Convergence of Softmax Policy Gradient: Incorporating Entropy Regularization and Handling Linear Function Approximation

by

Matin Aghaei

B.Sc., Amirkabir University of Technology, 2022

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Spring 2025

Declaration of Committee

Name:	**Matin Aghaei**
Degree:	**Master of Science**
Thesis title:	**Convergence of Softmax Policy Gradient: Incorporating Entropy Regularization and Handling Linear Function Approximation**
Committee:	**Chair:** Manolis Savva Associate Professor, Computing Science
	Sharan Vaswani Supervisor Assistant Professor, Computing Science
	Mo Chen Committee Member Associate Professor, Computing Science
	Hang Ma Examiner Assistant Professor, Computing Science

Abstract

Policy gradient methods are a fundamental tool in reinforcement learning, enabling the direct optimization of parameterized policies. This thesis studies the softmax policy gradient (PG) method in two novel contexts: when (i) using entropy regularization and (ii) linear function approximation. Entropy regularization helps maintain the stochasticity in the policy, and has been shown to aid policy optimization by smoothing the loss landscape. By focusing on multi-armed bandits and tabular Markov decision processes (MDPs), we demonstrate that using softmax PG with a decaying entropy coefficient ensures robust global convergence to the optimal policy. The second part of this thesis considers linear function approximation which is important for handling large state-action spaces. In particular, we consider the linear bandit setting, and prove that Softmax PG is guaranteed to converge to the globally optimal policy only when the features exhibit a specific structure. For both parts of the thesis, we empirically validate our theoretical results and illustrate how our findings can inform the design of more scalable and reliable Softmax PG algorithms in practice.

Keywords: Reinforcement Learning; Bandits; Softmax Policy Gradient; Entropy Regularization; Linear Function Approximation; Convergence Analysis

Preface

The main matter of this thesis is based on two publications.

- *Towards Principled, Practical Policy Gradient for Bandits and Tabular MDPs. Michael Lu, Matin Aghaei, Anant Raj, and Sharan Vaswani, Reinforcement Learning Conference (RLC), 2024*.

  Matin Aghaei is the primary contributor to the "Policy Gradient with Entropy Regularization" section which forms the basis of Chapter 3 of the thesis. Matin was responsible for the writing, theoretical results and experimental evaluation. Michael Lu and Anant Raj provided constructive feedback on the theoretical analysis.

- The second paper entitled *On the Global Convergence of Softmax Policy Gradient for Deterministic and Stochastic Linear Bandits. Qiushi Lin, Jincheng Mei, Matin Aghaei, Michael Lu, Bo Dai, Alekh Agarwal, Dale Schuurmans, Csaba Szepesvári, and Sharan Vaswani* extends and improves the results in Mei et al. (2024a). This paper is under preparation and forms the basis of Chapter 4 in this thesis.

  The first three sections of Chapter 4 motivate the problem and are mainly adapted from the original paper (Mei et al., 2024a). Matin Aghaei is the major contributor to the writing and theoretical results for the exact setting, while Qiushi Lin is the primary contributor to the theoretical results in the stochastic setting. Matin was responsible for the design and implementation of all experiments. Michael Lu assisted with the writing and theoretical results in both the exact and stochastic settings.

Both works were supervised by Sharan Vaswani, who provided valuable guidance, consistent feedback, and support throughout the development of this research.

Acknowledgements

I am profoundly grateful to many individuals whose guidance, support, and encouragement have been indispensable throughout the journey of completing this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Sharan Vaswani. His unwavering guidance, insightful feedback, and constant support have been the cornerstone of my research journey. Sharan's expertise in reinforcement learning and optimization, combined with his genuine encouragement, inspired me to push boundaries and strive for excellence in my work.

I am sincerely thankful to my collaborators who contributed significantly to this research. Michael Lu, Qiushi Lin, Jincheng Mei, and Anant Raj have been invaluable colleagues. Their expertise, constructive discussions, and collaborative spirit greatly enriched my research experience. I deeply appreciate Michael's assistance with writing and theoretical insights, Qiushi's and Jincheng's outstanding theoretical contributions, and Anant's guidance and feedback. Working alongside such talented and supportive peers has been both inspiring and enriching.

I also owe a great debt of gratitude to my family for their unconditional love, faith, and endless support. Their belief in me provided the strength and motivation to overcome challenges and stay focused on my goals. To my family, thank you for always standing by me and encouraging me to pursue my aspirations.

A special note of thanks goes to my old friend and current roommate, Shayan Shafaghi. Shayan's friendship, understanding, and constant encouragement provided not only a comforting presence but also a source of motivation and sanity amidst the rigors of research. His support and companionship have been invaluable through both the highs and lows of this journey.

This thesis stands as a testament to the collaborative efforts, mentorship, and support I have received from all of these individuals. I am deeply thankful to each one of them for their contributions, guidance, and unwavering belief in my potential.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Reinforcement Learning (RL) is a fundamental paradigm in machine learning where agents learn to make sequential decisions by interacting with an environment to maximize cumulative reward (Sutton and Barto, 2018). As agents learn from trial and error, they adapt their behavior to achieve better outcomes over time. For example, RL has enabled computer programs to master complex board games like Go through self-play (Silver et al., 2016), and has been successfully applied in robotics for tasks such as object grasping and navigation (Kober et al., 2013). These successes highlight RL's potential to solve complex decision-making problems that are difficult to address with traditional methods.

A central challenge in RL is the *policy optimization* problem, where the goal is to find an optimal policy that dictates the best action to take in each state to maximize expected returns. A widely used approach to this problem is the softmax policy gradient (Softmax PG) method, which directly parameterizes the policy using a softmax function, ensuring that the resulting policy remains a probability distribution over actions. This differentiable parameterization enables the use of gradient ascent to iteratively train policy parameters in the direction of increasing expected reward (Agarwal et al., 2021; Mei et al., 2020b). The smoothness of the softmax function facilitates stable updates, making Softmax PG naturally applicable to continuous action spaces and scenarios where maintaining stochasticity in the policy is beneficial.

Despite its success, analyzing the Softmax PG method is challenging due to the non-concave nature of the policy optimization objective (Agarwal et al., 2021). Recent theoretical work has made progress in understanding the convergence properties of Softmax PG in simplified settings, such as tabular representations with access to exact gradients (Mei et al., 2020b; Agarwal et al., 2021).

In practice, encouraging exploration is crucial due to non-convex optimization landscapes where policies may prematurely converge to suboptimal deterministic strategies. Entropy regularization augments the objective function to promote stochasticity, smoothing the

optimization surface, and helping the agent escape flat regions (Ahmed et al., 2019). However, incorporating entropy regularization introduces its own theoretical complexities, such as managing bias and determining appropriate decay strategies for the entropy coefficient.

In practical scenarios with large or continuous state and action spaces, representing policies exactly becomes infeasible. To address this, function approximation is employed to compactly represent policies and manage computational complexity. For instance, linear function approximation or neural networks are used to generalize across similar states and actions. Incorporating function approximation alters the structure of the optimization landscape and introduces additional complexities, making the analysis of convergence properties and performance guarantees more challenging and less understood.

By focusing on entropy regularization and linear function approximation, this thesis aims to advance the understanding of Softmax PG, developing practical theoretically-principled algorithms that navigate these challenges without relying on unrealistic assumptions, thereby paving the way for robust convergence in more realistic settings.

1.2 Related Works

The softmax policy gradient method parameterizes the policy using the softmax function, ensuring that the policy remains a valid probability distribution over actions. This method has been extensively studied in the tabular setting, where states and actions are finite.

Specifically, in the exact setting where the rewards and transition probabilities are known, Agarwal et al. (2021) proved that Softmax PG can attain asymptotic convergence to an optimal policy despite the non-concave nature of the PG objective. Mei et al. (2020b) improve this result and quantify the rate of convergence, proving that Softmax PG requires $\mathcal{O}(1/\epsilon)$ iterations to converge to an ϵ-optimal policy.

In the stochastic setting where the rewards and transition probabilities are unknown and algorithms require sampling from the environment, Zhang et al. (2020b) first proved that REINFORCE (Williams, 1992; Sutton et al., 1999) converges to a first-order stationary point at a rate of $\tilde{\mathcal{O}}(1/\epsilon^2)$. Mei et al. (2021a, 2022b) analyzed the convergence of stochastic Softmax PG, proving that it requires $\mathcal{O}(1/\epsilon^2)$ iterations to converge to an ϵ-optimal policy. However, the resulting algorithm requires the full gradient (which in turn requires knowledge of the environment) to set the algorithm parameters, making it impractical in the stochastic setting. Similarly, Yuan et al. (2022b) proved that stochastic Softmax PG converges to an optimal policy at a slower $\tilde{\mathcal{O}}(1/\epsilon^3)$ rate. However, this result requires knowledge of the optimal action, making it impractical. More recently, Mei et al. (2023) analyzed stochastic Softmax PG in the multi-armed bandit setting and proved that it converges to the optimal arm at an $\mathcal{O}(1/\epsilon)$ rate. Unfortunately, the algorithm requires knowledge of the reward gap which is typically unknown even in the simplified settings such as multi-armed bandit problems.

A related line of work (Lu et al., 2024) further addresses the dependence on unknown problem-dependent quantities by leveraging ideas from optimization—particularly, exponentially decreasing step-sizes—to design practical Softmax PG algorithms in both exact and stochastic settings. These methods avoid relying on oracle-like knowledge (e.g., the reward gap, the reward distributions, or the noise level) yet preserve strong theoretical guarantees. Specifically, they show that employing exponentially decreasing step-sizes, rather than fixed or gap-dependent learning rates, yields convergence results comparable to the state-of-the-art while making fewer assumptions on the environment. Empirical results confirm that these step-size strategies allow Softmax PG to perform competitively against methods that do require the knowledge of problem-specific constants.

Overall, although Softmax PG has appealing theoretical properties in simplified scenarios, adapting it realistic environments with partially known or entirely unknown transition dynamics and rewards has several challenges.

1.2.1 Limitations and Challenges

- **Non-Concavity**: The policy optimization objective is non-concave, making it difficult to design and analyze algorithms that ensure global convergence.
- **Dependence on Problem Constants**: Many theoretical guarantees assume access to environment-dependent constants, such as a reward gap, which are typically unknown in practice.
- **Sample Efficiency**: In the stochastic setting, estimating gradients from samples can require a large number of interactions, affecting practical performance.
- **Function Approximation**: Scaling to large or continuous state and action spaces necessitates function approximation, introducing additional complexities not addressed by standard analyses.

1.3 Thesis Contributions

Building on the challenges outlined above, this thesis extends the analysis of the standard Softmax Policy Gradient (PG) method in two novel contexts: entropy regularization and linear function approximation. These investigations are motivated by the need to handle exploration without relying on oracle-like knowledge and to understand global convergence beyond traditional approximation error assumptions.

The key contributions of this thesis are as follows:

- **Entropy-Regularized Softmax PG**: We introduce a multi-stage algorithm that iteratively decays the entropy regularization term. This approach enables the resulting

algorithm to escape flat regions thus enabling robust convergence to the optimal policy without requiring problem-dependent constants.

- **Global Convergence under Linear Approximation**: For the linear bandits setting, we analyze the convergence of Softmax PG when combined with linear function approximation. Our results show that global convergence does not solely rely achieving a small approximation error. Instead, we identify the necessary structural conditions – such as rank preservation and specific feature geometry—that ensure convergence to the globally optimal policy despite significant approximation errors.
- **Theoretical and Empirical Insights**: We provide rigorous theoretical analysis supporting these methods, including convergence rates and proofs of global optimality. We validate our theoretical findings in the bandit setting, and demonstrate the effectiveness of the proposed algorithms.

These contributions deepen our understanding of Softmax PG, offering practical algorithms and theoretical insights that advance its applicability to complex reinforcement learning tasks.

Chapter 2

Background and Preliminaries

2.1 Fundamentals of Reinforcement Learning

Reinforcement Learning (RL) is a framework for learning sequential decision-making policies by interacting with an environment. An RL problem is typically modeled as an infinite-horizon discounted Markov Decision Process (MDP) (Puterman, 2014), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho, \gamma)$, where:

- $\mathcal{S}$: A finite set of states, with $S = |\mathcal{S}|$.
- $\mathcal{A}$: A finite set of actions, with $A = |\mathcal{A}|$.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$: State transition probability function, where $\Delta_{\mathcal{S}}$ denotes the probability simplex over $\mathcal{S}$.
- $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$: Reward function, assigning a reward to each state-action pair.
- $\rho \in \Delta_{\mathcal{S}}$: Initial state distribution.
- $\gamma \in [0, 1)$: Discount factor.

We focus on *tabular MDPs*, assuming that the state and action spaces are finite. At each time step t, the agent observes a state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$ according to its policy π, receives a reward $r_t = r(s_t, a_t)$, and transitions to the next state s_{t+1} according to the transition probabilities $\mathcal{P}(s_{t+1} \mid s_t, a_t)$. We assume a uniform initial state distribution, that is, $\rho(s) = \frac{1}{S}$ for all $s \in \mathcal{S}$. This assumption is common in the policy gradient literature and simplifies the analysis by focusing on optimization aspects without dealing with exploration challenges (Mei et al., 2020b).

2.2 Policy Optimization

The central goal in reinforcement learning is to determine a policy that maximizes the expected cumulative reward. A *policy* π is a mapping from states to probability distributions over actions, where $\pi(a|s)$ denotes the probability of taking action a in state s.

For a given policy π, the *action-value function* $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ quantifies the expected return starting from state s, taking action a, and then following policy π:

$$Q^\pi(s,a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, s_0 = s,\, a_0 = a\Big].$$

The corresponding *value function* $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined as:

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\Big[Q^\pi(s,a)\Big].$$

The *advantage function* $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ measures how much better taking action a in state s is compared to the average.

For a state $s \in \mathcal{S}$, the *discounted state visitation distribution* starting from s_0 is:

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr^\pi[s_t = s \,|\, s_0],$$

representing the normalized discounted frequency of visiting state s when starting from s_0.

Given an initial state distribution ρ, the policy optimization objective is:

$$\max_{\pi\in\Pi} J(\pi) = \mathbb{E}_{s_0\sim\rho}\Big[V^\pi(s_0)\Big] = V^\pi(\rho).$$

We denote the optimal policy as $\pi^* = \arg\max_{\pi\in\Pi} J(\pi)$.

In the *bandit* setting, where $|S| = 1$ and $\gamma = 0$, the problem simplifies to:

$$J(\pi) = \mathbb{E}_{a\sim\pi}[r(a)] = \langle \pi, r\rangle,$$

with $\pi \in \Delta_{\mathcal{A}}$ as the probability distribution over actions and $r \in [0,1]^A$ as the reward vector.

2.3 Policy Parameterization

We consider policies with a *softmax tabular parameterization*. For parameters $\theta \in \mathbb{R}^{S\times A}$, the policy $\pi_\theta : S \to \Delta_A$ is defined using the softmax function:

$$\pi_\theta(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a'\in\mathcal{A}} \exp(\theta(s,a'))}. \tag{2.1}$$

This parameterization ensures that $\pi_\theta(\cdot|s)$ is a valid probability distribution for each state s. The softmax tabular parameterization has been used in recent theoretical analyses of policy gradient methods (Mei et al., 2020b; Agarwal et al., 2021). Throughout this thesis, we denote the objective function as $f(\theta)$, which depends on the setting:

- In the MDP setting: $f(\theta) = V^{\pi_\theta}(\rho)$.

- In the bandit setting: $f(\theta) = \langle \pi_\theta, r \rangle$.

By abstracting $f(\theta)$ in this way, our results can be generalized to other settings, such as constrained MDPs (Altman, 2021) or convex MDPs (Zahavy et al., 2021; Zhang et al., 2020a). The optimal policy π^* is deterministic in both the bandit and MDP settings (Puterman, 2014). Specifically, in the MDP setting, for each state $s \in \mathcal{S}$, there exists an action $a^*(s) \in \mathcal{A}$ such that:

$$\pi^*(a^*(s)|s) = 1, \quad \pi^*(a|s) = 0 \quad \text{for all } a \neq a^*(s). \tag{2.2}$$

In the softmax parameterization, this corresponds to $\theta^*(s, a^*(s)) \to \infty$ or $\theta^*(s, a) \to -\infty$ for all $a \neq a^*(s)$. This behavior is similar to logistic regression in classification tasks with linearly separable data (Ji and Telgarsky, 2018).

2.4 Properties of the Objective Function

Setting	$f(\theta)$	$[\nabla f(\theta)]_{s,a}$	L	$C(\theta)$
Bandits	$\langle \pi_\theta, r \rangle$	$\pi_\theta(a)\left[r(a) - \langle \pi_\theta, r \rangle\right]$	$5/2$	$\pi_\theta(a^*)$
MDPs	$V^{\pi_\theta}(\rho)$	$\frac{d^{\pi_\theta}(s)\,\pi_\theta(a\|s)\,A^{\pi_\theta}(s,a)}{1-\gamma}$	$\frac{8}{(1-\gamma)^3}$	$\frac{\min_s \pi_\theta(a^*(s)\|s)}{\sqrt{S}\left\|\frac{d_\rho^{\pi^*}}{d_\rho^{\pi_\theta}}\right\|_\infty}$

Table 2.1: Function and gradient expressions, uniform smoothness, and non-uniform Łojasiewciz properties for bandits and MDPs with $\xi = 0$ (Mei et al., 2020b). Here, a^* is index of the optimal arm in the bandit problem, and $a^*(s)$ is the optimal action at state s in the MDP problem.

2.4.1 Smoothness and Non-Concavity

The function $f(\theta)$ is non-concave with respect to θ in both bandit and MDP settings (Mei et al., 2020b, Proposition 1). However, it is twice differentiable and satisfies *uniform smoothness*, meaning there exists a constant $L > 0$ such that for all θ:

$$\nabla^2 f(\theta) \preceq L I_{SA}, \tag{2.3}$$

where I_{SA} is the identity matrix of size $S \times A$.

2.4.2 Łojasiewicz Conditions

The function $f(\theta)$ satisfies a *non-uniform Łojasiewicz (Ł) condition* (Mei et al., 2020b):

$$\|\nabla f(\theta)\|_2 \geq C(\theta)\left(f^* - f(\theta)\right)^{1-\xi}, \tag{2.4}$$

where $f^* = \max_\theta f(\theta)$, $C(\theta) > 0$, and $\xi \in [0, 1]$. When $C(\theta)$ is a constant and $\xi = \frac{1}{2}$, this reduces to the well-known Polyak-Łojasiewicz (PŁ) condition (Polyak, 1963; Karimi et al., 2016). The Łojasiewciz condition states that every stationary point $\tilde{\theta}$ (s.t. $\nabla f(\tilde{\theta}) = 0$) is also a global maximum s.t. $f(\tilde{\theta}) = f^*$. This condition enables the convergence of local ascent methods such as PG to an optimal solution $\theta^* := \arg\max_\theta f(\theta)$ despite the problem's non-concavity (Karimi et al., 2016; Mei et al., 2020b; Agarwal et al., 2021).

Table 2.1 summarizes both the uniform and non-uniform smoothness and Łojasiewciz properties for bandits and MDPs.

2.5 Softmax Policy Gradient

Softmax PG aims to optimize policies by directly maximizing the objective function $f(\theta)$ with respect to the policy parameters θ. This method performs gradient ascent on the objective function $f(\theta)$ using its gradient $\nabla f(\theta)$.

2.5.1 Exact Setting

In the exact setting, we assume that the reward function r and the transition probability function $\mathcal{P}$ are known. This allows for the exact computation of the gradient $\nabla f(\theta)$. With the exact gradient calculated, the policy parameters are updated using the following update.
Update 1. *(Softmax PG, True Gradient)* $\theta_{t+1} = \theta_t + \eta_t \nabla f(\theta_t)$.

Refer to Table 2.1 for the gradient expressions of the policy gradient $\nabla f(\theta)$ in both the bandit and MDP cases.

2.5.2 Stochastic Setting

In many reinforcement learning problems, the environment dynamics—such as the reward function r and the transition probabilities $\mathcal{P}$—are unknown to the agent. Instead, the agent must learn to optimize its policy solely through interactions with the environment, relying on sampled experiences to estimate the necessary quantities. This scenario is referred to as the *stochastic setting*. In this setting, the policy gradient cannot be computed exactly due to the lack of full knowledge about the environment. Instead, the agent must estimate the gradient using sampled data, which introduces randomness and potential variance in the estimates.

For simplicity, in this section, we focus on the bandit setting, which captures the essential aspects of the stochastic setting while being more tractable for analysis. In the stochastic multi-armed bandit problem, each arm $a \in \mathcal{A}$ has an unknown reward distribution P_a. At each iteration $t \in [1, T]$, the algorithm selects an arm $a_t \in \mathcal{A}$ according to its current policy π_t then receives a stochastic reward R_t sampled from the reward distribution of the selected arm: $R_t \sim P_{a_t}$. The algorithm then constructs an on-policy importance sampling (IS) reward

estimate for each action $a \in \mathcal{A}$:

$$\hat{r}_t(a) = \frac{\mathbb{I}\{a_t = a\}}{\pi_t(a)} R_t, \tag{2.5}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. The IS reward estimate is then used to form the stochastic gradient $\widetilde{\nabla} f(\theta_t)$ such that $\widetilde{\nabla} f(\theta_t)(a) = \pi_{\theta_t}(a)[\hat{r}_t(a) - \langle \pi_{\theta_t}, \hat{r}_t \rangle]$. Mei et al. (2021a, Lemma 5) showed that the resulting stochastic gradients are (i) unbiased i.e. $\mathbb{E}[\widetilde{\nabla} f(\theta)] = \nabla f(\theta)$ and have (ii) bounded variance i.e. $\mathbb{E}\left\|\widetilde{\nabla} f(\theta) - \nabla f(\theta)\right\|_2^2 \leq \sigma^2$. Similarly, we can construct gradient estimators that are unbiased and have bounded variance for MDPs. Given these estimators, the resulting stochastic Softmax PG algorithm has the following update:

Update 2. *(Stochastic Softmax PG, Importance Sampling)* $\theta_{t+1} = \theta_t + \eta_t \widetilde{\nabla} f(\theta_t)$.

This update rule allows the agent to improve its policy based on the sampled rewards, even without full knowledge of the reward distributions.

This foundation sets the stage for analyzing extensions of Softmax PG, such as incorporation of entropy regularization, which is the focus of the next chapter.

Chapter 3

Softmax Policy Gradient with Entropy Regularization

3.1 Introduction

As reinforcement learning agents tackle increasingly complex tasks, effective exploration becomes critical to avoid suboptimal behaviors and ensure robust performance. Entropy regularization is a promising technique for encouraging such exploration by maintaining policy diversity and smoothing the optimization landscape.

We will next consider adding entropy regularization to the objective in the exact and stochastic settings. Entropy regularization RL, also known as maximum entropy RL, uses entropy regularization to promote action diversity and prevent premature convergence to a deterministic policy (Williams, 1992; Haarnoja et al., 2018). Although it is widely believed to help with exploration, the addition of entropy regularization results in a smoother optimization landscape, allowing PG methods to escape flat regions within the optimization landscape (Ahmed et al., 2019). For example, in the bandit setting, flat regions occur when a policy commits to an arm. Mei et al. (2020b) showed entropy regularization helps escaping these regions when starting from a "bad" initialization, i.e. the initial policy selects a sub-optimal arm with high probability.

