left\{ \sum_{t=1}^{T} \log(1-u_t) \right\} && \text{(E.456)}\\
&\leq \exp\left\{ -\sum_{t=1}^{T} u_t \right\} \qquad (\log(1-x) < -x) && \text{(E.457)}\\
&= \exp\{-s_T\}, && \text{(E.458)}
\end{aligned}$$

which implies that,

$$s_T \leq -\log q < \infty. \tag{E.459}$$

Therefore, we have $\sum_{t=1}^{\infty} u_t$ converges to a finite value.

**Second part.** $\prod_{t=1}^{\infty}(1-u_t)$ converges to a positive value if $\sum_{t=1}^{\infty} u_t$ converges to a finite value.

Suppose $\sum_{t=1}^{\infty} u_t$ converges to a finite value. Then we have, $u_t \to 0$ as $t \to \infty$. There exists a finite number $t_0 \geq 1$, such that for all $t \geq t_0$, we have $u_t \leq 1/2$. Also, we have, for all $T \geq 1$,

$$s_T \leq s < \infty. \tag{E.460}$$

Then we have,

$$\prod_{t=t_0}^{T}(1-u_t) = \exp\left\{\sum_{t=t_0}^{T}\log(1-u_t)\right\} \tag{E.461}$$

$$\geq \exp\left\{-\sum_{t=t_0}^{T}2\cdot u_t\right\} \tag{E.462}$$

$$(-2\cdot x \leq \log(1-x) \text{ for all } x\in[0,1/2]) \tag{E.463}$$

$$= \exp\{-2\cdot s_T\}, \tag{E.464}$$

which implies that, for all large enough $T \geq 1$,

$$q_T = \left(\prod_{t=1}^{t_0-1}(1-u_t)\right)\cdot\left(\prod_{t=t_0}^{T}(1-u_t)\right) \tag{E.465}$$

$$\geq \left(\prod_{t=1}^{t_0-1}(1-u_t)\right)\cdot\exp\{-2\cdot s_T\} \tag{E.466}$$

$$\geq \left(\prod_{t=1}^{t_0-1}(1-u_t)\right)\cdot\exp\{-2\cdot s\} \tag{E.467}$$

$$> 0. \tag{E.468}$$

Therefore, we have $\prod_{t=1}^{\infty}(1-u_t)$ converges to a positive value. □

**Lemma 61.** *Let $u_t \in (0,1)$ for all $t \geq 1$. We have*

$$\prod_{t=1}^{\infty}(1-u_t) = \lim_{T\to\infty}\prod_{t=1}^{T}(1-u_t) = 0, \tag{E.469}$$

*if and only if the series $\sum_{t=1}^{\infty}u_t$ diverges to a positive infinity.*

*Proof.* **First part.** $\prod_{t=1}^{\infty}(1-u_t)$ diverges to 0 only if $\sum_{t=1}^{\infty}u_t$ diverges to positive infinity.

Suppose $\prod_{t=1}^{\infty}(1-u_t)$ diverges to 0. According to Lemma 60, $\sum_{t=1}^{\infty}u_t$ diverges. And since the partial sum $s_T \coloneqq \sum_{t=1}^{T}u_t$ is monotonically increasing, we have $\sum_{t=1}^{\infty}u_t$ diverges to positive infinity.

**Second part.** $\prod_{t=1}^{\infty}(1-u_t)$ diverges to 0 if $\sum_{t=1}^{\infty}u_t$ diverges to a positive infinity.

Suppose $\sum_{t=1}^{\infty}u_t$ diverges to positive infinity. According to Lemma 60, $\prod_{t=1}^{\infty}(1-u_t)$ diverges. And since the partial product $q_T \coloneqq \prod_{t=1}^{T}(1-u_t)$ is

non-negative and monotonically decreasing, we have $\prod_{t=1}^{\infty}(1-u_t)$ diverges to 0. □

**Lemma 62.** *Let $\pi_{\theta_t}(a) \in (0,1)$ be the probability of sampling action $a$ using online sampling $a_t \sim \pi_{\theta_t}(\cdot)$, for all $t \geq 1$. If $1-\pi_{\theta_t}(a) \in \Theta(1/t^{\alpha})$ with $\alpha \in [0,1]$, then $\prod_{t=1}^{\infty}\pi_{\theta_t}(a) = 0$.*

This lemma means if $\pi_{\theta_t}(a)$ approaches to 1 **slowly**, i.e., no faster than $O(1/t)$, then the probability of sampling $a$ forever using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$ is **zero**, i.e., the other actions $a' \neq a$ always have a chance to be sampled.

*Proof.* Suppose $1-\pi_{\theta_t}(a) \in \Theta(1/t^{\alpha})$ and $\alpha \in (0,1]$. Let $u_t \coloneqq 1-\pi_{\theta_t}(a) \in (0,1)$ for all $t \geq 1$. According to Lemma 59, we have,

$$\sum_{t=1}^{\infty} u_t = \sum_{t=1}^{\infty}(1-\pi_{\theta_t}(a)) = \infty, \tag{E.470}$$

i.e., the series diverges to positive infinity. According to Lemma 61, we have,

$$\prod_{t=1}^{\infty}\pi_{\theta_t}(a) = \prod_{t=1}^{\infty}(1-u_t) = 0, \tag{E.471}$$

which means it is impossible to sample $a$ forever using on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$. □