In the exact setting, where the full gradient can be computed, Mei et al. (2020b) showed Softmax PG with entropy regularization obtains a fast $\mathcal{O}(\log(1/\epsilon))$ rate to a biased ϵ-optimal policy. The resulting optimal policy is biased since the presence of entropy prevents convergence to a deterministic policy. Furthermore, in the same setting, Cen et al. (2022) showed NPG with entropy regularization achieves the same $\mathcal{O}(\log(1/\epsilon))$ convergence rate to a biased ϵ-optimal policy. To ensure that the resulting optimal policy is unbiased, the strength of the entropy regularization term must be decayed or removed. Mei et al. (2020b) introduced a two-stage approach to obtain the optimal policy when using Softmax PG with entropy regularization. In the first stage, entropy regularization obtains fast convergence close to the optimal policy. In the second stage, the regularizer is removed to guarantee

convergence to the optimal policy. Unfortunately, the final convergence rate is $\mathcal{O}(1/\epsilon)$, which is the same as the Softmax PG. Additionally, to transition from the first to the second stage, the reward gap is needed, making the resulting algorithm impractical.

In the stochastic setting, where the value function must be approximated, Ding et al. (2021) introduced a two-stage approach for stochastic Softmax PG with entropy regularization. Instead of modifying the strength of the entropy regularizer across stages, the batch size is modified. The resulting algorithm requires $\mathcal{O}(1/\epsilon)$ iterations in the second stage and $\tilde{\mathcal{O}}(1/\epsilon^2)$ samples to converge to a biased ϵ-optimal policy. The method allows for global convergence with arbitrary initiation. However, the strength of the entropy regularizer is not decayed, preventing convergence to the optimal policy. Additionally, the biased optimal policy to set the algorithm hyper-parameters making the resulting algorithm redundant. Moreover, in the stochastic setting with access to a generative model, using NPG with entropy regularization, Cen et al. (2022) achieved a linear convergence rate to a biased optimal policy with a $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity.

A recent line of work by Lu et al. (2024) addresses the reliance on unknown parameters in the non-regularized setting by leveraging exponentially decreasing step-sizes to design practical Softmax PG algorithms in both exact and stochastic settings. By using such step-size schedules, their methods avoid reliance on oracle-like knowledge (e.g., reward gaps or noise levels), yet preserve strong theoretical guarantees. Empirical results confirm that exponentially decreasing step-sizes allow Softmax PG to perform competitively against methods dependent on problem-specific constants, inspiring the application of similar techniques in the entropy-regularized setting.

In the following sections, we will present a multi-stage algorithm that iteratively reduces the strength of the entropy regularization term. This method obtains convergence to the optimal policy while eliminating the reliance on unknown quantities compared to the prior work. In Section 3.2 we first state how the objective's functional property changes when entropy regularization is added. In Section 3.3 we present the multi-stage algorithm in the exact setting and the algorithm achieves an $\mathcal{O}(1/\epsilon^p)$ rate, where p depends on the properties of the entropy-regularized objective. Next in Section 3.4, we extend the same multi-stage algorithm in the stochastic setting with exponentially decreasing step-sizes to obtain an also $\mathcal{O}(1/\epsilon^{2p+1})$ rate to the optimal policy. Finally, in Section 3.4.1 we compare the proposed our multi-stage algorithm to prior PG methods without entropy regularization and show that the multi-stage algorithm helps escape flat regions within the optimization landscape.

3.2 Problem Formulation

Following Chapter 2, for a policy π, the *entropy-regularized action-value function* is defined as $\tilde{Q}^\pi_\tau(s,a) := \mathbb{E}[\sum_{t=0}^\infty \gamma^t (r(s,a) - \tau \log \pi)]$ and the *entropy-regularized value function* is defined

as $\tilde{V}_\tau^\pi(s) := \mathbb{E}_{a\sim\pi}[\tilde{Q}_\tau^\pi(s,a)](s)$. The *entropy-regularized advantage function* is defined as $\tilde{A}_\tau^\pi(s,a) := \tilde{Q}_\tau^\pi(s,a) - \tau \log \pi(a|s) - \tilde{V}_\tau^\pi(s)$.

Furthermore, let $f^\tau(\theta) := f(\theta) + \tau\, \Lambda(\pi_\theta)$ denote the entropy-regularized objective, where $\Lambda(\pi_\theta)$ is the "discounted entropy" for a policy π_θ and $\tau \geq 0$ is the "temperature" or strength of the entropy regularization. Refer to Table 2.1, for definitions of discounted entropy in the bandit and MDP settings. For a fixed τ, f^τ is L^τ-uniform smooth and note that the smoothness now depends on τ. Furthermore, f^τ satisfies a non-uniform Łojasiewciz condition with $C_\tau(\theta)$ and $\xi = 1/2$. Compared to f, whose non-uniform Łojasiewciz degree is $\xi = 0$ (refer to Table 2.1), the increase to $\xi = 1/2$ allows for faster convergence. Table 3.1 summarizes the entropy regularizer, uniform smoothness and non-uniform Łojsaiewciz properties for the bandit and general MDP settings with entropy regularization. Finally, we will denote the maximum value of the regularized objective function as $f^{*\tau} := f^\tau(\theta_\tau^*)$, where $\theta_\tau^* := \arg\max_\theta f^\tau(\theta)$.

Setting	$\Lambda(\pi_\theta)$	$[\nabla f^\tau(\theta)]_{s,a}$	L^τ	$C_\tau(\theta)$
Bandits	$-\langle \pi_\theta, \log \pi_\theta \rangle$	$\pi_\theta(a)\,[r(a) - \langle \pi_\theta,\, r - \tau \log \pi_\theta \rangle]$	$5/2 + 5\,\tau\,(1 + \log A)$	$\sqrt{2\tau}\, \min_a \pi_\theta(a)$
MDPs	$\mathbb{H}(\pi_\theta)$	$\frac{d_\rho^{\pi_\theta}(s)\, \pi_\theta(a\|s)\, \tilde{A}^{\pi_\theta}(s,a)}{1-\gamma}$	$\frac{8+\tau\,(4+8\log A)}{(1-\gamma)^3}$	$\frac{\sqrt{\tau}\, \min_s \sqrt{\rho(s)}\ \min_{s,a} \pi_\theta(a\|s)}{S \left\| \frac{d_\rho^{\pi^{*\tau}}}{d_\rho^{\pi_\theta}} \right\|_\infty^{1/2}}$

Table 3.1: Entropy regularizer, uniform smoothness and non-uniform Łojasiewciz condition with $\xi = 1/2$ for bandit and general tabular MDP settings with entropy regularization. Here, $\mathbb{H}(\pi_\theta) := \mathbb{E}[\sum_{t=0}^\infty -\gamma^t \log \pi_\theta(a_t|s_t)]$.

With the above properties of f^τ, we next present how to principally decay τ for Softmax PG with entropy regularization to obtain convergence to the optimal policy.

3.3 Exact Setting

We first consider the exact setting as a test bed to analyze how to decay τ to obtain convergence to the optimal policy. Recall that for a constant $\tau > 0$, Softmax PG with entropy regularization is unable to converge to the optimal policy, since the regularizer prevents the final policy from becoming deterministic. Softmax PG with entropy regularization has the following update:

Update 3. *(Softmax PG with Entropy Regularization, True Gradient)* $\theta_{t+1} = \theta_t + \eta_t \nabla f^\tau(\theta_t)$.

Refer to Table 3.1 for the entropy-regularized policy gradient $\nabla f^\tau(\theta)$ in both the bandit and the general MDP cases. In this setting, Mei et al. (2020b) show that Softmax PG with entropy regularization converges to a biased optimal policy at a rate of $\mathcal{O}(\log 1/\epsilon)$ when using a fixed step-size of $\eta_t = \eta = \frac{1}{L^\tau}$. The optimal policy is biased since $\tau > 0$ is fixed. Consequently, it is necessary to to decay the regularization strength τ in order to converge

to the globally optimal policy. In the bandit setting, Mei et al. (2020b) proposed a two-stage approach to decay τ to obtain global convergence. A fixed $\tau > 0$ is used in the first stage but is then set to 0 in the second stage. However, the resulting algorithm requires knowledge of the reward gap $\Delta := \max_{a^* \neq a} r(a^*) - r(a)$ in order to transition from the first stage to the second stage, making the method impractical. Furthermore, Mei et al. (2020b) proposed an additional approach by allowing τ to be a function of t and slowly decreasing τ_t over time. This approach also obtains convergence to the global optimal policy. However, it required $\tau_t \propto \Delta$, meaning that the algorithm again requires the knowledge of the reward gap.

Algorithm 1: Multi-Stage Softmax PG with Entropy Regularization

Output: Policy $\pi_{\theta_t} = \text{softmax}(\theta_t)$
Initialize parameters $\theta_0, \tau_0, N_{\text{stages}}$
$t \leftarrow 0$
$\text{last}_0 \leftarrow t$
$i \leftarrow 1$
while $i \leq N_{stages}$ **do**
  $\tau_i \leftarrow \tau_{i-1}/2$
  $\eta_i \leftarrow 1/L^{\tau_i}$
  $T_i \leftarrow \frac{2}{\eta_i \mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}(1 + B_4)\right)$
  while $t - last_{i-1} < T_i$ **do**
    $\theta_{t+1} \leftarrow \theta_t + \eta_i \nabla f^{\tau_i}(\theta_t)$
    $t \leftarrow t + 1$
  end
  $\text{last}_i \leftarrow t$
  $i \leftarrow i + 1$
end

In order to design a theoretically principled algorithm that decays the entropy without a dependence on the reward gap, we assume that f^τ satisfies an alternative non-uniform Łojasiewciz condition with $\xi = 1/2$. The following assumption is similar to the non-uniform Łojasiewciz condition discussed in Section 3.2, only with a different non-uniform constant compared to the one presented in Table 3.1.

Assumption 1. *f^τ satisfies the non-uniform Łojasiewciz condition for some $C_\tau(\theta)$ and $\xi = \frac{1}{2}$ such that $\mu := \inf_{t \geq 1}[C_\tau(\theta_t)]^2 = \tau^p B_1$ for constants $p \geq 1$ and $B_1 > 0$.*

Subsequently, we provide empirical evidence justifying this assumption. Under Assumption 1, we propose a multi-stage algorithm (Algorithm 1) to decay τ that can obtain ϵ convergence to the globally optimal policy without knowledge of the reward gap or any other problem-dependent parameters. Algorithm 1 operates in multiple stages, each using a temperature τ_i for T_i iterations before halving it, i.e. $\tau_{i+1} = \frac{\tau_i}{2}$. To understand why this scheduling is useful, define the *suboptimality gap* as $f^* - f(\theta)$, where $f^* = \max_\theta f(\theta)$. In our entropy-regularized setting, this gap can be split into two parts:

- **Optimization Gap**: The part of the suboptimality gap that we can reduce through gradient-based updates.
- **Regularization Bias**: The additional error stems from the fact that $\tau > 0$ prevents the policy from reaching the optimal policy.

By decreasing τ at each stage, we progressively reduce the bias originating from entropy regularization. We run the algorithm at stage i until the optimization gap is comparable to the bias level introduced by τ_i. Although we halve τ for theoretical convenience, other decay schedules may also work in practice.

Moreover, if $p = 1$ in Assumption 1, then the non-uniform constant of the Łojasiewciz condition scales linearly with τ. Halving τ halves this constant, so we must roughly double the length T_i of each stage to keep the optimization gap sufficiently small throughout. This balancing of bias reduction and optimization progress ultimately allows the algorithm to converge to an unbiased optimal policy without relying on unknown constants.

To prove that the method achieves global convergence, we first make the following assumptions to relate the entropy regularization objective f^τ to the unregularized objective f:

Assumption 2. *f^τ is L^τ-smooth and $L^\tau \leq L^{\max}$, where $L^{\max} = \max_{\tau \in [0,1]} L^\tau$ is a constant. Furthermore, $L^\tau \geq L^{\min}$, where $L^{\min} = \min_{\tau \in [0,1]} L^\tau > 0$ is a constant.*

Assumption 3. *$f^* - f(\theta^*_\tau) \leq \tau B_2$, for a constant $B_2 > 0$.*

Assumption 4. *For a constant $B_3 > 0$, $f(\theta^*_\tau) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau B_3$.*

Assumption 5. *For $\tau_2 < \tau_1$ and a constant $B_4 > 0$, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 B_4$.*

The Assumptions 2 to 5 hold for both the bandit and tabular MDP settings and are proved in Section A.2.2 and Section A.2.3, respectively.

The following theorem (proved in Section A.2) shows that Algorithm 1 converges to the unbiased optimal policy at a rate of $\mathcal{O}(1/\epsilon^p)$.

Theorem 1. Assuming f^τ and f satisfy Assumptions 1 to 5, for a given $\epsilon \in (0, 1)$, Algorithm 1 achieves ϵ-suboptimality to the global optimal policy after $T_{\text{total}} = \frac{4\, L^{\max} C_1^p}{\epsilon^p\, B_1} \log\left(2\ (1 + B_4)\right)$ iterations, where $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3$.

The resulting $\mathcal{O}(1/\epsilon^p)$ rate depends on the constant p in Assumption 1. In the best case, when $p = 1$, we recover an $\mathcal{O}(1/\epsilon)$ convergence rate, similar to the non-regularized Softmax PG. This is verified by our experiments, in particular, in Figure 3.1, we observe that Algorithm 1 with $p = 1$ has a similar performance compared to Softmax PG. This provides empirical evidence that Assumption 1 holds with $p = 1$. In contrast to the above result, in the non-regularized bandit setting, the convergence rate of Softmax PG is inversely proportional to $\inf_{t \geq 1} \pi_{\theta_t}(a^*)^2$, which can be arbitrarily small due to poor initialization. We note that this advantage of entropy-regularized Softmax PG extends to the MDP setting. It is important

to note that compared to Mei et al. (2020b), Algorithm 3 can obtain ϵ-convergence without requiring knowledge of the reward gap.

In the bandit setting with $A = 10$, we compare Algorithm 1 (`PG-E-MS`), assuming $p = 1$ and setting $B_1 = 0.01$, and Softmax PG (`PG`) with a fixed step-size of $\eta_t = \frac{1}{L} = \frac{2}{5}$, and Softmax PG with entropy regularization (`PG-E`) with fixed $\tau = 0.1$ and $\eta_t = \eta = \frac{1}{L^\tau} = \frac{2}{5+10\,\tau(1+\log A)}$, and Softmax PG with decaying entropy regularization (`PG-DE`), which requires oracle knowledge (Theorem 8 from (Mei et al., 2020b)) with $\alpha = 1$. For `PG-E-MS`, the initial regularization strength is $\tau_0 = 1$, and p and B_1 were selected by using grid-search on a separate set of bandit instances. We test the algorithms on bandit settings of varying difficulty based on their minimum reward gap $\underline{\Delta} := \min_{a^* \neq a} r(a^*) - r(a)$. The easy, medium and hard environments correspond to $\underline{\Delta} = 0.2, 0.1, 0.05$ respectively. The figure plots the average and 95% confidence interval of 50 random mean reward vectors.

In most realistic scenarios, it is difficult to calculate the exact gradient of the objective function. In the next section, we investigate how to extend the presented multi-stage algorithm to the stochastic setting.

3.4 Stochastic Setting

Following Chapter 2, we can construct a stochastic policy gradient using on-policy importance sampling (IS) reward estimates for the entropy-regularized objective. Let $\widetilde{\nabla} f^\tau(\theta_t)$ denote the stochastic gradient with entropy regularization. By Lemma 29, the gradient estimators $\widetilde{\nabla} f^\tau(\theta)$ are (i) unbiased, i.e. $\mathbb{E}[\widetilde{\nabla} f^\tau(\theta)] = \nabla f^\tau(\theta)$ and have (ii) bounded variance, i.e. $\mathbb{E}\left\|\widetilde{\nabla} f^\tau(\theta) - \nabla f^\tau(\theta)\right\|_2^2 \leq \sigma^2$. The bound of the variance differs compared to $\widetilde{\nabla} f(\theta)$ as σ^2 depends on the regularization strength τ. In this setting, we will consider the following update,

Update 4. *(Stochastic Softmax PG with Entropy, Importance Sampling)* $\theta_{t+1} = \theta_t + \eta_t \widetilde{\nabla} f^\tau(\theta_t)$.

Under the same setting when using on-policy IS reward estimates, prior work (Ding et al., 2021) proposes a two-stage approach that converges to a biased optimal policy by modifying the batch size to counteract the variance. However, the method requires a $\tilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity and knowledge of the biased optimal policy to set the algorithm hyper-parameters. Furthermore, even with knowledge of the biased optimal policy, Ding et al. (2021) is unable to converge to the optimal policy.

To extend Algorithm 1 to the stochastic setting, we first require an additional assumption since $\inf_{t\geq 1}[C_\tau(\theta_t)]^2$ is now a random variable in the stochastic setting.

Assumption 6. *f^τ satisfies the non-uniform Łojasiewciz condition for some $C_\tau(\theta)$ and $\xi = \frac{1}{2}$ such that $\mu := \mathbb{E}\left[\inf_{t\geq 1}[C_\tau(\theta_t)]^2\right] = \tau^p B_1$ for constants $p \geq 1$ and $B_1 > 0$.*

Under Assumption 6, we will utilize *exponentially decaying step-sizes* (Li et al., 2021b; Vaswani et al., 2022; Lu et al., 2024) for each stage to handle the noise inherent in the

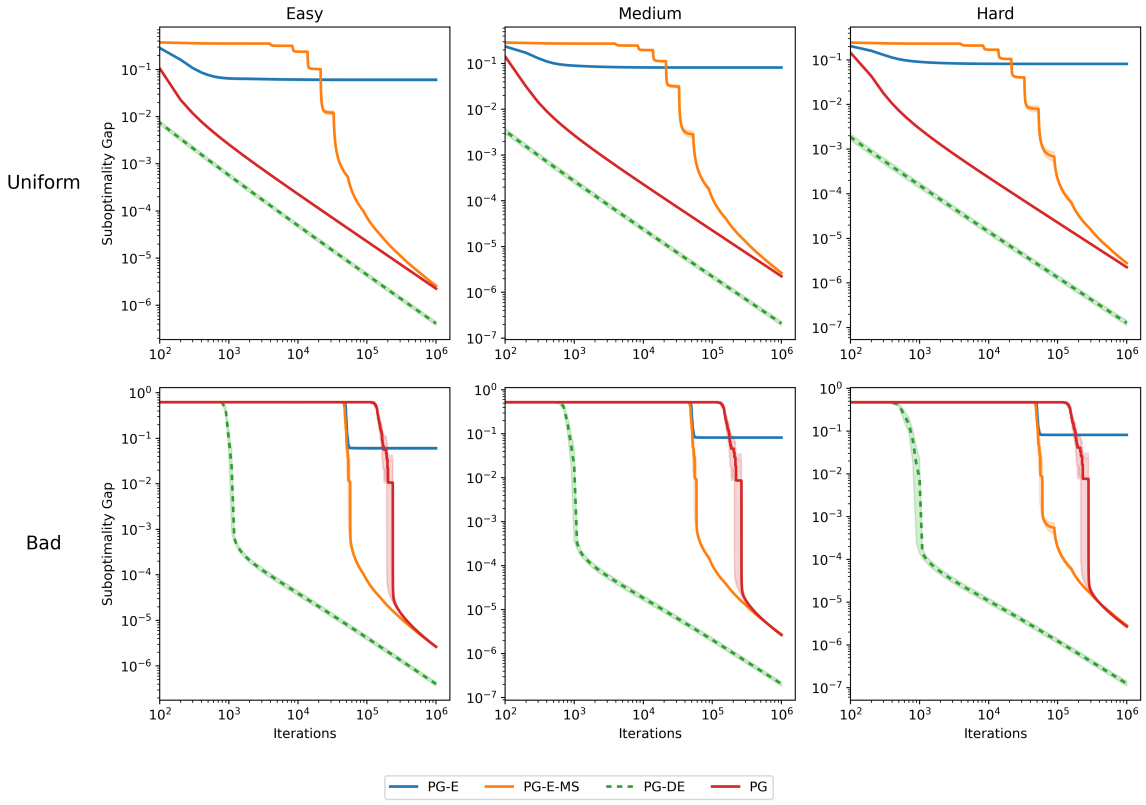


Figure 3.1: Sub-optimality gap across various environments and initializations. Top Row: the initial policy's parameters is uniform, i.e. $\theta_0(a) = 0 \quad \forall a$. Bottom Row: the initial policy's parameters is "bad", i.e. $\theta_0(a') = 12$ where $a' = \arg\min_a r(a)$. `PG-E-MS` can converge to the optimal policy unlike `PG-E` since the temperature τ is decreasing. Furthermore, under "bad" initialization, where the worst arm has a high probability of being chosen, `PG-E-MS` outperforms `PG` since the addition of entropy allows the method to escape the initial flat region. On the other hand, `PG-E` can escape the initial region quickly but cannot converge to the optimal policy since τ is fixed. `PG-DE` has a good performance in all settings, but requires oracle knowledge.

stochastic gradient estimates. In particular, even if the noise variance is bounded, using fixed step-sizes can prevent convergence to the optimum. By shrinking the step-size over time, the effect of the noise diminishes over the course of optimization and enables convergence to the solution. At stage i, the resulting step-size at iteration t is set as: $\eta_{i,t-1} = \frac{1}{L^{\tau_i}} \alpha_i^{t-\text{last}_{i-1}}$ where $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta \geq 1$, and T_i is the length of stage i. Together, this results in Algorithm 2.

Algorithm 2: Stochastic Multi-Stage Softmax PG with Entropy Regularization

Output: Policy $\pi_{\theta_t} = \text{softmax}(\theta_t)$
Choose constants c_0, c_1 according to the theory (see Section A.3.1)
Initialize parameters $\theta_0, \tau_0, N_{\text{stages}}, \beta = 1$
$t \leftarrow 0$
$\text{last}_0 \leftarrow t$
$i \leftarrow 1$
while $i \leq N_{stages}$ **do**
  $\tau_i \leftarrow \frac{\tau_{i-1}}{2}$
  $X_1 \leftarrow \exp\Big(\frac{\mu_i\,\beta}{L^\tau\,\log(T/\beta)}\Big)$
  $X_2 \leftarrow \frac{c_0}{L^\tau}$
  $X_3 \leftarrow \frac{5\,L^\tau\,X_1}{e^2}$
  $T_i^{'} \leftarrow \frac{2}{X_2\,\mu_i}\log\Big(\frac{2\,X_1\,\tau_{i-1}}{\tau_i}\,(1+B_4)\Big)$
  $T_i^{''} \leftarrow \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}$
  $T_i \leftarrow \max(c_1, 2\,T_i^{'}\,\log T_i^{'}, 4\,T_i^{''}\,\log^2 T_i^{''})$
  $\alpha_i \leftarrow \Big(\frac{\beta}{T_i}\Big)^{\frac{1}{T_i}}$
  $\eta_{i,t} \leftarrow \frac{\alpha_i}{L^{\tau_i}}$
  while $t - last_{i-1} < T_i$ **do**
    $\theta_{t+1} \leftarrow \theta_t + \eta_{i,t}\,\tilde{\nabla} f^\tau(\theta_t)$
    $\eta_{i,t+1} \leftarrow \eta_{i,t}\,\alpha_i$
    $t \leftarrow t+1$
  end
  $\text{last}_i \leftarrow t$
  $i \leftarrow i+1$
end

The following theorem (proved in Section A.3.1) shows that Algorithm 2 converges to the globally optimal policy at an $\tilde{\mathcal{O}}\left(1/\epsilon^p + \sigma^2/\epsilon^{2p+1}\right)$ rate.

Theorem 2. Assuming f^τ and f satisfy Assumptions 2 to 6, for a given $\epsilon \in (0,1)$, using Algorithm 2 with (a) unbiased stochastic gradients whose variance is bounded by σ^2 and (b) exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\,\alpha_i^{t-\text{last}_{i-1}+1}$, where $\eta_{i,\text{last}_{i-1}} = \frac{1}{L^{\tau_i}}$, $\alpha_i = \Big(\frac{\beta}{T_i}\Big)^{\frac{1}{T_i}}$, $\beta = 1$, and setting constants $c_0 = 0.69, c_1 = 5583$, achieves ϵ-sub-optimality to the globally optimal policy after $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

If $p = 1$, then the convergence rate matches the $\tilde{\mathcal{O}}(\sigma^2/\epsilon^3)$ rate for stochastic Softmax PG with exponentially decreasing step-sizes (Lu et al., 2024). We remark that this is the first stochastic Softmax PG algorithm to obtain ϵ-convergence to the optimal policy while using entropy regularization. Unlike in previous work (Ding et al., 2021), oracle-like knowledge of the environment is not necessary to obtain convergence while using entropy regularization in the stochastic setting.

In the next section, we will compare the multi-stage method with baseline methods in the bandit setting. To investigate whether entropy regularization is indeed useful, we will consider both uniform and "bad" initialization.

3.4.1 Experimental Evaluation

We evaluate the methods in multi-armed bandit environments with $A = 10$ in stochastic settings. For each environment, we compare the various algorithms based on their expected sub-optimality gap $\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r]$. We plot the average and 95% confidence interval of the expected sub-optimality gap across 25 independent bandit instances over $T = 10^6$ iterations. To counteract the randomness of each algorithm, for each bandit instance, we additionally run each algorithm 5 times. In total, for each algorithm, the corresponding plot is comprised of 125 runs. To investigate whether regularization of the entropy is helpful in escaping flat regions, we consider uniform and "bad" initialization. For experiments with uniform initialization, the initial policy is uniform, i.e. $\pi_{\theta_0}(a) = 1/A$ for all $a \in \mathcal{A}$. For experiments with bad initialization, the initial policy favors the worst arm, i.e. $\theta_0(a') = 9$ ($\pi_{\theta_0}(a') \approx 0.999$), where $a' := \arg\min_a r(a)$.

Environment Details

Each environment's underlying reward distribution is either a Bernoulli, Gaussian, or Beta distribution with a fixed mean reward vector $r \in \mathbb{R}^A$ and support $[0, 1]$. The difficulty of the environment is determined by the maximum reward gap $\bar{\Delta} := \max_a r(a^*) - r(a)$. In easy environments $\bar{\Delta} = 0.5$ and in hard environments $\bar{\Delta} = 0.1$. For each environment, r is randomly generated for each run.

Methods

We compare the presented stochastic Softmax PG multi-stage algorithm (Algorithm 2) (`SPG-E-MS`) to stochastic Softmax PG (`SPG-ESS`) and stochastic Softmax PG with entropy regularization (`SPG-E-ESS`) with exponentially decreasing step-sizes and when using the "doubling" trick (`SPG-ESS [D]`). We also compare with prior work that uses the full gradient (`SPG-O-G`) (Mei et al., 2021a) and the reward gap (`SPG-O-R`) (Mei et al., 2023) when setting the step-size. For `SPG-ESS` and `SPG-ESS [D]`, we select $\beta = 1$ and $\eta_0 = \frac{1}{18}$. For `SPG-E-ESS` we fix $\tau = 0.1$ and similarly select $\beta = 1$ and $\eta_0 = \frac{1}{L^\tau} = \frac{2}{5+10\,\tau\,(1+\log A)}$. Finally, for `SPG-E-MS`, we observed that the number of iterations T_i at each stage derived by Lemma 13 for the stochastic multistage algorithm is loose due to the exponentially-decreasing step-size analysis. Furthermore, in the exact setting, we observe that when $p = 1$, the number of iterations doubles after each stage. Therefore, instead of using the theoretical number of iterations at each stage, we use the "doubling trick" (Lu et al., 2024). For `SPG-E-ESS` set the hyper-parameters $T_1 = 5000, \tau_0 = 0.5$, $B_1 = 1$ by employing a grid-search on a separate validation

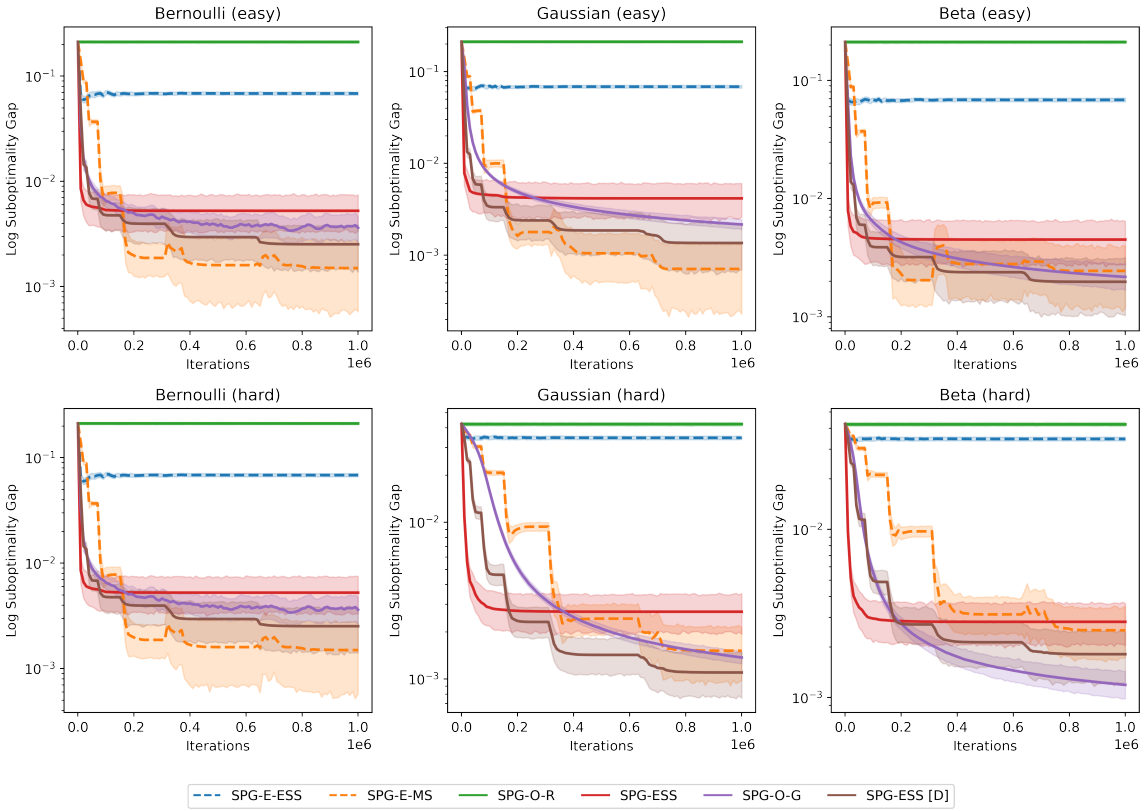


Figure 3.2: Expected sub-optimality gap across various environments with uniform initialization

set of bandit instances. To fairly compare against `SPG-ESS` and `SPG-ESS [D]` we also select $\beta = 1$.

Results

From Figure 3.2, with uniform initialization, the performance of `SPG-E-MS` is comparable to `SPG-ESS`, `SPG-ESS [D]` and `SPG-O-G`. However, in the "bad" initialization settings (Figure 3.3), due to the presence of entropy, `SPG-E-MS` out preforms all other methods. Here we also find that entropy regularization helps escaping from flat regions in the stochastic setting. Since `SPG-E-ESS` uses a fixed entropy regularization term, it is unable to converge to the optimal policy.

3.5 Discussion

We proposed a systematic method for the (stochastic) softmax policy gradient (PG) to utilize the benefits of entropy regularization while guaranteeing convergence to the optimal policy. Under Assumption 1, our proposed multi-stage algorithm achieves convergence to the

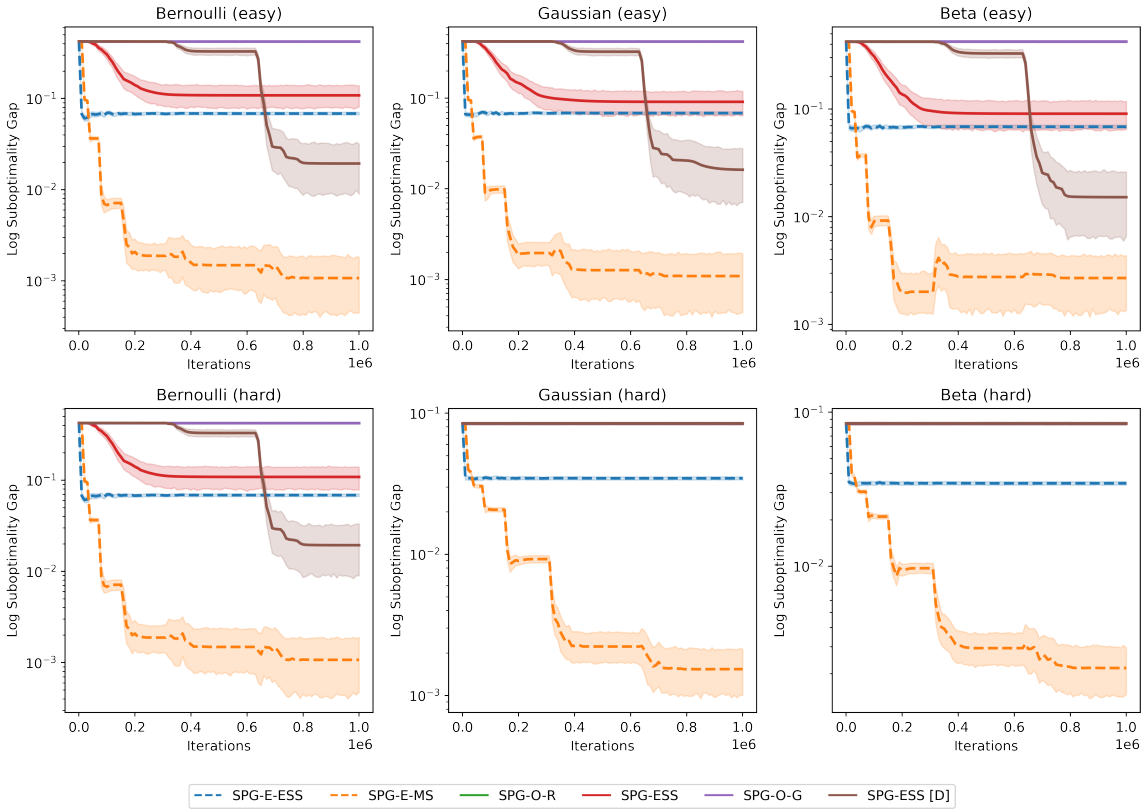


Figure 3.3: Expected sub-optimality gap across various environments with "bad" initialization

optimal policy without any oracle-like knowledge compared to prior methods. We empirically demonstrate that our multi-stage algorithm can escape flat regions in the exact and stochastic settings, due to entropy regularization.

A key direction for future research is to bridge the gap between the non-uniform Łojasiewicz conditions as the entropy regularization strength τ approaches zero. Investigating how the Łojasiewicz degree transitions from $\xi = 1/2$ to $\xi = 0$ could refine theoretical convergence guarantees and enhance our understanding of the interplay between entropy regularization and policy optimization. Additionally, extending these approaches to more general MDPs and exploring their applicability to other policy gradient methods remain promising avenues for further inquiry.

Chapter 4

Linear Softmax Policy Gradient

4.1 Introduction

When dealing with large-scale or continuous action spaces, policy optimization in reinforcement learning often relies on function approximation to generalize across similar actions. In this chapter, we focus specifically on the convergence behavior of the Softmax Policy Gradient (PG) method under linear approximation in the multi-armed bandits. We establish a surprising result: Softmax PG converges globally whenever there exists an adequate linear function that both preserves the ordering of the ground-truth reward vector and satisfies certain geometric feature conditions. This requirement is strictly weaker than zero approximation error, as even large approximation error can admit global convergence if the reward ordering is maintained and the features meet these geometric constraints.

Understanding the behavior of PG methods under function approximation is crucial for describing the behavior of RL in practice since one rarely faces domains small enough to explicitly enumerate over states and actions in parameterizing the policy. It is well known that standard Softmax PG converges to stationary points if a "compatible" function approximation is used (Sutton et al., 2000); i.e., one that is able to exactly represent policy value functions. However, when exact policy values are non-realizable, the "approximation error" is typically considered to be the key quantity for characterizing how well a function approximation captures relevant problem quantities, including transition dynamics, rewards, and policy values. This chapter shows that such an approximation error perspective is *overly demanding* when attempting to characterize the conditions that lead to global convergence of PG methods.

Using the concept of approximation error, global convergence results for PG methods have been recently established in an additive form,

$$\text{sub-optimality gap} \leq \text{optimization error} + \text{approximation error}, \tag{4.1}$$

implying that if the approximation error is small, a diminishing optimization error implies a small sub-optimality gap. A representative result is the global convergence of the natural

policy gradient (NPG) (Agarwal et al., 2021, Table 2), where the optimization error will diminish as the algorithm updates. There have also been global convergence results for other PG variants under linear function approximation that follow a similar approximation error analysis (Agarwal et al., 2020; Cayci et al., 2021; Chen et al., 2022; Yuan et al., 2022a; Alfano and Rebeschini, 2022; Abbasi-Yadkori et al., 2019a,b). However, an additive bound like Equation (4.1) has the inherent weakness that the approximation error will never be zero if the function approximation is not able to perfectly represent the desired quantities. This prevents such a strategy from establishing global convergence in cases where the approximation error is non-zero but a PG method still reaches the best representable solution.

Therefore, in spite of this recent progress, the use of approximation error in PG global convergence with function approximations has left two major gaps in the literature. First, it has not been investigated whether a small approximation error is *necessary* to achieve convergence to an optimal representable policy (Agarwal et al., 2021), diverting attention from feature designs that achieve useful properties beyond the small approximation error. Second, it is not clear whether the standard Softmax PG converges globally under small approximation errors. In particular, NPG contains a least squares regression step (Agarwal et al., 2021, Eq. (17)) that can be naturally characterized with an approximation error quantity. However, the standard Softmax PG does not have such a projection step (Sutton et al., 2000), and the results in (Agarwal et al., 2021) do not apply to this update. Whether standard Softmax PG can achieve global convergence with even linearly realizable rewards (zero approximation error) is still an open problem.

In this chapter, we address the above questions and contribute the following results. First, we provide negative answers to questions on the role of approximation error in determining global convergence of the standard Softmax PG update:

(i) Global convergence can be achieved under linear function approximation with non-zero approximation error.

(ii) The approximation error is not a key quantity for characterizing global convergence.

Second, these results lead us to the question whether approximation error is an appropriate quantity to consider the global convergence of the standard Softmax PG update under linear function approximation. We establish new general results that characterize the conditions for global convergence:

(i) We show that the global convergence of Softmax PG follows if the representation (i.e., the feature matrix) preserves the ranking of the rewards *and* satisfies a suitable geometric feature condition. These assumptions go well beyond exact (or even approximate) realizability of the reward vector, thereby highlighting that approximation error alone does not determine global convergence.

(ii) As a byproduct, we resolve an open question by showing that Softmax PG can fail to converge globally if any of the conditions mentioned above are violated even when the reward is perfectly realizable.

(iii) We extend these convergence guarantees to the *stochastic* setting, showing that the Softmax PG converges almost surely to a globally optimal policy under mild conditions analogous to those in the deterministic case.

4.2 Setting and Background

We study the policy optimization problem for stochastic K-armed bandits (Lattimore and Szepesvári, 2020) specified by a true mean reward vector $r \in \mathbb{R}^K$, where for each action $a \in [K]$,

$$r(a) = \int_{-R_{\max}}^{R_{\max}} x \, P_a(x) \mu(dx),$$

where $R_{\max} > 0$ is the reward range, μ is a finite measure over $[-R_{\max}, R_{\max}]$, and $P_a(x) \geq 0$ is the probability density function with respect to μ. Let R_a denote the reward distribution for the action a defined by the density P_a and the base measure μ. The objective is to find a parametric policy $\pi_\theta \in [0,1]^K$ to maximize the expected reward.

$$\sup_{\theta \in \mathbb{R}^d} \langle \pi_\theta, r \rangle, \tag{4.2}$$

where $\theta \in \mathbb{R}^d$ with $d < K$ is the parameter, and $\pi_\theta = \text{softmax}(X\theta)$ is called a "log-linear policy" (Agarwal et al., 2021; Yuan et al., 2022a) such that for all action $a \in [K] \coloneqq \{1, 2, \ldots, K\}$,

$$\pi_\theta(a) = \text{softmax}([X\theta](a)) = \frac{\exp([X\theta](a))}{\sum_{a' \in [K]} \exp([X\theta](a'))}, \tag{4.3}$$

where $X \in \mathbb{R}^{K \times d}$ is the feature matrix with full column rank $d < K$.

There are two major difficulties with the policy optimization problem. First, Equation (4.2) is a non-concave maximization w.r.t. θ, due to the softmax transform (Mei et al., 2020b, Proposition 1). Second, the policy and reward can be unrealizable, in the sense that the parametric log-linear policy $\pi_\theta = \text{softmax}(X\theta)$ cannot well approximate every policy π in the K-dimensional probability simplex, and the score $X\theta \in \mathbb{R}^K$ cannot well approximate the true mean reward $r \in \mathbb{R}^K$. Such limitations arise in the linear function approximation case because π_θ and $X\theta$ are restricted to low-dimensional manifolds via $\theta \in \mathbb{R}^d$ for $d < K$.

As a test bed to analyze the necessary and sufficient conditions for global convergence, we will first consider the exact setting, where the true reward vector is known. To solve Equation (4.2), we consider the standard Softmax PG (Sutton et al., 2000) method. Softmax PG

is an instance of gradient ascent, obtained by the chain rule,

$$\frac{d\,\langle \pi_{\theta_t}, r\rangle}{d\theta_t} = \frac{d\,X\theta_t}{d\theta_t}\left(\frac{d\,\pi_{\theta_t}}{d\,X\theta_t}\right)^{\top} \frac{d\,\langle \pi_{\theta_t}, r\rangle}{d\pi_{\theta_t}} = X^{\top}(\mathrm{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^{\top})\; r. \tag{4.4}$$

As a representative policy-based method, in its general form, Softmax PG lays the foundation for widely used RL methods, including REINFORCE (Williams, 1992), actor-critic (Konda and Tsitsiklis, 1999; Bhatnagar et al., 2009; Haarnoja et al., 2018), TRPO and PPO (Schulman et al., 2015, 2017).

In the exact setting, we are given the true reward vector r at each iteration. Using Equation (4.4), we can have Softmax PG for the Exact Linear Bandits as shown in Algorithm 3.

Algorithm 3 Softmax PG for Linear Bandits

Input: Initial parameters $\theta_1 \in \mathbb{R}^d$, learning rate $\eta > 0$
Output: Policies $\pi_{\theta_t} = \mathrm{softmax}(X\theta_t)$
while $t \geq 1$ **do**
  $\theta_{t+1} \leftarrow \theta_t + \eta\, X^{\top}(\mathrm{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^{\top})r$
end while

To understand the difficulty of the optimization problem in Equation (4.2), it is helpful to consider previous work that has analyzed the convergence of PG methods.

In the tabular setting, where $d = K$, $X = \mathbf{I}_d$, and $\pi_\theta = \mathrm{softmax}(\theta)$ with $\theta \in \mathbb{R}^K$, both the rewards and optimal policy can be arbitrarily well approximated. In this case, it is known that Softmax PG achieves global convergence asymptotically, i.e., $\langle \pi_{\theta_t}, r\rangle \to r(a^*)$ as $t \to \infty$ (Agarwal et al., 2021), with an $O(1/T)$ rate of convergence that exhibits undesirable problem and initialization dependent constants (Mei et al., 2020a; Li et al., 2021a). Directly extending this global convergence result to the case of function approximation, i.e., log-linear policies, is impossible without any additional assumptions on the features, since there can be exponentially many sub-optimal local maxima in the worst case (Chen et al., 2019). In fact, even with linearly realizable rewards (zero approximation error), whether the standard Softmax PG achieves global convergence still remains unsolved (Agarwal et al., 2021). One intuitive reason why this is a difficult result to establish is that the standard Softmax PG uses the gradient Equation (4.4) rather than projection (regression) to perform updates, which is less directly connected to the concept of approximation error.

4.3 The Limitations of Approximation Error in Characterizing Convergence

It is known that there exist representations $X \in \mathbb{R}^{K\times d}$ with $d < K$ and $r \in \mathbb{R}^K$ that create exponentially many sub-optimal local maxima in Eq. (4.2) (Chen et al., 2019, Theorem 1),

which makes it impossible to ensure global convergence of PG methods without imposing any structure on the function approximation. Before identifying specific conditions that ensure global convergence, we first explain how approximation error cannot be a useful structural measure for this purpose, by demonstrating that zero approximation error is not a necessary condition for global convergence, and illustrating problem instances with comparable approximation error that render starkly different convergence behaviors across different PG methods. Specifically, we illustrate these points with a set of concrete scenarios, each with 4 actions and 2-dimensional feature vectors describing each action. Since $d < K$, not every policy can be expressed in these representations, hence the problem instances are unrealizable.

4.3.1 Global Convergence is Achievable with Non-zero Approximation Error

The results of (Chen et al., 2019, Theorem 1) do not imply that sub-optimal local maxima always appear, as shown in the following.

Example 1. *$K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ and $r = (9, 8, 7, 6)^\top$. The approximation error is $\epsilon_{approx} = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2 = \left\|X \left(X^\top X\right)^{-1} X^\top r - r\right\|_2 = \sqrt{202.6} \approx 14.2338$.*

The approximation error is larger than any sub-optimality gap, i.e., for any policy π,

$$\langle \pi^* - \pi, r \rangle \leq 3 < \epsilon_{\text{approx}},$$

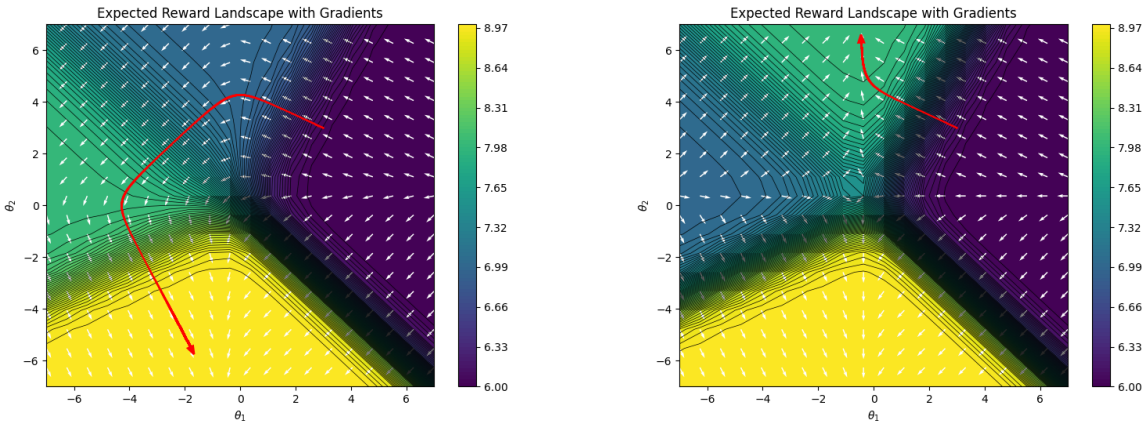
Despite the non-zero approximation error, Algorithm 3 can be shown to reach a global maximum.

Proposition 3. *Denote $a^* \coloneqq \arg\max_{a \in [K]} r(a)$. With constant $\eta > 0$ and any initialization $\theta_1 \in \mathbb{R}^d$, Algorithm 3 guarantees $\langle \pi_{\theta_t}, r \rangle \to r(a^*)$ as $t \to \infty$ on Example 1.*

The fact that Softmax PG achieves global convergence in Example 1 is harder to establish since Equation (4.4) involves a complex non-linearity given the presence of the softmax. To illustrate the intuition behind Proposition 3 we use a visualization of the optimization landscape.

Visualization. A visualization of the optimization landscape of Example 1 is shown in Figure 4.1(a). The colors visualize the expected reward $\langle \pi_\theta, r \rangle$ over the parameter space $\mathbb{R}^d$ where $d = 2$. For each $\theta \in \mathbb{R}^d$, we calculate π_θ using Equation (4.3) and $\langle \pi_\theta, r \rangle$ using Equation (4.2).

To verify Proposition 3, we run Softmax PG on Example 1 with $\theta_1 = (3, 3)^\top \in \mathbb{R}^2$. In Figure 4.1(a), the optimization trajectories show 10^4 iterations of Softmax PG, with a learning rate $\eta = 0.2$. It can be clearly seen that Softmax PG eventually achieves the expected reward $\langle \pi_{\theta_t}, r \rangle \to 9 = r(a^*)$, demonstrating global convergence.



(a) Algorithm 3 running on Example 1

(b) Algorithm 3 running on Example 2

Figure 4.1: Visualizing the landscapes in the example problem instances.

In summary, Example 1 shows that Softmax PG can achieve global convergence on unrealizable problem instances with non-zero approximation error. This raises the question:

Is non-zero approximation error useful for characterizing global convergence?

4.3.2 Global Convergence is Irrelevant to Non-zero Approximation Error

We answer the above question negatively. By comparing alternative problem instances with similar approximation errors but different convergence behaviors, we illustrate how approximation error is not able to distinguish between scenarios where global versus local convergence is obtained.

Example 2. $K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d\times K}$, *and* $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$.
The approximation error is $\left\| X \left(X^\top X \right)^{-1} X^\top r - r \right\|_2 = \sqrt{205} \approx 14.3178$.

The only difference between Examples 1 and 2 is that the second and third columns of $X^\top$ have been exchanged. The approximation error remains similar to that of Example 1. However, as shown in Example 4.1(b), using the same initialization and learning rate, Softmax PG obtains $\langle \pi_{\theta_t}, r \rangle \to 8 = r(2) < r(a^*)$ as it converges to a sub-optimal deterministic policy.

In summary, Examples 1 and 2 have similar approximation errors, yet Softmax PG achieves global convergence on Example 1 but reaches a bad local maxima on Example 2. Note that these examples can be re-scaled to have exactly the same approximation errors while demonstrating the same convergence behavior of the algorithms. From these findings, we conclude that, if there is any quantity that can predict whether global versus local convergence is obtained by Softmax PG, that quantity cannot be the approximation error alone. This motivates us to investigate the question: what is the right quantity to characterize global convergence for unrealizable problems?

4.4 Global Convergence For Linear Bandits In The Exact Setting

In this section, we analyze the conditions under which the Softmax Policy Gradient (PG) method for linear bandits achieves global convergence in the exact setting where the full gradient can be computed. Our objective is to identify sufficient characteristics of the feature representation and the reward structure to ensure convergence to the optimal policy. To support our analysis, we introduce the following assumptions:

Assumption 7 (Unique True Mean Reward)**.** *For all $i, j \in [K]$, if $i \neq j$, $r(i) \neq r(j)$.*

Remark. Assumption 7 ensures that the mean rewards for all arms are distinct, thus guaranteeing a unique optimal arm. This assumption has been widely used by existing works (Mei et al., 2024a,b) to ensure convergence to strict one-hot policies. Moreover, assuming a unique optimal action simplifies the formulation of subsequent feature-related assumptions. We believe that Softmax PG can work without Assumption 7, and removing it remains an open question for future work.

Assumption 8 (Reward Ordering Preservation)**.** *There exists a $w \in \mathbb{R}^d$ such that $r' = X\,w$ preserves the ordering of the reward r, that is, $r'(i) > r'(j)$ if and only if $r(i) > r(j)$.*

Remark. Assumption 8 implies that the feature representation X is expressive enough to retain the relative ordering of the true rewards through a linear transformation. Note that this condition is weaker than requiring the exact realization of the true rewards.

Intuition. Consider Example 1, where Softmax PG achieves global convergence. From the landscape shown in Figure 4.1(a), there appears to be a monotonic path from any initialization point that allows the gradient ascent to reach the optimal plateau with reward $r(a^*) = 9$. Intuitively, this arises because the rewards of the actions seem to be nicely "ordered". For example, starting from $\theta_1 = (6, 8)^\top \in \mathbb{R}^d$ such that $\langle \pi_{\theta_1}, r \rangle \approx 6$, Softmax PG is able to improve its expected reward eventually to $\langle \pi_{\theta_t}, r \rangle \approx 7$, since there exists a suboptimal plateau with higher reward 7 right beside the lowest plateau with reward 6. Next, Softmax PG continues to improve its expected reward eventually to $\langle \pi_{\theta_t}, r \rangle \approx 8$ by "climbing" toward another neighboring plateau with a higher reward. Finally, this process ends with the Softmax PG successfully reaching the optimal plateau with reward $r(a^*) = 9$.

In contrast, in Example 2, as shown in Figure 4.1(b), Softmax PG is stuck on a bad plateau with a local maximum reward of 8. Visually, Softmax PG stops improving its expected reward on this suboptimal plateau, because it is "surrounded" by two lower plateaus with rewards 6 and 7, which breaks the nice "ordering" of the expected reward landscape and traps the gradient ascent trajectory on a suboptimal plateau from which there is no monotonic ascent to global optimality.

Verifying reward order preservation. Based on the above intuition and observations, we conjecture that the ordering structure between the different rewards is a key property behind the global convergence of Softmax PG. We can verify this conjecture in each of Examples 1 and 2 by determining whether the feature matrix $X \in \mathbb{R}^{K\times d}$ allows the same action ordering as the reward vector $r \in \mathbb{R}^K$. For Example 1, note that with $w = (-1,-1)^\top \in \mathbb{R}^d$, we have

$$r' \coloneqq Xw = (2,1,-1,-2)^\top \in \mathbb{R}^K,$$

which preserves the ordering of $r \in \mathbb{R}^K$, such that for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$. In this example, Softmax PG converges to a globally optimal reward.

In contrast, for Example 2, it is impossible to find any $w \in \mathbb{R}^d$ such that Xw preserves the order of rewards r. To see why, consider any $w = (w(1), w(2))^\top$ and note that

$$r' \coloneqq Xw = (-2\cdot w(2), w(2), -w(1), 2\cdot w(1))^\top.$$

To preserve the reward order, we require both $-2\cdot w(2) > w(2)$ (which would imply $w(2) < 0$) and $-w(1) > 2\cdot w(1)$ (which would imply $w(1) < 0$), but these two conditions imply $w(2) < 0 < -w(1)$, which must reverse the order of the second and third actions. This is an example where PG can fail to reach a global optimum.

Under the above assumptions, we establish the monotonic improvement property of the Softmax PG method.

Lemma 1. *Assuming Assumptions 7 and 8 are satisfied, using Algorithm 3 with the following constant learning rate,*

$$0 < \eta < \frac{4}{9\,\|r\|_\infty\,\lambda_{\max}(X^\top X)}, \tag{4.5}$$

ensures (i) $\langle \pi_{\theta_{t+1}}, r\rangle > \langle \pi_{\theta_t}, r\rangle$ for all $t \geq 1$ and (ii) $\pi_{\theta_t}(a) \to 1$ for an arm $a \in [K]$ as $t \to \infty$.

Lemma 1 establishes that, under our assumptions and with an appropriate learning rate, the Softmax PG method not only consistently improves the expected reward, but also converges to a deterministic policy focused on a single arm. However, this lemma does not specify which arm the policy converges to, leading us to further investigate the conditions that ensure convergence to the optimal arm from any initialization. To guarantee that the Softmax PG method converges to the optimal policy regardless of initialization, we introduce an additional condition on the feature matrix X.

4.4.1 Warm up: Global Convergence when $K = 3$

Assumption 8 only demands the reward ordering preservation along one arbitrary direction w. This assumption ensures that there are no finite stationary points in the optimization landscape, and therefore, Softmax PG will commit to one of the arms. However, as we will show later, Assumption 8 is not sufficient for ensuring convergence to the optimal arm. A

natural way to strengthen the assumption is to require the features to preserve the reward ordering in more than one direction. Consider a simple case when $\theta \in \mathbb{R}^2$. Assume that there exist two orthogonal directions u and v such that $r^u \coloneqq \langle r, u\rangle$ and $r^v \coloneqq \langle r, v\rangle$ both preserve the ordering of the reward. Then, we can rewrite the feature as: for $i \in [3]$, $x_i = r_i^u u + r_i^v v$. This implies that

$$\begin{aligned}
\langle x_2 - x_3, x_1 - x_3\rangle &= \langle (r_2^u - r_3^u)u + (r_2^v - r_3^v)v, (r_1^u - r_3^u)u + (r_1^v - r_3^v)v\rangle \\
&= (r_2^u - r_3^u)(r_1^u - r_3^u)u^2 + (r_2^v - r_3^v)(r_1^v - r_3^v)v^2 && (\langle u, v\rangle = 0) \\
&> 0 && (r^u \text{ and } r^v \text{ preserve the reward ordering})
\end{aligned}$$

Given the above observations, we state another key feature condition that is required for the guarantee of global convergence.

Assumption 9 (Feature Condition ($K = 3$))**.** *The given feature matrix, X, satisfies that $\langle x_2 - x_3, x_1 - x_3\rangle > 0$.*

The next result shows that in the three-armed bandit setting, the above assumptions are sufficient to ensure convergence to the optimal action.

Theorem 4. Given a reward vector $r \in \mathbb{R}^3$ and a feature matrix $X \in \mathbb{R}^{3\times d}$ such that Assumptions 7, 8, and 9 are satisfied, Algorithm 3 with a constant learning rate as in Eq. 4.5 is guaranteed to converge to the optimal policy.

Refer to Figure B.1a for empirical evaluation of Softmax PG for 3-armed linear bandits. Given Assumptions 7 to 9, we next investigate which assumptions are required for global convergence.

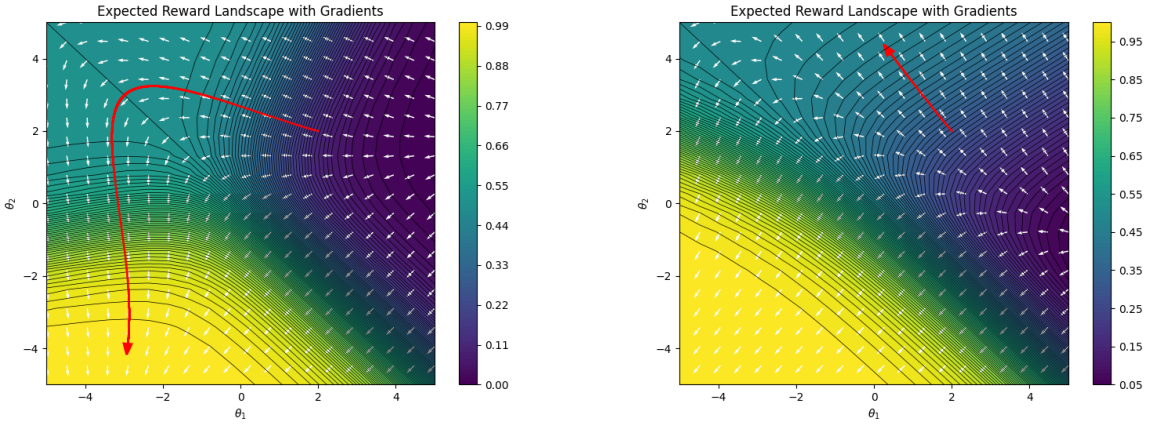
Example 3. *Let $K = 3$ $d = 2$, $X^\top = \begin{bmatrix} 0 & -0.3 & 1 \\ -1 & 0.6 & 0 \end{bmatrix}$ and $r = (1, 0.5, 0)^\top$. Assumption 8 is satisfied since we have $r' = (1, 0, -2)^\top = Xw$ for $w = (-2, -1)^\top$, and Assumption 9 is satisfied since $\langle x_2 - x_3, x_1 - x_3\rangle = 0.7 > 0$.*

In the above example, Algorithm 3 is guaranteed to converge to the optimal policy for any initialization. Furthermore, we can prove that Assumption 9 is a necessary condition for global convergence in 3-armed bandits. By "necessary," we do not claim that a violation of this condition guarantees failure of the algorithm in all cases. Rather, we assert that if this condition is omitted while the others are satisfied, it is always possible to construct a specific counterexample on which the algorithm fails to converge. In other words, each condition is essential in the sense that leaving any one of them out allows for the existence of a problem instance that breaks global convergence.

Proposition 5. *Given a reward vector $r \in \mathbb{R}^3$ and a feature matrix $X \in \mathbb{R}^{3\times d}$ such that Assumptions 7 and 8 are satisfied but Assumption 9 is not. Using Algorithm 3 with a constant learning rate as in Equation (4.5) and initialization $\theta_1 = c\,(x_3 - x_1)$, such that $c > \frac{-\log(m)}{\|x_3 - x_1\|_2^2}$, where $m = \frac{\langle x_3 - x_2, x_1 - x_3\rangle}{\langle x_1 - x_2, x_1 - x_3\rangle} \frac{\langle \pi_{\theta_1}, r\rangle - r(3)}{r(1) - \langle \pi_{\theta_1}, r\rangle}$ fails to converge to the optimal policy.*

The next example is only slightly different from Example 3. In Example 4, setting $c = 2$, results in $\theta_1 = c\,(x_3 - x_1) = [2, 2]^\top$, and $c = 2 > \frac{-\log(m)}{\|x_3 - x_1\|_2^2} \approx 1.61$. This satisfies the condition in Proposition 5, thereby demonstrating that Softmax PG must fail in this specific scenario (Figure 4.2b).

Example 4. *Let* $K = 3$, $d = 2$, $X^\top = \begin{bmatrix} 0 & 0.6 & 1 \\ -1 & 0.6 & 0 \end{bmatrix}$, *and* $r = (1, 0.5, 0)^\top$. *Assumption 8 is satisfied since* $r' = (1, -1.8, -2)^\top = Xw$ *for* $w = (-2, -1)^\top$, *but Assumption 9 is not since* $\langle x_2 - x_3, x_1 - x_3 \rangle = -0.2 < 0$.



(a) Algorithm 3 running on Example 3

(b) Algorithm 3 running on Example 4

Figure 4.2: The effect of feature conditions on convergence

We also show that Assumption 8 is still required, even if Assumption 9 is satisfied, thus reinforcing that *each* of these assumptions is independently necessary in the same sense described above.

Proposition 6. *Let* $K = 3$, $d = 2$, $X^\top = \begin{bmatrix} 3 & 5 & 1 \\ 4 & 6 & 2 \end{bmatrix} \in \mathbb{R}^{d \times K}$, *and* $r = (3, 2, 1)^\top$. *In this case, Assumptions 7 and 9 are satisfied, but Assumption 8 is not, and the features do not allow the optimal reward to be achieved for any set of finite or infinite parameters. Therefore, Algorithm 3 does not achieve global convergence for any initialization* θ_1.

4.4.2 Global Convergence for all $K \geq 3$

To extend the analysis to any number of arms, we generalize the previous feature condition.

Assumption 10 (Feature Conditions (Backward Compatible))**.** *The feature matrix* X *satisfies*

$$\langle x_i - x_j, x_{a^*} - x_k \rangle \begin{cases} > 0 & \text{If } i = a^* \text{ or } j = k \\ \geq 0 & \text{Otherwise} \end{cases}.$$

for any three arms i, j, and k such that $r(i) > r(j)$ and $r(i) > r(k)$.

Under Assumption 10, we prove in the following theorem that Softmax PG can achieve global convergence for $K \geq 3$.

Theorem 7. Given a reward vector $r \in \mathbb{R}^K$ and a feature matrix $X \in \mathbb{R}^{K \times d}$ such that Assumptions 7, 8 and 10 are satisfied,using Algorithm 3 with a constant learning rate as in Equation (4.5) converges to the optimal policy.

Theorem 7 obtains the same convergence guarantee as in Mei et al. (2024a). Unlike Mei et al. (2024a)], we consider a slightly different set of assumptions with a similar level of constraints on the feature matrix. Refer to Figure B.1b for empirical evaluation of Softmax PG for multi-armed linear bandits. In Section 4.5, we show that such assumptions can be generalized to stochastic linear bandits.

Discussion. The assumptions we have introduced play a critical role in guaranteeing the global convergence of the Softmax PG method. Assumption 7 ensures a unique optimal arm, simplifying the convergence analysis. Assumption 8 emphasizes the importance of the feature representation in capturing the correct action ordering. Assumption 10 imposes geometric constraints on features to facilitate consistent progress towards the optimal policy. These assumptions collectively highlight that the structure of the feature representation and its alignment with the reward ordering are essential for the success of the Softmax PG method.

Furthermore, Lemma 1 establishes the foundation for our convergence analysis by ensuring monotonic improvement and convergence to a deterministic policy. However, without the feature conditions, the algorithm may converge to a suboptimal arm, as demonstrated in the examples. Although the monotonicity of the expected reward is guaranteed in the deterministic setting under our assumptions, extending these results to the stochastic setting introduces additional complexities.

In stochastic environments, the observed rewards are noisy estimates of the true mean rewards, and the monotonicity property may not hold at every iteration. This requires careful analysis to ensure convergence, which we explore in the next section.

4.5 Global Convergence For Linear Bandits In The Stochastic Setting

In this section, we extend our analysis to the stochastic setting. Here, the full gradient cannot be computed since the rewards are drawn from unknown distributions and only a single arm is selected in each iteration. We will show that the insights and conditions from the exact setting can be used to ensure global convergence in the presence of stochasticity. We consider the Softmax PG method in the stochastic setting, as outlined in Algorithm 4.

Algorithm 4 Softmax PG for Linear Bandits

Input: Initial parameters $\theta_1 \in \mathbb{R}^K$, learning rate $\eta > 0$
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$
while $t \geq 1$ **do**
  Sample an action $a_t \sim \pi_{\theta_t}(\cdot)$ and observe reward $R_t(a_t) \sim P_{a_t}$
  for all $a \in [K]$ **do**
    if $a = a_t$ **then**
      $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta\ (1 - \pi_{\theta_t}(a))\ R_t(a_t)$
    else
      $\theta_{t+1}(a) \leftarrow \theta_t(a) - \eta\, \pi_{\theta_t}(a)\, R_t(a_t)$
    end if
  end for
end while

Proposition 8. *[Proposition 2.3 of (Mei et al., 2023)] Algorithm 4 is equivalent to the following update:*

$$\theta_{t+1} = \theta_t + \eta\, X^\top (\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top)\, \hat{r}_t,$$

where $\mathbb{E}_t\left[\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right] = \frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t}$*, and* $\mathbb{E}_t[\cdot]$ *is defined with respect to randomness from on-policy sampling* $a_t \sim \pi_{\theta_t}(\cdot)$ *and reward sampling* $R_t(a_t) \sim P_{a_t}$*. The Jacobian of* $\theta \to \pi_\theta := \text{softmax}(X\theta)$ *is defined as* $\left(\frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t}\right)^\top := X^\top(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top) \in \mathbb{R}^{d\times d}$ *and* $\hat{r}_t(a) := \frac{\mathbb{1}\{a=a_t\}}{\pi_{\theta_t}(a)} R_t(a)$ *for all* $a \in [K]$ *is the importance sampling (IS) estimator, and we set* $R_t(a) = 0$ *for all* $a \neq a_t$*.*

Remark. The update rule in Algorithm 4 is an unbiased estimate of the gradient of the expected reward with respect to the parameters θ_t. This stochastic approximation introduces randomness into the optimization process, which must be carefully managed to ensure convergence.

4.5.1 Decomposition of Stochastic Process

To show that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$, we must have, almost surely, $\theta_t(a^*) \to \infty$ and $\theta_t(a) \to -\infty$ for all arms $a \neq a^*$ as $t \to \infty$. To establish this fact, we will consider the stochastic process of a logit $z_t(a) := [X\theta](a) \in \mathbb{R}^K$ for any arm $a \in [K]$. Let us express $z_t(a)$ in terms of the "cumulative noise" and "progress" terms. Define $\mathcal{F}_t$ as the σ-algebra generated by $\{a_1, R_1(a_1), \cdots, a_{t-1}, R_{t-1}(a_{t-1})\}$, i.e.

$$\mathcal{F}_t = \sigma(\{a_1, R_1(a_1), \cdots, a_{t-1}, R_{t-1}(a_{t-1})\})\,.$$

Note that θ_t, z_t, are $\mathcal{F}_t$-measurable and $\hat{r}_t$ is $\mathcal{F}_{t+1}$-measurable for all $t \geq 1$. Let $\mathbb{E}_t$ denote the conditional expectation with respect to $\mathcal{F}_t$, which implies that $\mathbb{E}_t[X] = \mathbb{E}[X|\mathcal{F}_t]$. We define the following notation,

$$W_t(a) := z_t(a) - \mathbb{E}_{t-1}[z_t(a)], \qquad (\text{“noise”})$$
$$P_t(a) := \mathbb{E}_t[z_{t+1}(a)] - z_t(a). \qquad (\text{“progress”})$$

For $a \in [K]$, $t \geq 2$, we have the following decomposition of the stochastic process of $z_t(a)$:

$$z_t(a) = W_t(a) + P_{t-1}(a) + z_{t-1}(a), \tag{4.6}$$

and,

$$z_1(a) = \underbrace{z_1(a) - \mathbb{E}[z_1(a)]}_{W_1(a)} + \mathbb{E}[z_1(a)],$$

where $\mathbb{E}[z_1(a)]$ accounts for possible randomness in initializing $\theta_1 \in \mathbb{R}^d$. By recursing Equation (4.6), we obtain

$$z_t(a) = z_1(a) + \underbrace{\sum_{s=1}^{t-1} P_s(a)}_{\text{“cumulative progress”}} + \underbrace{\sum_{s=1}^{t} W_s(a)}_{\text{“cumulative noise”}}. \tag{4.7}$$

We will compare the logits between any two distinct arms a_1 and a_2 ($a_1 \neq a_2$) to measure which arm will be dominant. Using Equation (4.7), we have

$$z_t(a_1) - z_t(a_2) = z_1(a_1) - z_1(a_2) + \underbrace{\sum_{s=1}^{t-1} [P_s(a_1) - P_s(a_2)]}_{\text{(i)}} + \underbrace{\sum_{s=1}^{t-1} [W_{s+1}(a_1) - W_{s+1}(a_2)]}_{\text{(ii)}}. \tag{4.8}$$

In the following sections, we will use the above decomposition to show that Algorithm 4 achieves global convergence to the optimal policy, almost surely.

4.5.2 Asymptotic Global Convergence

In the stochastic setting, we need to adjust our analysis to account for the randomness introduced by sampling. Similarly to Lemma 1, by setting the learning rate to be sufficiently small, we can have the monotonicity of the expected reward.

Lemma 2. *We set the constant learning rate as:*

$$\eta = \min\left\{ \frac{1}{6\,(\lambda_{\max}[X^\top X])^{3/2}\sqrt{2\,R_{\max}}}, \frac{\lambda_{\min}[X^\top X]}{6\,\rho\,[\lambda_{\max}[X^\top X]]^2} \right\}, \tag{4.9}$$

where $\rho := \frac{8\, R_{\max}^3 K^{3/2}}{\Delta^2}$, $\Delta := \min_{i \neq j} |r(i) - r(j)|$, $\kappa := \frac{\lambda_{\max}[X^\top X]}{\lambda_{\min}[X^\top X]}$ *is the condition number of* $X^\top X$, *and* $\mu := [\mathbb{E}[\inf_{t \geq 1} [\pi_\theta(a^*)]^{-2}]]^{-1} > 0$. *Algorithm 4 with the above learning rate assures that, for all* $t \geq 1$,

$$\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] - \langle \pi_{\theta_t}, r \rangle \geq \frac{1}{6\, \rho\, \kappa^2} \left\| J(z_t) \right\|_2^2.$$

Lemma 2 indicates that, in expectation, the algorithm makes progress towards increasing the expected reward. However, the presence of stochasticity means that the reward may not increase at every iteration, and additional techniques are required to handle the variance in the updates. Moreover, with the same learning rate, Algorithm 4 can converge to a one-hot policy almost surely.

Lemma 3. *Using Algorithm 4 with a constant step-size as in Equation* (4.9) *will converge to a one-hot policy (i.e. there exists an (possibly random) arm* $k \in [K]$ *such that* $\pi_{\theta_t}(k) \to 1$ *as* $t \to \infty$*) almost surely.*

Using the same feature conditions, we show that

Theorem 9. Given a reward vector $r \in \mathbb{R}^K$ and a feature matrix $X \in \mathbb{R}^{K \times d}$ such that Assumptions 7 and 10 are satisfied, Algorithm 4 with the constant learning rate as in Equation (4.9), we have, almost surely, $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$.

By setting $d = K$ and $X = \mathbf{I}_d$ (that is, reducing to the tabular setting), we can recover the same convergence guarantee as in Mei et al. (2023).

4.5.3 Rates of Convergence Convergence

Using the learning rate as in Equation (4.9), Algorithm 4 can achieve a sub-linear convergence rate of $O(1/T)$.

Theorem 10. Given a reward vector $r \in \mathbb{R}^K$ and a feature matrix $X \in \mathbb{R}^{K \times d}$ such that Assumptions 7 and 10 are satisfied, Algorithm 4 with the constant learning as in Equation (4.9) results in the following sub-linear convergence rate:

$$\mathbb{E}[\langle \pi^*, r \rangle - \langle \pi_{\theta_1}, r \rangle] \leq \frac{6\, \rho\, \kappa^2}{\mu\, T},$$

where $\rho := \frac{8\, R_{\max}^3 K^{3/2}}{\Delta^2}$, $\kappa := \frac{\lambda_{\max}[X^\top X]}{\lambda_{\min}[X^\top X]}$, and $\mu := [\mathbb{E}[\inf_{t \geq 1} [\pi_{\theta_t}(a^*)]^{-2}]]^{-1} > 0$.

4.6 Discussion

We believe that this work opens up new directions for understanding PG-based methods under function approximation, going well beyond the conventional approximation error-based

analysis. It identifies ordering-based conditions that guarantee global convergence in both the deterministic and stochastic settings.

Extending the results and techniques to general MDPs is an important and challenging next step. Investigating how these global convergence conditions can be used to achieve better representation learning is of great interest for algorithm design. Generalizing the proof techniques to other scenarios where non-linear transforms (activation functions) interact with low-dimensional features through gradient descent, such as in neural networks, is another direction of future work.

Chapter 5

Conclusion

The first part of this thesis analyzed the softmax policy gradient method with entropy regularization, addressing both exact and stochastic settings. We proposed a multi-stage algorithm, which iteratively decays the entropy term and avoids reliance on problem-dependent constants. Our findings demonstrated that:

- Entropy regularization improves the optimization landscape, enabling the policy to escape flat regions and improving convergence behavior.
- Systematically decaying the entropy term prevents bias and guarantees convergence to the true optimal policy without requiring oracle-like knowledge.
- Exponentially decreasing step-sizes can be effectively employed to further enhance the practicality of the approach in stochastic environments.

The second part of this thesis examined the softmax policy gradient method under linear function approximation in the bandit setting. We showed that:

- Global convergence of Softmax PG can occur even with non-zero approximation error, provided that the feature representation preserves the ordering of rewards and satisfies certain geometric conditions.
- Specific feature conditions guarantee convergence to the globally optimal policy, while slight violations can lead to suboptimal outcomes.
- Our theoretical analysis extends to the stochastic setting, showing that under similar feature conditions and with appropriately chosen learning rates, Softmax PG converges almost surely to the globally optimal policy even in the presence of noise.

Combining the Two Parts. Although each line of work was treated separately in this thesis, we believe that combining entropy regularization with linear softmax policy gradient is a promising next step. Such a unified approach could potentially leverage the benefits of both

lines of research, promoting more effective exploration through entropy while still accommodating large-scale or continuous action spaces via linear approximation. However, merging these methods entails considerable technical work and we leave a detailed investigation of this direction as an avenue for future work.

We hope that our results will inform the design of more robust, scalable reinforcement learning algorithms capable of handling practical, high-dimensional problems and will motivate further research into the interplay between entropy regularization and linear function approximation.

Bibliography

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019a.

Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*, 2019b.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR, 2019.

Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Semih Cayci, Niao He, and Rayadurgam Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70 (4):2563–2578, 2022.

Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. *Advances in Neural Information Processing Systems*, 32, 2019.

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

Yuhao Ding, Junzi Zhang, Hyunin Lee, and Javad Lavaei. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021a.

Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pages 6553–6564. PMLR, 2021b.

Michael Lu, Matin Aghaei, Anant Raj, and Sharan Vaswani. Towards principled, practical policy gradient for bandits and tabular mdps. *arXiv preprint arXiv:2405.13136*, 2024.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33:21130–21140, 2020a.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020b.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a.

Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pages 7555–7564. PMLR, 2021b.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 35:17818–17830, 2022a.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 35:17818–17830, 2022b.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24325–24360. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/mei23a.html`.

Jincheng Mei, Bo Dai, Alekh Agarwal, Mohammad Ghavamzadeh, Csaba Szepesvári, and Dale Schuurmans. Ordering-based conditions for global convergence of policy gradient methods. *Advances in Neural Information Processing Systems*, 36, 2024a.

Jincheng Mei, Bo Dai, Alekh Agarwal, Sharan Vaswani, Anant Raj, Csaba Szepesvári, and Dale Schuurmans. Small steps no more: Global convergence of stochastic gradient bandits for arbitrary learning rates. *Advances in Neural Information Processing Systems*, 2024b.

B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(63)90382-3. URL `https://www.sciencedirect.com/science/article/pii/0041555363903823`.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad. Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In *International Conference on Machine Learning*, pages 22015–22059. PMLR, 2022.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022a.

Rui Yuan, Robert M. Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient, 2022b.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34: 25746–25759, 2021.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020a.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020b.

Supplementary Material

Organization of the Appendix

Appendix A

Proofs of Chapter 3

A.1 Definitions

A function f is L-smooth if for all θ and θ'

$$|f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle| \leq \frac{L}{2} \|\theta - \theta'\|_2^2. \tag{A.1}$$

A function f is L_1-non-uniform smooth if for all θ and θ'

$$|f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle| \leq \frac{L_1 \|\nabla f(\theta')\|}{2} \|\theta - \theta'\|_2^2. \tag{A.2}$$

A function f satisfies the non-uniform Łojasiewciz condition of degree ξ for $\xi \in [0, 1]$ is defined as

$$\|\nabla f(\theta)\| \geq C(\theta) |f^* - f(\theta)|^{1-\xi} \qquad (f^* := \sup_\theta f(\theta))$$

where $C : \theta \to \mathbb{R} > 0$.

A function f satisfies the reversed Łojasiewciz condition if for all θ

$$\|\nabla f(\theta)\| \leq \nu \left[f^* - f(\theta) \right] \tag{A.3}$$

where $\nu > 0$.

A.2 Proofs of Section 3.3

A.2.1 Proof of Theorem 1

Theorem 1. Assuming f^τ and f satisfy Assumptions 1 to 5, for a given $\epsilon \in (0,1)$, Algorithm 1 achieves ϵ-suboptimality to the global optimal policy after $T_{\text{total}} = \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1} \log\left(2\,(1+B_4)\right)$ iterations, where $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3$.

Proof. Observe that in Algorithm 1, we use τ_i and η_i at stage $i \geq 1$, which starts at iteration $\text{last}_{i-1}+1$, runs for $T_i = \frac{2}{\eta_i\,\mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)$ iterations, and ends at iteration last_i. Now, we prove by induction that $f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ for all $i \geq 0$:
Base Case: For $i = 0$, we have

$$f^{*\tau_0} - f^{\tau_0}(\theta_0) \leq \max(\tau_0, f^{*\tau_0} - f^{\tau_0}(\theta_0)) = \tau_0 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{A.4}$$

Induction Step: Suppose $f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}}) \leq \tau_{i-1} \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds.

Since $f^{\tau_i}(\theta)$ is L^{τ_i}-smooth and satisfies the non-uniform Łojasiewciz condition with $\mu_i := \inf_{t\geq 1} C_\tau^2(\theta_t)$, we use Lemma 4 for stage i:

$$f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \exp(-\frac{\eta_i\,\mu_i}{2}\,T_i)[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})] \tag{A.5}$$

If $T_i \geq \frac{2}{\eta_i\,\mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)$, we have

$$= \frac{f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})}{\exp\left(\log\left(\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)\right)\right)} \tag{A.6}$$

Under Assumption 5

$$\leq \frac{f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}}) + \tau_{i-1}B_4}{\frac{\tau_{i-1}}{\tau_i}\,(1+B_4)} \tag{A.7}$$

Using the inductive hypothesis

$$\leq \frac{\tau_i\,\tau_{i-1}\left(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_4\right)}{\tau_{i-1}\,(1+B_4)} \tag{A.8}$$

$$\leq \frac{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)(1+B_4)}{1+B_4} \tag{A.9}$$

$$= \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{A.10}$$

Therefore, for all $i \geq 0$

$$f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{A.11}$$

Define $\epsilon_i := f^* - f(\theta_{\text{last}_i})$ as the sub-optimality at the end of stage i. We have

$$\epsilon_i = f^* - f(\theta_{\text{last}_i}) \tag{A.12}$$

$$= \Big[f^* - f(\theta^*_{\tau_i})\Big] + \Big[f(\theta^*_{\tau_i}) - f(\theta_{\text{last}_i})\Big] \tag{A.13}$$

Under Assumption 4

$$\leq \Big[f^* - f(\theta^*_{\tau_i})\Big] + f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i}) + \tau_i B_3 \tag{A.14}$$

By Equation (A.11),

$$\leq \Big[f^* - f(\theta^*_{\tau_i})\Big] + \tau_i \bigg(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_3\bigg) \tag{A.15}$$

Using Assumption 3,

$$\leq \tau_i\, B_2 + \tau_i \bigg(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_3\bigg) \tag{A.16}$$

$$= \tau_i \underbrace{\left(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3\right)}_{:=C_1} \tag{A.17}$$

$$= 2^{-i}\, \tau_0\, C_1. \qquad (\tau_i = 2^{-i}\, \tau_0)$$

Therefore, the number of stages N_{stages} required to obtain an ϵ sub-optimality is given as:

$$2^{N_{\text{stages}}} \geq \frac{\tau_0\, C_1}{\epsilon} \implies N_{\text{stages}} \geq \log_2\left(\frac{\tau_0\, C_1}{\epsilon}\right). \tag{A.18}$$

On the other hand, the sufficient number of iterations at stage i is:

$$T_i \geq \frac{2}{\eta_i\, \mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1 + B_4)\right) \tag{A.19}$$

Since $\eta_i = \frac{1}{L^{\tau_i}}$

$$= \frac{2\, L^{\tau_i}}{\mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1 + B_4)\right), \tag{A.20}$$

Since $L^{\tau_i} \leq L^{\max}$, it is sufficient to set T_i as:

$$T_i = \frac{2\, L^{\max}}{\mu_i} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1 + B_4)\right) \tag{A.21}$$

Under Assumption 1, $\mu_i = \tau_i^p B_1$

$$= \frac{2\, L^{\max}}{\tau_i^p\, B_1} \log\left(\frac{\tau_{i-1}}{\tau_i}\,(1 + B_4)\right) \tag{A.22}$$

Since $\tau_i = 2^{-i}\,\tau_0$, we have

$$= \frac{2\,L^{\max}\,2^{ip}}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.23}$$

Consequently, we can calculate the sufficient total number of iterations T_{Total} in terms of ϵ:

$$T_{\text{Total}} \geq \sum_{i=1}^{N_{\text{stages}}} T_i = \sum_{i=1}^{N_{\text{stages}}} \left[\frac{2\,L^{\max}\,2^{ip}}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right)\right] \tag{A.24}$$

$$= \frac{2\,L^{\max}\,\sum_{i=1}^{N_{\text{stages}}}(2^p)^i}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.25}$$

Since for all $x > 1, n \geq 0$, $\sum_{i=0}^{n} x^i = \frac{x^{n+1}-1}{x-1}$

$$= \frac{2\,L^{\max}\,\left[\frac{(2^p)^{N_{\text{stages}}+1}-1}{2^p-1} - 1\right]}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.26}$$

Therefore, it is sufficient that

$$T_{\text{Total}} \geq \frac{2\,L^{\max}\,\frac{(2^p)^{N_{\text{stages}}+1}}{2^p-1}}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.27}$$

$$= \frac{2\,L^{\max}\,\frac{2^p\,(2^p)^{N_{\text{stages}}}}{2^p-1}}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.28}$$

Since $p \geq 1$, we have $\frac{2^p}{2^p-1} \leq 2$. Hence, it is sufficient to use

$$T_{\text{Total}} = \frac{4\,L^{\max}\,(2^p)^{N_{\text{stages}}}}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.29}$$

$$= \frac{4\,L^{\max}\,(2^{N_{\text{stages}}})^p}{\tau_0^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.30}$$

Using Equation (A.18),

$$\geq \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1}\log\left(2\left(1+B_4\right)\right) \tag{A.31}$$

in order to guarantee $f^* - f(\theta_{T_{\text{total}}}) \leq \epsilon$. □

Corollary 1. In the bandit setting, assuming for each stage i, $\mu_i = \tau_i^p B_1$ for constants $p \geq 1$ and $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 1 with $\eta_i = \frac{2}{5+10\,\tau_i\,(1+\log A)}$ achieves ϵ-sub-optimality after $T_{\text{total}} = \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1}\log\left(2\left(1+W\left(\frac{A-1}{e}\right)+\log A\right)\right)$ iterations, where $L^{\max} = \frac{5}{2} + 5\,(1+\log A)$ and $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + W\left(\frac{A-1}{e}\right) + \log A$.

Proof. Set $f(\theta) = {\pi_\theta}^\top r$ and $f^\tau(\theta) = {\pi_\theta}^\top(r - \tau\log\pi_\theta)$. We can extend Theorem 1 to the bandit setting since:

- by Lemma 18, f^τ is L^τ-smooth and since $\tau \in [0,1]$

$$\frac{5}{2} = L^{\min} \leq L^\tau = \frac{5}{2} + \tau\, 5\,(1+\log A) \leq \frac{5}{2} + 5\,(1+\log A) = L^{\max} \tag{A.32}$$

- by Lemma 6, we have $f^* - f(\theta_\tau^*) \leq \tau W\left(\frac{A-1}{e}\right)$
- by Lemma 7, we have for all θ, $f(\theta_\tau^*) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau \log A$
- by Lemma 8, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \log A$

□

Corollary 2. In the tabular MDP setting, assuming for each stage i, $\mu_i = \tau_i^p B_1$ for constants $p \geq 1$ and $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 1 with $\eta_i = \frac{(1-\gamma)^3}{8+\tau_i\,(4+8\,\log A)}$ achieves ϵ-sub-optimality after $T_{\text{total}} = \frac{4\,L^{\max}\,C_1^p}{\epsilon^p\,B_1} \log\left(2\left(1+\frac{2\,\log A}{1-\gamma}\right)\right)$ iterations, where $L^{\max} = \frac{12+8\log A}{(1-\gamma)^3}$ and $C_1 = \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + \frac{2\,\log A}{1-\gamma}$.

Proof. Set $f(\theta) = V^{\pi_\theta}(\rho)$ and $f^\tau(\theta) = \tilde{V}_\tau^{\pi_\theta}(\rho)$. We can extend Theorem 1 to the tabular MDP setting since:

- by Lemma 20, $f^\tau(\theta)$ is L^τ-smooth and since $\tau \in [0,1]$

$$L^{\min} = \frac{8}{(1-\gamma)^3} \leq L^\tau = \frac{8+\tau(4+8\log A)}{(1-\gamma)^3} \leq \frac{12+8\log A}{(1-\gamma)^3} = L^{\max} \tag{A.33}$$

- by Lemma 9, we have $f^* - f(\theta_\tau^*) \leq \tau \frac{\log A}{1-\gamma}$
- by Lemma 11, we have for all θ, $f(\theta_\tau^*) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau\,\frac{\log A}{1-\gamma}$
- by Lemma 12, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 \frac{2\log A}{1-\gamma}$

□

Additional Lemmas

Lemma 4. *Assuming f^τ satisfies Assumptions 1 and 2, using Update 3 with $\eta_t = \frac{1}{L^\tau}$, we have*

$$f^{*\tau} - f^\tau(\theta_{t_2}) \leq \exp\left(-\frac{\eta_t\,\mu}{2}\,(t_2 - t_1)\right)\,[f^{*\tau} - f^\tau(\theta_{t_1})] \tag{A.34}$$

where $t_1 < t_2$.

Proof.

Since f^τ is L^τ-smooth

$$f^\tau(\theta_{t+1}) \geq f^\tau(\theta_t) + \langle \nabla f^\tau(\theta_t),\, \theta_{t+1} - \theta_t \rangle - \frac{L^\tau}{2}\|\theta_{t+1} - \theta_t\|_2^2 \tag{A.35}$$

Using Update 3, $\theta_{t+1} = \theta_t + \eta_t \nabla f^\tau(\theta_t)$

$$= f^\tau(\theta_t) + \eta \|\nabla f^\tau(\theta_t)\|_2^2 - \frac{L^\tau {\eta_t}^2}{2} ||\nabla f^\tau(\theta_t)||_2^2 \tag{A.36}$$

Using $\eta_t = \frac{1}{L^\tau}$

$$= f^\tau(\theta_t) + \frac{\eta_t}{2} \|\nabla f^\tau(\theta_t)\|_2^2 \tag{A.37}$$

Assuming Assumption 1 is satisfied, $\|\nabla f^\tau(\theta)\|_2^2 \geq \mu \, |f^{*\tau} - f^\tau(\theta)|$

$$\geq f^\tau(\theta_t) + \frac{\eta \, \mu}{2} \left[f^{*\tau} - f^\tau(\theta_t)\right] \tag{A.38}$$

Multiplying both sides by -1 and adding f^*

$$\implies f^{*\tau} - f^\tau(\theta_{t+1}) \leq \left(1 - \frac{\eta_t \, \mu}{2}\right) \left[f^{*\tau} - f^\tau(\theta_t)\right] \tag{A.39}$$

Using $1 - x \leq \exp(-x)$

$$\leq \exp\left(-\frac{\eta_t \, \mu}{2}\right) \left[f^{*\tau} - f^\tau(\theta_t)\right]. \tag{A.40}$$

Therefore,

$$f^{*\tau} - f^\tau(\theta_{t_2}) \leq \exp\left(-\frac{\eta_t \, \mu}{2} \, (t_2 - t_1)\right) \left[f^{*\tau} - f^\tau(\theta_{t_1})\right]. \tag{A.41}$$

□

A.2.2 Lemmas for the Bandit Setting

Verifying Assumption 3

Lemma 5. *If* $\nabla_r \left[(\pi^* - \pi_\tau^*)^\top r\right] = \mathbf{0}$*, then all suboptimal rewards must be equal.*

Proof. Setting gradient of the bias of softmax optimal policy $(\pi^* - \pi_\tau^*)^\top r$ with respect to the reward vector r equal to a zero vector, the derivative of the bias with respect to an

arbitrary suboptimal reward $r(\hat{a})$, where $\hat{a}$ is a suboptimal action, should be 0:

$$\frac{d}{dr(\hat{a})}(\pi^* - \pi^*_\tau)^\top r = 0 \implies \frac{d}{dr(\hat{a})} \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = 0 \tag{A.42}$$

$$\implies \frac{\left(\frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}[r(a^*) - r(\hat{a})] - e^{\frac{r(\hat{a})}{\tau}}\right)\left(\sum_a e^{\frac{r(a)}{\tau}}\right) - \frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}\left(\sum_a e^{\frac{r(a)}{\tau}}[r(a^*) - r(a)]\right)}{\left(\sum_{a'} e^{\frac{r(a')}{\tau}}\right)^2} = 0 \tag{A.43}$$

$$\implies \frac{\frac{e^{\frac{r(\hat{a})}{\tau}}}{\tau}\left(\sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}) - \tau]\right)}{\left(\sum_{a'} e^{\frac{r(a')}{\tau}}\right)^2} = 0 \implies \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}) - \tau] = 0 \tag{A.44}$$

Now, for any two suboptimal actions $\hat{a}_i$ and $\hat{a}_j$, we have

$$\implies \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}_i) - \tau] - \sum_a e^{\frac{r(a)}{\tau}}[r(a) - r(\hat{a}_j) - \tau] = 0 - 0 \tag{A.45}$$

$$\implies \sum_a e^{\frac{r(a)}{\tau}}[r(\hat{a}_j)) - r(\hat{a}_i)] = 0 \implies r(\hat{a}_j) = r(\hat{a}_i). \tag{A.46}$$

Therefore, all suboptimal rewards must be equal. □

Lemma 6. *We have* $(\pi^* - \pi^*_\tau)^\top r \leq \tau W\left(\frac{A-1}{e}\right)$*, where* $W \colon \mathbb{R}^+ \mapsto \mathbb{R}^+$ *is the principal branch of the Lambert W function, which is defined by* $W(x)e^{W(x)} = x \quad \forall x \geq 0$.

Proof. We want to find an upper bound on the difference between the expected reward achieved by the optimal policy π^* and the softmax optimal policy $\pi^*_\tau = \text{softmax}(r/\tau)$. Denoting $\Delta(a) = r(a^*) - r(a)$, $\Delta = \min_{a \neq a^*} \Delta(a)$, and a^* is the optimal action, we have

$$(\pi^* - \pi^*_\tau)^\top r = \sum_a \pi^*_\tau(a)\, r(a^*) - \sum_a \pi^*_\tau(a)\, r(a) = \sum_{a \neq a^*} \pi^*_\tau(a)\, \Delta(a) = \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}}. \tag{A.47}$$

To find the upper bound, it is enough to find a reward vector $r \in \mathbb{R}^A$ that maximizes the bias. To do so, we find a unique stationary point and then prove that it is the reward vector with the maximum bias. First, we show that decreasing all rewards by a constant value c does not change the bias:

$$(\pi^* - \pi^*_\tau)^\top (r - c\mathbf{1}) = \frac{\sum_{a \neq a^*} e^{\frac{r(a)-c}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')-c}{\tau}}} = \frac{e^{-\frac{c}{\tau}} \sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{e^{-\frac{c}{\tau}} \sum_{a'} e^{\frac{r(a')}{\tau}}} \tag{A.48}$$

$$= \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = (\pi^* - \pi^*_\tau)^\top r \tag{A.49}$$

Therefore, without loss of generality, we assume that the smallest reward value equals 0. Furthermore, according to Lemma 5, stationary reward vectors must have equal values for

all non-optimal actions. Therefore, we assume that the reward vector has a value of $r_{a^*} = \Delta$ for the optimal action and 0 values for all other actions. In this case,

$$(\pi^* - \pi^*_\tau)^\top r = \frac{\sum_{a \neq a^*} e^{\frac{r(a)}{\tau}} \Delta(a)}{\sum_{a'} e^{\frac{r(a')}{\tau}}} = \frac{(A-1)\Delta}{e^{\frac{\Delta}{\tau}} + A - 1}. \tag{A.50}$$

Now, we find the reward gap Δ that makes the first derivative of the bias with respect to Δ equal to 0:

$$\frac{d}{d\Delta} \frac{(A-1)\Delta}{e^{\frac{\Delta}{\tau}} + A - 1} = 0 \implies \frac{(A-1)\left(e^{\frac{\Delta}{\tau}} + A - 1\right) - \frac{(A-1)\Delta e^{\frac{\Delta}{\tau}}}{\tau}}{\left(e^{\frac{\Delta}{\tau}} + A - 1\right)^2} = 0 \tag{A.51}$$

$$\implies (A-1)\left(e^{\frac{\Delta}{\tau}} + A - 1\right) - \frac{(A-1)\Delta e^{\frac{\Delta}{\tau}}}{\tau} = 0 \implies \tau\left(e^{\frac{\Delta}{\tau}} + A - 1\right) = \Delta e^{\frac{\Delta}{\tau}} \tag{A.52}$$

$$\implies \tau(A-1) = (\Delta - \tau)e^{\frac{\Delta}{\tau}} \implies \frac{\Delta - \tau}{\tau} e^{\frac{\Delta}{\tau}} = A - 1 \implies \frac{\Delta - \tau}{\tau} e^{\frac{\Delta - \tau}{\tau}} = \frac{A-1}{e} \tag{A.53}$$

$$\implies W\left(\frac{A-1}{e}\right) = \frac{\Delta - \tau}{\tau} \implies \Delta = \tau\left(W\left(\frac{A-1}{e}\right) + 1\right), \tag{A.54}$$

where $W : \mathbb{R} \mapsto \mathbb{R}$ is the principal branch of the Lambert W function. Since this value is the only stationary point of the bias with respect to the rewards vector, $\Delta = \tau\left(W\left(\frac{A-1}{e}\right) + 1\right)$ is either the global maximum or the global minimum point. Since π^* is the optimal policy, the bias $(\pi^* - \pi^*_\tau)^\top r$ is always non-negative. For $\Delta = 0$, the bias is equal to 0, so the unique stationary point must yield the global maximum. Substituting it in Equation (A.50), we get

$$(\pi^* - \pi^*_\tau)^\top r \leq \frac{(A-1)\tau\left(W\left(\frac{A-1}{e}\right) + 1\right)}{e^{W\left(\frac{A-1}{e}\right)+1} + A - 1}. \tag{A.55}$$

Now, since $e^{W(x)} = \frac{x}{W(x)}$,

$$= \frac{(A-1)\tau\left(W\left(\frac{A-1}{e}\right) + 1\right)}{\frac{A-1}{W\left(\frac{A-1}{e}\right)} + A - 1} \tag{A.56}$$

$$= \tau W\left(\frac{A-1}{e}\right). \tag{A.57}$$

□

Verifying Assumption 4

Lemma 7. *For a fixed θ and τ, we have*

$$(\pi_\tau^* - \pi_\theta)^\top r \le {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau \log A. \tag{A.58}$$

Proof.

$$(\pi_\tau^* - \pi_\theta)^\top r = {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau(\pi_\tau^* \log \pi_\tau^* - \pi_\theta \log \pi_\theta) \tag{A.59}$$

For all θ, $\log \frac{1}{A} \le {\pi_\theta}^\top \log \pi_\theta \le 0$

$$\le {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau \left(0 - \log \frac{1}{A}\right) \tag{A.60}$$

$$= {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_\theta}^\top (r - \tau \log \pi_\theta) + \tau \log A. \tag{A.61}$$

□

Verifying Assumption 5

Lemma 8. *Set $f^\tau(\theta) = {\pi_\theta}^\top (r - \tau \log \pi_\theta)$. For a fixed θ, if $\tau_2 < \tau_1$, then*

$$f^{*\tau_2} - f^{\tau_2}(\theta) \le f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \tau_1 \log A. \tag{A.62}$$

Proof. Assuming $\tau_2 < \tau_1$, we have

$$[f^{*\tau_2} - f^{\tau_2}(\theta)] - [f^{*\tau_1} - f^{\tau_1}(\theta)] = [f^{*\tau_2} - f^{*\tau_1}] - [f^{\tau_2}(\theta) - f^{\tau_1}(\theta)] \tag{A.63}$$

$$= \left[{\pi_{\tau_2}^*}^\top (r - \tau_2 \log \pi_{\tau_2}^*) - {\pi_{\tau_1}^*}^\top (r - \tau_1 \log \pi_{\tau_1}^*)\right] - [{\pi_\theta}^\top (r - \tau_2 \log \pi_\theta) - {\pi_\theta}^\top (r - \tau_1 \log \pi_\theta)] \tag{A.64}$$

$$= (\pi_{\tau_2}^* - \pi_{\tau_1}^*)^\top r - \left[\tau_2 {\pi_{\tau_2}^*}^\top \log \pi_{\tau_2}^* - \tau_1 {\pi_{\tau_1}^*}^\top \log \pi_{\tau_1}^*\right] + (\tau_2 - \tau_1) {\pi_\theta}^\top \log \pi_\theta \tag{A.65}$$

For all θ, $\log \frac{1}{A} \le {\pi_\theta}^\top \log \pi_\theta \le 0$

$$\le (\pi_{\tau_2}^* - \pi_{\tau_1}^*)^\top r - \left[\tau_2 \log \frac{1}{A} - \tau_1 0\right] + (\tau_2 - \tau_1) \log \frac{1}{A} \le (\pi^* - \pi_{\tau_1}^*)^\top r + \tau_1 \log A. \tag{A.66}$$

By Lemma 6

$$\implies f^{*\tau_2} - f^{\tau_2}(\theta) \le f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \tau_1 \log A. \tag{A.67}$$

□

A.2.3 Lemmas for Tabular MDP Setting

Verifying Assumption 3

Lemma 9 (Equation (12) in (Cen et al., 2022))**.** $V^*(\rho) - V^{\pi^*_\tau}(\rho) \leq \tau \frac{\log A}{1-\gamma}$.

Verifying Assumption 4

Lemma 10. *For any π and ρ, we have*

$$\mathbb{H}(\pi) \leq \frac{\log A}{1-\gamma}, \tag{A.68}$$

where

$$\mathbb{H}(\pi) := \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log\pi(a_t|s_t)\right]. \tag{A.69}$$

Proof.

$$\begin{aligned}
\mathbb{H}(\pi) =& \mathop{\mathbb{E}}_{\substack{s_0\sim\rho, a_t\sim\pi(\cdot|s_t),\\ s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log\pi(a_t|s_t)\right] && \text{(A.70)}\\
=& \frac{1}{1-\gamma}\sum_{s,a} d^\pi_\rho(s)\,\pi(a|s)\,[-\log\pi(a|s)] && \text{(A.71)}\\
=& \frac{1}{1-\gamma}\sum_{s} d^\pi_\rho(s)\left[-\sum_a \pi(a|s)\,\log\pi(a|s)\right] && \text{(A.72)}
\end{aligned}$$

Since for all π, $\log\frac{1}{A} \leq \sum_a \pi(a|s)\,\log\pi(a|s) \leq 0$

$$\begin{aligned}
\leq& \frac{1}{1-\gamma}\sum_{s} d^\pi_\rho(s)\left[-\log\frac{1}{A}\right] && \text{(A.73)}\\
=& \frac{1}{1-\gamma}\sum_{s} d^\pi_\rho(s)\,\log A && \text{(A.74)}\\
=& \frac{\log A}{1-\gamma} && \text{(A.75)}
\end{aligned}$$

□

Lemma 11. *For a fixed θ and τ, we have*

$$V^{\pi^*_\tau}(\rho) - V^{\pi_\theta}(\rho) \leq \tilde{V}^*_\tau(\rho) - \tilde{V}^{\pi_\theta}_\tau(\rho) + \frac{\tau\log A}{1-\gamma}. \tag{A.76}$$

Proof.

$$V^{\pi^*_\tau}(\rho) - V^{\pi_\theta}(\rho) = (V^{\pi^*_\tau}(\rho) + \tau\mathbb{H}(\rho, \pi^*_\tau)) - (V^{\pi_\theta}(\rho) + \tau\mathbb{H}(\pi_\theta)) + \tau(\mathbb{H}(\pi_\theta) - \mathbb{H}(\pi^*_\tau)) \quad \text{(A.77)}$$

$$= \tilde{V}^*_\tau(\rho) - \tilde{V}^{\pi_\theta}_\tau(\rho) + \tau(\mathbb{H}(\pi_\theta) - \mathbb{H}(\pi^*_\tau)) \quad \text{(A.78)}$$

Since for all π, $\mathbb{H}(\pi) \geq 0$

$$\leq \tilde{V}^*_\tau(\rho) - \tilde{V}^{\pi_\theta}_\tau(\rho) + \tau\mathbb{H}(\pi_\theta) \quad \text{(A.79)}$$

By Lemma 10

$$\leq \tilde{V}^*_\tau(\rho) - \tilde{V}^{\pi_\theta}_\tau(\rho) + \frac{\tau \log A}{1-\gamma} \quad \text{(A.80)}$$

□

Verifying Assumption 5

Lemma 12. *For a fixed θ, if $\tau_2 < \tau_1$, then*

$$\tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_2}(\rho) \leq \tilde{V}^*_{\tau_1}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho) + \frac{2\,\tau_1 \log A}{1-\gamma}. \quad \text{(A.81)}$$

Proof. Assuming $\tau_2 < \tau_1$, we have

$$\tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_2}(\rho) - \tilde{V}^*_{\tau_1}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho) = [\tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^*_{\tau_1}(\rho)] - [\tilde{V}^{\pi_\theta}_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho)] \quad \text{(A.82)}$$

$$= \left[\left(V^{\pi^*_{\tau_2}}(\rho) + \tau_2\,\mathbb{H}(\pi^*_{\tau_2})\right) - \left(V^{\pi^*_{\tau_1}}(\rho) + \tau_1\,\mathbb{H}(\pi^*_{\tau_1})\right)\right]$$

$$- [(V^{\pi_\theta}(\rho) + \tau_2\,\mathbb{H}(\pi_\theta)) - (V^{\pi_\theta}(\rho) + \tau_1\,\mathbb{H}(\pi_\theta))] \quad \text{(A.83)}$$

$$= \left[V^{\pi^*_{\tau_2}}(\rho) - V^{\pi^*_{\tau_1}}(\rho)\right] + [\tau_2\,\mathbb{H}(\pi^*_{\tau_2}) - \tau_1\,\mathbb{H}(\pi^*_{\tau_1})]$$

$$+ (\tau_1 - \tau_2)\,\mathbb{H}(\rho, \pi_\theta). \quad \text{(A.84)}$$

By Lemma 10, $0 \leq \mathbb{H}(\pi) \leq \frac{\log A}{1-\gamma}$

$$\leq \left[V^{\pi^*_{\tau_2}}(\rho) - V^{\pi^*_{\tau_1}}(\rho)\right] + \left[\tau_2 \frac{\log A}{1-\gamma} - \tau_1\, 0\right] + (\tau_1 - \tau_2)\frac{\log A}{1-\gamma} \quad \text{(A.85)}$$

$$\leq V^*(\rho) - V^{\pi^*_{\tau_1}}(\rho) + \tau_1 \frac{\log A}{1-\gamma}. \quad \text{(A.86)}$$

By Lemma 9,

$$\implies \tilde{V}^*_{\tau_2}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_2}(\rho) \leq \tilde{V}^*_{\tau_1}(\rho) - \tilde{V}^{\pi_\theta}_{\tau_1}(\rho) + \frac{2\tau_1 \log A}{1-\gamma}. \quad \text{(A.87)}$$

□

A.3 Proofs of Section 3.4

A.3.1 Proof of Theorem 2

Theorem 2. Assuming f^{τ} and f satisfy Assumptions 2 to 6, for a given $\epsilon \in (0,1)$, using Algorithm 2 with (a) unbiased stochastic gradients whose variance is bounded by σ^2 and (b) exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\, \alpha_i^{t-\text{last}_{i-1}+1}$, where $\eta_{i,\text{last}_{i-1}} = \frac{1}{L^{\tau_i}}$, $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, and setting constants $c_0 = 0.69, c_1 = 5583$, achieves ϵ-sub-optimality to the globally optimal policy after $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

Proof. Observe that in Algorithm 2, we use τ_i at stage $i \geq 1$, which starts at iteration $\text{last}_{i-1}+1$, ends at iteration last_i, and runs for $T_i = \max(5583, 2\,T_i^{'}\,\log T_i^{'}, 4\,T_i^{''}\,\log^2 T_i^{''})$ iterations, where

$$T_i^{'} = \frac{2\,\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i}, \quad T_i^{''} = \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}, \tag{A.88}$$

where $X_1 = \exp\left(\frac{\mu_i\,\beta}{L^{\tau_i}\,\log(T/\beta)}\right)$, $X_2 = \frac{0.69}{L^{\tau_i}}$, and $X_3 = \frac{5\,L^{\tau_i}\,X_1}{e^2}$. Now, we will prove by induction that $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ for all $i \geq 0$:
Base Case: For $i = 0$, we have

$$f^{*\tau_0} - f^{\tau_0}(\theta_0) \leq \max(\tau_0, f^{*\tau_0} - f^{\tau_0}(\theta_0)) = \tau_0 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right). \tag{A.89}$$

Induction Step: Suppose $\mathbb{E}[f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}})] \leq \tau_{i-1} \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds. At stage i, by Lemma 13, using exponentially decreasing step-size $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\, \alpha_i^{t-\text{last}_{i-1}+1}$, where $\eta_{i,\text{last}_{i-1}} = \frac{1}{L^{\tau_i}}$, $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$ with $\beta = 1$, for $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)$ to hold, it suffices that $T_i \geq \max(5583, 2\,Y_i\,\log Y_i, 4\,Y_i^{'}\,\log^2 Y_i^{'})$, where

$$Y_i = \frac{2\,\log\left(\frac{2\,X_1\,\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_{i-1}})]}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i}, \quad Y_i^{'} = \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2 \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}. \tag{A.90}$$

Under Assumption 5,

$$Y_i \leq \frac{2\,\log\left(\frac{2\,X_1\left(\mathbb{E}[f^{*\tau_{i-1}} - f^{\tau_{i-1}}(\theta_{\text{last}_{i-1}})] + \tau_{i-1}B_4\right)}{\tau_i \max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i} \tag{A.91}$$

Using the inductive hypothesis

$$\leq \frac{2\,\log\left(\frac{2\,X_1\left(\tau_{i-1}\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)+\tau_{i-1}B_4\right)}{\tau_i\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i} \tag{A.92}$$

$$\leq \frac{2\,\log\left(\frac{2\,X_1\,\tau_{i-1}\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)(1+B_4)}{\tau_i\max\left(1,\frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)}\right)}{X_2\,\mu_i} \tag{A.93}$$

$$= \frac{2\,\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i} = T_i^{'}. \tag{A.94}$$

On the other hand, we have

$$Y_i^{'} \leq \frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2} = T_i^{''}. \tag{A.95}$$

Therefore, $T_i = \max(5583, 2\,T_i^{'}\,\log T_i^{'}, 4\,T_i^{''}\,\log^2 T_i^{''}) \geq \max(5583, 2\,Y_i\,\log Y_i, 4\,Y_i^{'}\,\log^2 Y_i^{'})$. This implies $\mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] \leq \tau_i\,\max\left(1, \frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right)$ holds for all $i \geq 0$. As a result, under Assumption 4, we have

$$\mathbb{E}[f(\theta^*_{\tau_i}) - f(\theta_{\text{last}_i})] \leq \mathbb{E}[f^{*\tau_i} - f^{\tau_i}(\theta_{\text{last}_i})] + \tau_i\,B_3 \tag{A.96}$$

$$\leq \tau_i\,\left(\max\left(1, \frac{f^{*\tau_0} - f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_3\right) \tag{A.97}$$

Denote $\epsilon_i := \mathbb{E}[f^* - f(\theta_{\text{last}_i})]$ as the suboptimality at the end of stage i. We have

$$\epsilon_i = \mathbb{E}[f^* - f(\theta_{\text{last}_i})] \tag{A.98}$$

$$= f^* - f(\theta^*_{\tau_i}) + \mathbb{E}[f(\theta^*_{\tau_i}) - f(\theta_{\text{last}_i})] \tag{A.99}$$

Under Assumption 3

$$\leq \tau_i\,C_1 \tag{A.100}$$

where $C_1 = \max\left(1, \frac{f^{*\tau_0}-f^{\tau_0}(\theta_0)}{\tau_0}\right) + B_2 + B_3$. Therefore, ϵ_i has an upper bound that is proportional to τ_i. Now, since $\tau_i = 2^{-i}\,\tau_0$, the sub-optimality ϵ_i has an exponential rate in terms of the number of executed stages:

$$= 2^{-i}\,\tau_0\,C_1 \tag{A.101}$$

Therefore, the required number of stages N_{stages} in terms of the final sub-optimality $\epsilon := \epsilon_{N_{\text{stages}}}$ is

$$2^{N_{\text{stages}}} \geq \frac{\tau_0\,C_1}{\epsilon} \implies N_{\text{stages}} \geq \log_2\left(\frac{\tau_0\,C_1}{\epsilon}\right). \tag{A.102}$$

On the other hand, we have the sufficient number of iterations at stage i:

$$T_i \geq \max\left(5583, \frac{4\,\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i}\,\log\left(\frac{\log\left(\frac{2\,X_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{X_2\,\mu_i}\right), \frac{8\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}\,\log^2\left(\frac{2\,X_3\,\sigma^2}{\tau_i\,\mu_i^2}\right)\right) \tag{A.103}$$

Since $\tau_i \leq 1$, under Assumption 6, we have $\mu_i = \tau_i^p\,B_1 \leq B_1$. Furthermore, $\log\left(\frac{T_i}{\beta}\right) \geq 1$, and under Assumption 2, we have $0 < L^{\min} \leq L^{\tau_i} \leq L^{\max}$. Therefore,

$$X_1 \leq A_1 = \exp\left(\frac{B_1\,\beta}{L^{\min}}\right), \tag{A.104}$$

$$X_2 \geq A_2 = \frac{0.69}{L^{\max}}, \tag{A.105}$$

$$X_3 \leq A_3 = \frac{5\,L^{\max}\,A_1}{e^2}. \tag{A.106}$$

Hence, we can safely substitute variables X_1, X_2, X_3 with their corresponding constants A_1, A_2, A_3. Therefore, it is sufficient to set T_i as

$$T_i \geq \max\left(5583, \frac{4\,\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\mu_i}\,\log\left(\frac{\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\mu_i}\right), \frac{8\,A_3\,\sigma^2}{\tau_i\,\mu_i^2}\,\log^2\left(\frac{2\,A_3\,\sigma^2}{\tau_i\,\mu_i^2}\right)\right) \tag{A.107}$$

Under Assumption 6, $\mu_i = \tau_i^p\,B_1$

$$= \max\left(5583, \frac{4\,\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\tau_i^p\,B_1}\,\log\left(\frac{\log\left(\frac{2\,A_1\,\tau_{i-1}(1+B_4)}{\tau_i}\right)}{A_2\,\tau_i^p\,B_1}\right), \frac{8\,A_3\,\sigma^2}{\tau_i^{2p+1}\,B_1^2}\,\log^2\left(\frac{2\,A_3\,\sigma^2}{\tau_i^{2p+1}\,B_1^2}\right)\right) \tag{A.108}$$

Since $\tau_i = 2^{-i}\,\tau_0$

$$= \max\left(5583, \frac{4\,\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\,\log\left(\frac{\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\right), \frac{8\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\,\log^2\left(\frac{2\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\right)\right) \tag{A.109}$$

Since $i \leq N_{\text{stages}}$, it is sufficient that

$$T_i = \max\left(5583, \frac{4\,\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.110}$$

where $Y_1 = \log\left(\frac{\log(4\,A_1\,(1+B_4))\,(2^{N_{\text{stages}}})^p}{A_2\,\tau_0^p\,B_1}\right)$ and $Y_2 = \log^2\left(\frac{2\,A_3\,\sigma^2\,(2^{N_{\text{stages}}})^{2p+1}}{\tau_0^{2p+1}\,B_1^2}\right)$. Consequently, we can calculate the sufficient total number of iterations T_{Total} in terms of ϵ:

$$T_{\text{Total}} \geq \sum_{i=1}^{N_{\text{stages}}} T_i \tag{A.111}$$

$$= \sum_{i=1}^{N_{\text{stages}}} \max\left(5583, \frac{4\,\log(4\,A_1\,(1+B_4))\,2^{ip}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,2^{i(2p+1)}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.112}$$

$$= \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\sum_{i=1}^{N_{\text{stages}}}(2^p)^i}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,\sum_{i=1}^{N_{\text{stages}}}(2^{2p+1})^i}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.113}$$

Since $\forall x > 1, n \geq 0,\ \sum_{i=0}^{n} x^i = \frac{x^{n+1}-1}{x-1}$

$$= \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\left[\frac{(2^p)^{N_{\text{stages}}+1}-1}{2^p-1}-1\right]}{A_2\,\tau_0^p\,B_1}\,Y_1,\right.$$
$$\left.\frac{8\,A_3\,\sigma^2\,\left[\frac{(2^{2p+1})^{N_{\text{stages}}+1}-1}{2^{2p+1}-1}-1\right]}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.114}$$

Therefore, it is sufficient that

$$T_{\text{Total}} \geq \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\frac{(2^p)^{N_{\text{stages}}+1}}{2^p-1}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,\frac{(2^{2p+1})^{N_{\text{stages}}+1}}{2^{2p+1}-1}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.115}$$

$$= \max\left(5583\,N_{\text{stages}}, \frac{4\,\log(4\,A_1\,(1+B_4))\,\frac{2^p\,(2^p)^{N_{\text{stages}}}}{2^p-1}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{8\,A_3\,\sigma^2\,\frac{2^{2p+1}\,(2^{2p+1})^{N_{\text{stages}}}}{2^{2p+1}-1}}{\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.116}$$

Since $p \geq 1$, we have $\frac{2^p}{2^p-1} \leq 2$ and $\frac{2^{2p+1}}{2^{2p+1}-1} \leq \frac{8}{7}$. Hence, it is sufficient to use

$$T_{\text{Total}} = \max\left(5583\,N_{\text{stages}}, \frac{8\,\log(4\,A_1\,(1+B_4))\,(2^p)^{N_{\text{stages}}}}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{64\,A_3\,\sigma^2\,(2^{2p+1})^{N_{\text{stages}}}}{7\,\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.117}$$

$$= \max\left(5583\,N_{\text{stages}}, \frac{8\,\log(4\,A_1\,(1+B_4))\,(2^{N_{\text{stages}}})^p}{A_2\,\tau_0^p\,B_1}\,Y_1, \frac{64\,A_3\,\sigma^2\,(2^{N_{\text{stages}}})^{2p+1}}{7\,\tau_0^{2p+1}\,B_1^2}\,Y_2\right) \tag{A.118}$$

Using Equation (A.102)

$$\geq \max\left(5583\,\log_2\left(\frac{\tau_0\,C_1}{\epsilon}\right), \frac{8\,\log(4\,A_1\,(1+B_4))\,C_1^p\,\log\left(\frac{\log(4\,A_1\,(1+B_4))\,C_1^p}{A_2\,B_1\,\epsilon^p}\right)}{A_2\,B_1\,\epsilon^p}, \frac{64\,A_3\,C_1^{2p+1}\,\log^2\left(\frac{2\,A_3\,C_1^{2p+1}\,\sigma^2}{B_1^2\,\epsilon^{2p+1}}\right)\,\sigma^2}{7\,B_1^2\,\epsilon^{2p+1}}\right) \tag{A.119}$$

$$\implies T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right). \tag{A.120}$$

□

Corollary 3. In the bandit setting, assuming for each stage i, $\mu_i = \tau_i^p B_1$ for constants $p \geq 1$, $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 2 with exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\,\alpha_i^{t-\text{last}_{i-1}+1}$ where $\eta_{i,\text{last}_{i-1}} = \frac{2}{5+10\,\tau_i\,(1+\log A)}$ and $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, achieves ϵ-suboptimality after $T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

Proof. Set $f(\theta) = {\pi_\theta}^\top r$ and $f^\tau(\theta) = {\pi_\theta}^\top(r - \tau\log\pi_\theta)$. We can extend Theorem 2 to the bandit setting since:

- by Lemma 18, f^τ is L^τ-smooth and $\tau \in [0,1]$

$$\frac{5}{2} = L^{\min} \leq L^\tau = \frac{5}{2} + \tau\,5\,(1+\log A) \leq \frac{5}{2} + 5\,(1+\log A) = L^{\max} \tag{A.121}$$

- by Lemma 6, we have $f^* - f(\theta_\tau^*) \leq \tau W\left(\frac{A-1}{e}\right)$
- by Lemma 7, we have for all θ, $f(\theta_\tau^*) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau\,\log A$
- by Lemma 8, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 W\left(\frac{A-1}{e}\right) + \log A$
- by Lemma 30, the gradient estimator is unbiased and have bounded variance where $\sigma^2 = 8\,(1 + (\tau\,\log A)^2)$.

□

Corollary 4. In the tabular MDP setting, assuming for each stage i , $\mu_i = \tau_i^p\,B_1$ for constants $p \geq 1$, $B_1 > 0$, for a given $\epsilon \in (0,1)$, using Algorithm 2 with exponentially decreasing step-sizes $\eta_{i,t} = \eta_{i,\text{last}_{i-1}}\,\alpha_i^{t-\text{last}_{i-1}+1}$, where $\eta_{i,\text{last}_{i-1}} = \frac{(1-\gamma)^3}{8+\tau_i(4+8\log A)}$ and $\alpha_i = \left(\frac{\beta}{T_i}\right)^{\frac{1}{T_i}}$, $\beta = 1$, achieves ϵ-sub-optimality after $T_{\text{Total}} \in \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^p} + \frac{\sigma^2}{\epsilon^{2p+1}}\right)$ iterations.

Proof. Set $f(\theta) = V^{\pi_\theta}(\rho)$ and $f^\tau(\theta) = \tilde{V}_\tau^{\pi_\theta}(\rho)$. We can extend Theorem 2 to the MDP setting since:

- by Lemma 20, f^τ is L^τ-smooth and since $\tau \in [0,1]$

$$L^{\min} = \frac{8}{(1-\gamma)^3} \leq L^\tau = \frac{8 + \tau(4 + 8\log A)}{(1-\gamma)^3} \leq \frac{12 + 8\log A}{(1-\gamma)^3} = L^{\max} \tag{A.122}$$

- by Lemma 9, we have $f^* - f(\theta_\tau^*) \leq \tau \frac{\log A}{1-\gamma}$
- by Lemma 11, we have for all θ, $f(\theta_\tau^*) - f(\theta) \leq f^{*\tau} - f^\tau(\theta) + \tau \frac{\log A}{1-\gamma}$
- by Lemma 12, we have for all θ, $f^{*\tau_2} - f^{\tau_2}(\theta) \leq f^{*\tau_1} - f^{\tau_1}(\theta) + \tau_1 \frac{2\log A}{1-\gamma}$
- by Lemma 29, the gradient estimators are unbiased and have bounded variance where $\sigma^2 = \frac{8}{(1-\gamma)^2}\left(\frac{1+(\tau \log A)^2}{(1-\gamma^{1/2})^2}\right)$.

□

Additional Lemmas

Lemma 13. *Assuming f^τ satisfies Assumptions 2 and 6 and the gradient estimators $\widetilde{\nabla} f^\tau(\theta_t)$ are unbiased and have bounded variance σ^2, for a given $\epsilon \in (0,1)$, using Update 4 from iteration t_1+1 to t_2 with exponentially decreasing step-sizes $\eta_t = \eta_0\,\alpha^{t-t_1+1}$, where $\eta_t = \frac{1}{L^\tau}$ and $\alpha = (\frac{\beta}{T})^{\frac{1}{T}}$, $\beta \geq 1$, and $T = t_2 - t_1 > 0$, is achieved in ϵ-sub-optimality is achieved in $\max(\beta+1, 5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$ iterations, where $Y_1 = \frac{2\,\log\left(\frac{2\,X_1\,\mathbb{E}[f^{*\tau}-f^\tau(\theta_{t_1})]}{\epsilon}\right)}{X_2\,\mu}$, $Y_2 = \frac{2\,X_3\,\sigma^2}{\mu^2\,\epsilon}$,$X_1 = \exp\left(\frac{\mu\,\beta}{L^\tau\,\log(T/\beta)}\right)$, $X_2 = \frac{0.69}{L^\tau}$, and $X_3 = \frac{5\,L^\tau\,X_1}{e^2}$.*

Proof. From (Li et al., 2021b, Theorem 1), using Update 4 with exponentially decreasing step-sizes results from iterations t_1+1 to t_2 results in the following convergence

$$\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq X_1\,\exp\left(-\frac{X_2\,\mu}{2}\,\frac{T}{\log\frac{T}{\beta}}\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] + \frac{X_3\,\sigma^2}{\mu^2\,\frac{T}{\log^2\frac{T}{\beta}}}, \tag{A.123}$$

where

$$X_1 = \exp\left(\frac{\mu\,\beta}{L^\tau\,\log\frac{T}{\beta}}\right), \quad X_2 = \frac{0.69}{L^\tau}, \quad X_3 = \frac{5\,L^\tau\,X_1}{e^2} \tag{A.124}$$

and $\mu := \inf_{t\geq 1} C_\tau(\theta)$ with $T = t_2 - t_1$. We show that if the inequalities $\frac{T}{\log\frac{T}{\beta}} \geq Y_1$ and $\frac{T}{\log^2\frac{T}{\beta}} \geq Y_2$ are satisfied, where

$$Y_1 = \frac{2\,\log\left(\frac{2\,X_1\,\mathbb{E}[f^{*\tau}-f^\tau(\theta_{t_1})]}{\epsilon}\right)}{X_2\,\mu}, \quad Y_2 = \frac{2\,X_3\,\sigma^2}{\mu^2\,\epsilon}, \tag{A.125}$$

then $\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] \leq \epsilon$ holds since

$$\begin{aligned}
&\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_2})] && \text{(A.126)}\\
&\quad \leq X_1\,\exp\left(-\frac{X_2\,\mu}{2}\,\frac{2}{X_2\,\mu}\log\left(\frac{2\,X_1\,[f^{*\tau} - f^\tau(\theta_{t_1})]}{\epsilon}\right)\right)\mathbb{E}[f^{*\tau} - f^\tau(\theta_{t_1})] + \frac{X_3\,\sigma^2}{\mu^2\,\frac{2\,X_3\,\sigma^2}{\mu^2\,\epsilon}} && \text{(A.127)}\\
&\quad = \frac{\epsilon}{2} + \frac{\epsilon}{2} && \text{(A.128)}\\
&\quad = \epsilon. && \text{(A.129)}
\end{aligned}$$

By Lemma 14 and since $1 \leq \beta < T$, for $\frac{T}{\log(T/\beta)} \geq \frac{T}{\log T} \geq Y_1$ to hold, it suffices that $T \geq \max(2, 2\,Y_1\,\log Y_1)$. Furthermore, according to Lemma 15 and since $1 \leq \beta < T$, for $\frac{T}{\log^2(T/\beta)} \geq \frac{T}{\log^2 T} \geq Y_2$ to hold, it suffices that $T \geq \max(5583, 4\,Y_2\,\log^2 Y_2)$. Therefore, the required number of iterations to achieve ϵ-sub-optimality is $\max(5583, 2\,Y_1\,\log Y_1, 4\,Y_2\,\log^2 Y_2)$. □

Lemma 14. *For all $C > 0$, if $T \geq \max(2, 2\,C\,\log C)$, then $\frac{T}{\log T} \geq C$.*

Proof. If $C < 2$, knowing that $T \geq 2$, we have

$$\frac{T}{\log T} > 2 > C \tag{A.130}$$

Otherwise, if $C \geq 2$,

$$2\,C\,\log C = C(\log C + \log C) \tag{A.131}$$

Since $\forall C > 0$, $C \geq 2\,\log C$,

$$\geq C(\log C + \log(2\log C)) \tag{A.132}$$

$$= C\log(2\,C\,\log C) \tag{A.133}$$

$$\implies \frac{2\,C\,\log C}{\log(2\,C\,\log C)} \geq C. \tag{A.134}$$

Therefore, knowing that $T \geq 2\,C\,\log C$, since $2\,C\,\log C \geq 4\log 2 > 2.72$, we have

$$\frac{T}{\log T} \geq \frac{2\,C\,\log C}{\log(2\,C\,\log C)} \geq C. \tag{A.135}$$

□

Lemma 15. *For all $C > 0$, if $T \geq \max(5583, 4\,C\,\log^2 C)$, then $\frac{T}{\log^2 T} \geq C$.*

Proof. If $C < 75$, knowing that $T \geq 5583$, we have

$$\frac{T}{\log^2 T} > 75 > C. \tag{A.136}$$

Otherwise, if $C \geq 75$,

$$4\,C\,\log^2 C = C(\log C + \log C)^2 \tag{A.137}$$

Since $C \geq 4\,\log^2 C \quad \forall C \geq 75$,

$$\geq C(\log C + \log(4\log^2 C))^2 = C\log^2(4\,C\,\log^2 C) \tag{A.138}$$

$$\implies \frac{4\,C\,\log^2 C}{\log^2(4\,C\,\log^2 C)} \geq C. \tag{A.139}$$

Therefore, knowing that $T \geq 4\,C\,\log^2 C$, since $4\,C\,\log^2 C \geq 300\log^2 75 > 8$, we have

$$\frac{T}{\log^2 T} \geq \frac{4\,C\,\log^2 C}{\log^2(4\,C\,\log^2 C)} \geq C. \tag{A.140}$$

□

A.4 Additional Lemmas

For completeness, we append external lemmas here.

A.4.1 Smoothness

Lemma 16 (Lemma 2 in Mei et al. (2020b)). *$\forall r \in [0,1]^A$ $\theta \mapsto \langle \pi_\theta, r\rangle$ is $\frac{5}{2}$-smooth.*

Lemma 17 (Lemma 14 in (Mei et al., 2020b)). *$\theta \to -\langle \pi_\theta, \log \pi_\theta\rangle$ is $5\,(1+\log K)$-smooth.*

Lemma 18. *$\theta \to \langle \pi_\theta, r - \tau \log \pi_\theta\rangle$ is $\frac{5}{2} + \tau\, 5\,(1+\log K)$-smooth.*

Proof. By Lemma 16 and Lemma 17. □

Lemma 19 (Lemma 7 in Mei et al. (2020b)). *$\theta \to V^{\pi_\theta}(\rho)$ is $\frac{8}{(1-\gamma)^3}$-smooth.*

Lemma 20 (Lemmas 7 and 14 in (Mei et al., 2020b)). *$\theta \to V^{\pi_\theta}(\rho) + \tau\, \mathbb{H}(\pi_\theta)$ is $\frac{8+\tau\,(4+8\,\log A)}{(1-\gamma)^3}$-smooth.*

Lemma 21 (Lemma 2 in (Mei et al., 2021b)). *In the bandits setting, for any $r \in [0,1]^A$, $\theta \to \langle \pi_\theta, r\rangle$ is 3-non-uniform smooth.*

Lemma 22 (Lemma 6 in (Mei et al., 2021b)). *In the tabular MDP setting, assuming $\min_{s\in\mathcal{S}} \rho(s) > 0$, $\theta \to V^{\pi_\theta}(\rho)$ is C-non-uniform smooth with where $C := \left[3 + \frac{2\,C_\infty - (1-\gamma)}{(1-\gamma)\,\gamma}\right]\sqrt{S}$ and $C_\infty := \max_\pi \left\|\frac{d_\rho^\pi}{\rho}\right\|_\infty \leq \frac{1}{\min_s \rho(s)} < \infty$.*

Non-uniform Łojasiewicz condition

Lemma 23 (Lemma 3 in Mei et al. (2020b)). *Let $\pi^* := \max_{\pi\in\Pi} \langle \pi, r\rangle$. Then*

$$\left\|\frac{d\langle \pi_\theta, r\rangle}{d\theta}\right\|_2 \geq C(\theta)\,\langle \pi^* - \pi_\theta, r\rangle \tag{A.141}$$

where $C(\theta) := \pi_\theta(a^*)$.

Lemma 24 (Lemma 8 in Mei et al. (2020b)). *Let $V^*(\rho) := \max_{\pi\in\Pi} V^\pi(\rho)$. Then*

$$\left\|\frac{\partial V^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2 \geq C(\theta)\,(V^*(\rho) - V^{\pi_\theta}(\rho)) \tag{A.142}$$

where $C(\theta) := \frac{\min_s \pi_\theta(a^*(s)\,|\,s)}{\sqrt{S}\left\|\frac{d_\rho^{\pi^*}}{d_\rho^{\pi_\theta}}\right\|_\infty}$.

Lemma 25 (Proposition 5 in (Mei et al., 2020b)). *In the bandits setting, the non-uniform Łojasiewicz condition is*

$$\left\|\frac{d\langle \pi_\theta, (r-\tau\log\pi_\theta)\rangle}{d\theta}\right\|_2 \geq C_\tau(\theta)\left(\mathbb{E}_{a\sim\pi_\tau^*}[r(a)-\tau\log\pi_\tau^*]-\mathbb{E}_{a\sim\pi_\theta}[r(a)-\tau\log\pi_\theta]\right)^{\frac{1}{2}} \tag{A.143}$$

with

$$C_\tau(\theta) := \sqrt{2\tau}\,\min_a \pi_\theta(a). \tag{A.144}$$

Lemma 26 (Lemma 15 in (Mei et al., 2020b)). *In the tabular MDP setting, supposing $\rho(s) > 0$ for all states $s \in \mathcal{S}$, the non-uniform Łojasiewicz condition is*

$$\left\|\frac{\partial \tilde{V}_\tau^{\pi_\theta}(\rho)}{\partial\theta}\right\|_2 \geq C_\tau(\theta)\,\left[\tilde{V}_\tau^*(\rho)-\tilde{V}_\tau^{\pi_\theta}(\rho)\right]^{\frac{1}{2}} \tag{A.145}$$

with

$$C_\tau(\theta) := \frac{\sqrt{2\tau}}{\sqrt{S}}\,\min_s\sqrt{\rho(s)}\,\min_{s,a}\pi_\theta(a|s)\left\|\frac{d_\rho^{\pi_\tau^*}}{d_\rho^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}}. \tag{A.146}$$

A.4.2 Stochastic Policy Gradients

Lemma 27 (Lemma 5 from (Mei et al., 2021a)). *Let $\hat{r}$ be the IS estimator using on-policy sampling $a \sim \pi_\theta(\cdot)$. Then stochastic softmax PG estimator is:*
***Unbiased**: $\mathbb{E}_{a\sim\pi_\theta}\left[\widetilde{\nabla} f(\theta)\right] = \nabla f(\theta)$*
***Bounded Variance**: $\mathbb{E}_{a\sim\pi_\theta}\left\|\widetilde{\nabla} f(\theta)\right\|_2^2 \leq 2$*
$\Rightarrow \sigma^2 := \mathbb{E}_{a\sim\pi_\theta}\left[\widetilde{\nabla} f(\theta) - \nabla f(\theta)\right] = \mathbb{E}_{a\sim\pi_\theta}\left\|\widetilde{\nabla} f(\theta)\right\|_2^2 - \mathbb{E}_{a\sim\pi_\theta}\|\nabla f(\theta)\|_2^2 \leq 2.$

Lemma 28 (Lemma 11 from (Mei et al., 2021a)). *Let $\hat{Q}^{\pi_\theta}$ be the IS estimator using on-policy sampling $a(s) \sim \pi_\theta(\cdot|s)$. Then stochastic softmax PG estimator is:*
***Unbiased**: $\mathbb{E}\left[\widetilde{\nabla} f^\tau(\theta)\right] = \nabla f^\tau(\theta)$.*
***Bounded Variance**: $\mathbb{E}\left\|\widetilde{\nabla} f(\theta)\right\|_2^2 \leq \frac{2S}{(1-\gamma)^4} \Rightarrow \sigma^2 := \mathbb{E}\left[\widetilde{\nabla} f(\theta) - \nabla f(\theta)\right] \leq \frac{2S}{(1-\gamma)^4}$.*

Lemma 29 (Lemma 3 and Lemma 4 from (Ding et al., 2021)). *Let $\hat{Q}_\tau^{\pi_\theta}$ be the entropy regularized IS estimator using on-policy sampling $a(s) \sim \pi_\theta(\cdot|s)$. Then stochastic softmax PG estimator using entropy regularization is:*
***Unbiased**: $\mathbb{E}\left[\widetilde{\nabla} f^\tau(\theta)\right] = \nabla f^\tau(\theta)$.*
***Bounded Variance**: $\mathbb{E}\left\|\widetilde{\nabla} f^\tau(\theta) - \mathbb{E}[\widetilde{\nabla} f^\tau(\theta)]\right\|_2^2 \leq \sigma^2$, where $\sigma^2 = \frac{8}{(1-\gamma)^2}\left(\frac{1+(\tau\,\log A)^2}{(1-\gamma^{1/2})^2}\right)$.*

Lemma 30 (Instantiation of Lemma 29 in the bandits setting)**.** *Let $\hat{r}$ be the entropy regularized IS estimator using on-policy sampling $a \sim \pi_\theta(\cdot)$. Then stochastic softmax PG estimator using entropy regularization is:*
Unbiased*:* $\mathbb{E}\left[\widetilde{\nabla} f^\tau(\theta)\right] = \nabla f^\tau(\theta)$.
Bounded Variance*:* $\mathbb{E}\left\|\widetilde{\nabla} f^\tau(\theta) - \mathbb{E}[\widetilde{\nabla} f^\tau(\theta)]\right\|_2^2 \leq \sigma^2$*, where* $\sigma^2 = 8\left(1 + (\tau \log A)^2\right)$.

Appendix B

Proofs of Chapter 4

B.1 Definitions

- **[Smoothness]**. A function f is L-smooth if for all θ and θ'

$$|f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle| \leq \frac{L}{2} \|\theta - \theta'\|_2^2.$$

- **[Non-uniform smoothness]** A function f is L-non-uniform smooth if for all θ and θ'

$$|f(\theta) - f(\theta') - \langle \nabla f(\theta'), \theta - \theta' \rangle| \leq \frac{L\|\nabla f(\theta')\|}{2} \|\theta - \theta'\|_2^2.$$

- **[Polyak-Łojasiewciz condition]**. A function f satisfies the non-uniform Polyak-Łojasiewciz condition of degree $\xi \in [0, 1]$ if for all θ

$$\|\nabla f(\theta)\| \geq C(\theta) |f^* - f(\theta)|^{1-\xi},$$

where $f^* := \sup_\theta f(\theta)$ and $C : \theta \to \mathbb{R} > 0$.

B.2 Proofs of Section 4.3

B.2.1 Proof of Proposition 3

Proposition 3. *Denote $a^* := \arg\max_{a \in [K]} r(a)$. With constant $\eta > 0$ and any initialization $\theta_1 \in \mathbb{R}^d$, Algorithm 3 guarantees $\langle \pi_{\theta_t}, r \rangle \to r(a^*)$ as $t \to \infty$ on Example 1.*

Proof. Let $w = (-1, -1)^\top \in \mathbb{R}^d$. We have

$$r' \coloneqq Xw = (2, 1, -1, -2)^\top,$$

which preserves the ordering of $r \in \mathbb{R}^K$, such that for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$, which means Example 1 satisfies the conditions in Theorem 7. The results then follow by using Theorem 7. □

B.3 Proofs of Section 4.4

B.3.1 Warm up: Global Convergence when $K = 3$

Theorem 4. Given a reward vector $r \in \mathbb{R}^3$ and a feature matrix $X \in \mathbb{R}^{3\times d}$ such that Assumptions 7, 8, and 9 are satisfied, Algorithm 3 with a constant learning rate as in Eq. 4.5 is guaranteed to converge to the optimal policy.

Proof. Under Assumptions 7 and 8, according to Lemma 1, for all finite $t \geq 1$.

$$\langle \pi_{\theta_{t+1}}, r\rangle > \langle \pi_{\theta_t}, r\rangle, \tag{B.1}$$

and $\pi_{\theta_t}(a) \to 1$ as $t \to \infty$ for some action $a \in \{1, 2, 3\}$. We will prove $\pi_{\theta_t}(1) \to 1$ as $t \to \infty$ by showing that $\pi_{\theta_t}(2) \not\to 1$ and $\pi_{\theta_t}(3) \not\to 1$ as $t \to \infty$.

For any bounded initialization θ_1, we have $\langle \pi_{\theta_1}, r\rangle > r(3)$. From Equation (B.1), we know that for all finite $t \geq 1$,

$$\langle \pi_{\theta_t}, r\rangle > \langle \pi_{\theta_1}, r\rangle > r(3).$$

Therefore, $\pi_{\theta_t}(3) \not\to 1$ as $t \to \infty$.

Suppose that $\pi_{\theta_t}(2) \to 1$ as $t \to \infty$. Given this assumption and Equation (B.1), we know that for all finite $t \geq 1$, $\langle \pi_{\theta_t}, r\rangle < r(2)$. In this case, we will show that,

$$\lim_{t\to\infty} \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} = \infty,$$

and prove that this implies that for all large enough t, $\langle \pi_{\theta_t}, r\rangle > r(2)$. Hence, this results in a contradiction proving that $\pi_{\theta_t}(2) \not\to 1$. To start, we consider the following ratio,

$$\begin{aligned}
\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} &= \exp([X\,\theta_{t+1}](1) - [X\theta_{t+1}](3)) \\
&= \exp\left([X\,\theta_t](1) - [X\theta_t](3) + \eta\left(\sum_{i=1}^{3} \langle x_i, x_1 - x_3\rangle\, \pi_{\theta_t}(i)\,(r(i) - \langle \pi_{\theta_t}, r\rangle)\right)\right) \\
&\qquad\qquad \text{(By the update in Algorithm 3)} \\
&= \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \exp\left(\eta \underbrace{\left(\sum_{i=1}^{3} \langle x_i, x_1 - x_3\rangle\, \pi_{\theta_t}(i)\,(r(i) - \langle \pi_{\theta_t}, r\rangle)\right)}_{:=P_t}\right),
\end{aligned} \tag{B.2}$$

and the sign of P_t will dictate whether $\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)}$ will increase or decrease. For all finite $t \geq 1$, we have

$$\begin{aligned}
P_t &= \sum_{i=1}^{3} \langle x_i, x_1 - x_3 \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) \\
&= \langle x_1 - x_3, x_1 - x_3 \rangle \, \pi_{\theta_t}(1) \, (r(1) - \langle \pi_{\theta_t}, r \rangle) + \langle x_2 - x_3, x_1 - x_3 \rangle \, \pi_{\theta_t}(2) \, (r(2) - \langle \pi_{\theta_t}, r \rangle) \\
&\qquad\qquad \text{(Since } \textstyle\sum_{i=1}^{3} \langle x_3, x_1 - x_3 \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) = 0\text{)} \\
&> \langle x_1 - x_3, x_1 - x_3 \rangle \, \pi_{\theta_t}(1) \, (r(1) - r(2)) \\
&\quad \text{(Under Assumption 9, } \langle x_2 - x_3, x_1 - x_3 \rangle > 0 \text{ and for all finite } t \geq 1, \, r(2) > \langle \pi_{\theta_t}, r \rangle\text{)} \\
&= \|x_1 - x_3\|_2^2 \, \pi_{\theta_t}(1) \, (r(1) - r(2)) \qquad \text{(B.3)} \\
&> 0.
\end{aligned}$$

By recursing Equation (B.2), we get that,

$$\begin{aligned}
\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} &= \frac{\pi_{\theta_1}(1)}{\pi_{\theta_1}(3)} \exp(\eta \sum_{s=1}^{t-1} P_s) \\
&> \frac{\pi_{\theta_1}(1)}{\pi_{\theta_1}(3)} \exp(\eta \, \|x_1 - x_3\|_2^2 \, (r(1) - r(2)) \sum_{s=1}^{t-1} \pi_{\theta_s}(1)) \qquad \text{(By Equation (B.3))}
\end{aligned}$$

Next, we will prove $\sum_{s=1}^{\infty} \pi_{\theta_s}(1) = \infty$. Since $P_t > 0$, $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}$ is monotonically increasing. Hence, we have that $\frac{\pi_{\theta_{t+1}}(3)}{\pi_{\theta_{t+1}}(1)} < \frac{\pi_{\theta_t}(3)}{\pi_{\theta_t}(1)}$ for all finite $t \geq 1$. As a result,

$$\begin{aligned}
\sum_{s=1}^{t} (1 - \pi_{\theta_s}(2)) &= \sum_{s=1}^{t} (\pi_{\theta_s}(1) + \pi_{\theta_s}(3)) \\
&= \sum_{s=1}^{t} \left(\pi_{\theta_s}(1) + \pi_{\theta_s}(1) \frac{\pi_{\theta_s}(3)}{\pi_{\theta_s}(1)} \right) \\
&< \sum_{s=1}^{t} \left(\pi_{\theta_s}(1) + \pi_{\theta_s}(1) \frac{\pi_{\theta_1}(3)}{\pi_{\theta_1}(1)} \right) \\
&= \left(1 + \frac{\pi_{\theta_1}(3)}{\pi_{\theta_1}(1)} \right) \sum_{s=1}^{t} \pi_{\theta_s}(1),
\end{aligned}$$

For the LHS, Lemma 32 shows that $\sum_{s=1}^{\infty}(1 - \pi_{\theta_s}(2)) = \infty$. Therefore, we have that $\sum_{s=1}^{\infty} \pi_{\theta_s}(1) = \infty$. Using the equation above, we conclude that $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \to \infty$ as $t \to \infty$. Moreover,

$$\begin{aligned}
r(2) - \langle \pi_{\theta_t}, r \rangle &= \pi_{\theta_t}(1) \, (r(2) - r(1)) + \pi_{\theta_t}(3) \, (r(2) - r(3)) \\
&= \pi_{\theta_t}(3) \, (r(2) - r(3)) \Big[- \underbrace{\frac{r(1) - r(2)}{r(2) - r(3)}}_{>0} \underbrace{\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}}_{\to \infty} + 1 \Big] \\
&< 0. \qquad \text{(for large enough } t\text{)}
\end{aligned}$$

Therefore, we know that $\langle \pi_{\theta_t}, r \rangle > r(2)$ for all large enough t. This, combined with Equation (B.1), contradicts our assumption that $\pi_{\theta_t}(2) \to 1$ as $t \to \infty$.

Putting everything together, we can draw the conclusion that $\pi_{\theta_t}(1) \to 1$ as $t \to \infty$. □

Proposition 5. *Given a reward vector $r \in \mathbb{R}^3$ and a feature matrix $X \in \mathbb{R}^{3\times d}$ such that Assumptions 7 and 8 are satisfied but Assumption 9 is not. Using Algorithm 3 with a constant learning rate as in Equation (4.5) and initialization $\theta_1 = c\,(x_3 - x_1)$, such that $c > \frac{-\log(m)}{\|x_3 - x_1\|_2^2}$, where $m = \frac{\langle x_3 - x_2, x_1 - x_3\rangle}{\langle x_1 - x_2, x_1 - x_3\rangle} \frac{\langle \pi_{\theta_1}, r\rangle - r(3)}{r(1) - \langle \pi_{\theta_1}, r\rangle}$ fails to converge to the optimal policy.*

Proof. Consider

$$X = \begin{bmatrix} x_1^\top \\ x_2^\top \\ x_3^\top \end{bmatrix}, \text{ and } r = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \end{bmatrix},$$

where $x_i \in \mathbb{R}^d$ for all $i \in [K]$ and $r(1) > r(2) > r(3)$. Based on Algorithm 3, we have,

$$X\theta_{t+1} = X\theta_t + \eta\, XX^\top \left(\operatorname{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right),$$

where

$$XX^\top = \begin{bmatrix} x_1^\top x_1 & x_1^\top x_2 & x_1^\top x_3 \\ x_2^\top x_1 & x_2^\top x_2 & x_2^\top x_3 \\ x_3^\top x_1 & x_3^\top x_2 & x_3^\top x_3 \end{bmatrix}.$$

We show that if $\langle x_2 - x_3, x_1 - x_3\rangle < 0$, then there exists an initialization such that global convergence cannot happen. To show this, we choose an appropriate initialization θ_1 such that $\frac{\pi_{\theta_1}(1)}{\pi_{\theta_1}(3)} < m$, where

$$m = \frac{\langle x_3 - x_2, x_1 - x_3\rangle}{\langle x_1 - x_2, x_1 - x_3\rangle} \frac{\langle \pi_{\theta_1}, r\rangle - r(3)}{r(1) - \langle \pi_{\theta_1}, r\rangle},$$

and that if $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < m$ then $\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} < \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < m$. This would mean that $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < m$ for all t and $\pi_{\theta_t}(1) \not\to 1$ as $t \to \infty$.

We have,

$$\begin{aligned}
\frac{\pi_{\theta_1}(1)}{\pi_{\theta_1}(3)} &= \exp([X\theta_1](1) - [X\theta_1](3)) \\
&= \exp(\langle x_1 - x_3, \theta_1\rangle) \\
&= \exp(-c\,\|x_3 - x_1\|_2^2) && \text{(Since } \theta_1 = c\,(x_3 - x_1)\text{)} \\
&< \exp(\log(m)) && \text{(Since } c > \tfrac{-\log(m)}{\|x_3 - x_1\|_2^2}\text{)} \\
&= m.
\end{aligned}$$

Now, suppose that $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < m$. We have

$$\begin{aligned}\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} <& \frac{\langle x_3 - x_2, x_1 - x_3\rangle}{\langle x_1 - x_2, x_1 - x_3\rangle} \frac{\langle \pi_{\theta_1}, r\rangle - r(3)}{r(1) - \langle \pi_{\theta_1}, r\rangle} \\ \leq& \frac{\langle x_3 - x_2, x_1 - x_3\rangle}{\langle x_1 - x_2, x_1 - x_3\rangle} \frac{\langle \pi_{\theta_t}, r\rangle - r(3)}{r(1) - \langle \pi_{\theta_t}, r\rangle}. \qquad \text{(Due to monotonicity)}\end{aligned}$$

Furthermore,

$$\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} = \exp([X\theta_{t+1}](1) - [X\theta_{t+1}](3)),$$

and (from the expression of $XX^\top$ and the update),

$$\begin{aligned}[X\theta_{t+1}](1) - [X\theta_{t+1}](3) = [X\theta_t](1) - [X\theta_t](3) \\ + \eta \sum_{i=1}^{3} \langle x_i, x_1 - x_3\rangle \pi_{\theta_t}(i) \cdot (r(i) - \pi_{\theta_t}^\top r).\end{aligned}$$

If $\langle x_2 - x_3, x_1 - x_3\rangle < 0$, then we have, $\langle x_3 - x_2, x_1 - x_3\rangle > 0$, and,

$$\begin{aligned}\langle x_1 - x_2, x_1 - x_3\rangle &= \langle x_1 - x_3, x_1 - x_3\rangle + \langle x_3 - x_2, x_1 - x_3\rangle \\ &\geq \langle x_3 - x_2, x_1 - x_3\rangle > 0.\end{aligned}$$

Therefore, we have,

$$\begin{aligned}&\sum_{i=1}^{3} \langle x_i, x_1 - x_3\rangle \pi_{\theta_t}(i) \left(r(i) - \langle \pi_{\theta_t}, r\rangle\right. \\ &= \langle x_1 - x_2, x_1 - x_3\rangle \pi_{\theta_t}(1) \left(r(1) - \langle \pi_{\theta_t}, r\rangle\right) + \langle x_3 - x_2, x_1 - x_3\rangle \pi_{\theta_t}(3) \left(r(3) - \langle \pi_{\theta_t}, r\rangle\right) \\ &= -\underbrace{\langle x_3 - x_2, x_1 - x_3\rangle}_{>0} \pi_{\theta_t}(3) \left(\langle \pi_{\theta_t}, r\rangle - r(3)\right) \left[- \frac{\langle x_1 - x_2, x_1 - x_3\rangle}{\langle x_3 - x_2, x_1 - x_3\rangle} \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \frac{r(1) - \langle \pi_{\theta_t}, r\rangle}{\langle \pi_{\theta_t}, r\rangle - r(3)} + 1 \right] \\ &< -\underbrace{\langle x_3 - x_2, x_1 - x_3\rangle}_{>0} \pi_{\theta_t}(3) \left(\langle \pi_{\theta_t}, r\rangle - r(3)\right) [-1 + 1] \\ &\qquad\qquad \text{(Since } \tfrac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < \tfrac{\langle x_3 - x_2, x_1 - x_3\rangle}{\langle x_1 - x_2, x_1 - x_3\rangle} \tfrac{\langle \pi_{\theta_t}, r\rangle - r(3)}{r(1) - \langle \pi_{\theta_t}, r\rangle}) \\ &= 0\end{aligned}$$

which implies that,

$$\begin{aligned}\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} &= \exp([X\theta_{t+1}](1) - [X\theta_{t+1}](3)) \\ &= \exp([X\theta_t](1) - [X\theta_t](3)) + \eta \sum_{i=1}^{3} \langle x_i, x_1 - x_3\rangle \pi_{\theta_t}(i) \left(r(i) - \langle \pi_{\theta_t}, r\rangle\right) \\ &< \exp([X\theta_t](1) - [X\theta_t](3)) = \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}.\end{aligned}$$

□

Proposition 6. *Let $K = 3$, $d = 2$, $X^\top = \begin{bmatrix} 3 & 5 & 1 \\ 4 & 6 & 2 \end{bmatrix} \in \mathbb{R}^{d \times K}$, and $r = (3, 2, 1)^\top$. In this case, Assumptions 7 and 9 are satisfied, but Assumption 8 is not, and the features do not allow the optimal reward to be achieved for any set of finite or infinite parameters. Therefore, Algorithm 3 does not achieve global convergence for any initialization θ_1.*

Proof. We first show that Assumption 9 is satisfied, but Assumption 8 is not. We have $\langle x_2 - x_3, x_1 - x_3 \rangle = 16 > 0$, so Assumption 9 is satisfied. Now, suppose that $r' = Xw$ preserves the reward ordering for $w = (w(1), w(2))^\top$. In that case, the order of the optimal arm must also be preserved, i.e. $r'(1) > r'(2)$ and $r'(1) > r'(3)$. Therefore,

$$\begin{aligned} & \langle x_1, w \rangle > \langle x_2, w \rangle \quad \text{and} \quad \langle x_1, w \rangle > \langle x_3, w \rangle \\ \implies & 3\,w(1) + 4\,w(2) > 5\,w(1) + 6\,w(2) \quad \text{and} \quad 3\,w(1) + 4\,w(2) > w(1) + 2\,w(2) \\ \implies & w(1) + w(2) < 0 \quad \text{and} \quad w(1) + w(2) > 0 \end{aligned}$$

Therefore, there is no w that preserves the order of the optimal arm, so Assumption 8 is not satisfied. Furthermore, to achieve the optimal reward, we need parameters θ, such that

$$\begin{aligned} & \pi_\theta(1) >> \pi_\theta(2) \quad \text{and} \quad \pi_\theta(1) >> \pi_\theta(3) \\ \implies & [X\theta](1) >> [X\theta](2) \quad \text{and} \quad [X\theta](1) >> [X\theta](3) \\ \implies & \langle x_1, \theta \rangle >> \langle x_2, \theta \rangle \quad \text{and} \quad \langle x_1, \theta \rangle >> \langle x_3, \theta \rangle \\ \implies & 3\,\theta(1) + 4\,\theta(2) >> 5\,\theta(1) + 6\,\theta(2) \quad \text{and} \quad 3\,\theta(1) + 4\,\theta(2) >> \theta(1) + 2\,\theta(2) \\ \implies & \theta(1) + \theta(2) << 0 \quad \text{and} \quad \theta(1) + \theta(2) >> 0 \end{aligned}$$

Therefore, such a θ also does not exist, and the optimal reward cannot be achieved for any set of parameters. Hence, Algorithm 3 does not achieve global convergence for any initialization. □

B.3.2 Global Convergence for all $K \geq 3$

Theorem 7. Given a reward vector $r \in \mathbb{R}^K$ and a feature matrix $X \in \mathbb{R}^{K \times d}$ such that Assumptions 7, 8 and 10 are satisfied,using Algorithm 3 with a constant learning rate as in Equation (4.5) converges to the optimal policy.

Proof. Without the loss of generality, we assume $r(1) > r(2) > \cdots > r(K)$ as ties between distinct actions do not occur under Assumption 7. Additionally under Assumption 8, according to Lemma 1, we know that for all finite $t \geq 1$,

$$\langle \pi_{\theta_{t+1}}, r \rangle > \langle \pi_{\theta_t}, r \rangle, \tag{B.4}$$

and $\pi_{\theta_t}(a) \to 1$ as $t \to \infty$ for some action $a \in [K]$. For any bounded initialization θ_1, we have

$$\langle \pi_{\theta_1}, r \rangle > r(K),$$

The above two inequalities imply that $\langle \pi_{\theta_t}, r \rangle \not\to r(K)$ and hence $\pi_{\theta_t}(K) \not\to 1$ as $t \to \infty$. Next, we show that $\langle \pi_{\theta_t}, r \rangle \not\to r(a)$ for any $a \in \{2, 3, \ldots, K-1\}$ as $t \to \infty$.

We will prove this by contradiction. For this, in the subsequent proof, we assume that $\langle \pi_{\theta_t}, r \rangle \to r(a)$ as $t \to \infty$ for some $a \in \{2, 3, \ldots, K-1\}$. Therefore, there exists a large enough finite τ such that for all finite $t \geq \tau$, $r(a) > \langle \pi_{\theta_t}, r \rangle > r(a+1)$.

We will first prove that $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(k)} \to \infty$ as $t \to \infty$ for all $k \in [a+1, K]$. Considering a fixed arm $k \in [a+1, K]$, we have, for all finite $t \geq \tau$,

$$\begin{aligned}
\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(k)} &= \exp([X\,\theta_{t+1}](1) - [X\theta_{t+1}](k)) \\
&= \exp\left([X\,\theta_t](1) - [X\theta_t](k) + \eta \left(\sum_{i=1}^{K} \langle x_i, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) \right) \right) \\
&\qquad\qquad \text{(By the update in Algorithm 3)} \\
&= \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(k)} \exp\left(\eta \underbrace{\left(\sum_{i=1}^{K} \langle x_i, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) \right)}_{:= P_t} \right), \qquad \text{(B.5)}
\end{aligned}$$

and the sign of P_t will dictate whether $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(k)}$ will increase or decrease.

Next, we have, for all finite $t \geq \tau$,

$$\begin{aligned}
P_t &= \sum_{i=1}^{K} \langle x_i, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) \\
&= \sum_{\substack{i=1 \\ i \neq a}}^{K} \langle x_i - x_a, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) \\
&\qquad\qquad \text{(Since } \textstyle\sum_{i=1}^{K} \langle x_a, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle) = 0\text{)} \\
&= \sum_{i=1}^{a-1} \underbrace{\langle x_i - x_a, x_1 - x_k \rangle}_{\substack{>0 \text{ due to Assumption 10} \\ (\text{since } i<a \text{ and } k \geq a+1 > a)}} \pi_{\theta_t}(i) \underbrace{(r(i) - \langle \pi_{\theta_t}, r \rangle)}_{>0 \text{ (since } i<a)} \\
&\quad + \sum_{i=a+1}^{K} \underbrace{\langle x_a - x_i, x_1 - x_k \rangle}_{\substack{>0 \text{ due to Assumption 10} \\ (\text{since } i>a \text{ and } k \geq a+1 > a)}} \pi_{\theta_t}(i) \underbrace{(\langle \pi_{\theta_t}, r \rangle - r(i))}_{>0 \text{ (since } i>a)} \\
&> \sum_{i=1}^{a-1} \langle x_i - x_a, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (r(i) - r(a)) \\
&\quad + \sum_{i=a+1}^{K} \langle x_a - x_i, x_1 - x_k \rangle \, \pi_{\theta_t}(i) \, (\langle \pi_{\theta_\tau}, r \rangle - r(i)) \\
&\qquad\qquad \text{(Since } r(a) > \langle \pi_{\theta_t}, r \rangle \text{ and } \langle \pi_{\theta_t}, r \rangle \geq \langle \pi_{\theta_\tau}, r \rangle \text{ for all finite } t \geq \tau\text{.)}
\end{aligned}$$

Define

$$C_1 := \min_{1\le i\le a-1} \langle x_i - x_a, x_1 - x_k\rangle \left(r(i) - r(a)\right) > 0,$$
$$C_2^\tau := \min_{a+1\le i\le K} \langle x_a - x_i, x_1 - x_k\rangle \left(\langle \pi_{\theta_\tau}, r\rangle - r(i)\right) > 0.$$

Hence, we have

$$\begin{aligned} P_t &> C_1 \sum_{i=1}^{a-1} \pi_{\theta_t}(i) + C_2^\tau \sum_{i=a+1}^{K} \pi_{\theta_t}(i) \\ &> C^\tau \sum_{i\neq a} \pi_{\theta_t}(i) && (\text{Let } C^\tau := \min\{C_1, C_2^\tau\} > 0) \\ &= C^\tau \left(1 - \pi_{\theta_t}(a)\right). && \text{(B.6)} \end{aligned}$$

By recursing Equation (B.5), we get that, for all finite $t \ge \tau$,

$$\begin{aligned} \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(k)} &= \frac{\pi_{\theta_\tau}(1)}{\pi_{\theta_\tau}(k)} \exp(\eta \sum_{s=\tau}^{t-1} P_s) \\ &> \frac{\pi_{\theta_\tau}(1)}{\pi_{\theta_\tau}(k)} \exp(\eta\, C^\tau \sum_{s=\tau}^{t-1} (1 - \pi_{\theta_s}(a))). && (\text{By Equation (B.6)}) \end{aligned}$$

Lemma 32 shows that for any $i \in [K]$, $\sum_{s=1}^{\infty}(1-\pi_{\theta_s}(i)) = \infty$. Combining the above equations, we conclude that $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(k)} \to \infty$ and hence $\frac{\pi_{\theta_t}(k)}{\pi_{\theta_t}(1)} \to 0$ as $t \to \infty$ for all $k \in [a+1, K]$. As a result, there exists a $\tau' \ge \tau$ such that

$$\begin{aligned} & r(a) - \left\langle \pi_{\theta_{\tau'}}, r \right\rangle \\ &= \sum_{i=1}^{K} \pi_{\theta_{\tau'}}(i) \left(r(a) - r(i)\right) = \sum_{i=1}^{a-1} \pi_{\theta_{\tau'}}(i) \underbrace{\left(r(a) - r(i)\right)}_{<0} + \sum_{i=a+1}^{K} \pi_{\theta_{\tau'}}(i) \underbrace{\left(r(a) - r(i)\right)}_{>0} \\ &< \pi_{\theta_{\tau'}}(1) \left(r(a) - r(1)\right) + \sum_{i=a+1}^{K} \pi_{\theta_{\tau'}}(i) \left(r(a) - r(i)\right) \\ &= \pi_{\theta_{\tau'}}(1) \left(r(1) - r(a)\right) \left[\sum_{i=a+1}^{K} \underbrace{\frac{\pi_{\theta_{\tau'}}(i)}{\pi_{\theta_{\tau'}}(1)}}_{\to 0} \underbrace{\frac{r(a) - r(i)}{r(1) - r(a)}}_{>0} - 1 \right] \\ &< 0. && (\tau' \text{ is large enough}) \end{aligned}$$

Therefore, we know that $\left\langle \pi_{\theta_{\tau'}}, r \right\rangle > r(a)$. Combined with Equation (B.4), we know that for all $t \ge \tau'$, $\langle \pi_{\theta_t}, r\rangle > r(a)$. This contradicts the assumption that $\langle \pi_{\theta_t}, r\rangle \to r(a)$ as $t \to \infty$. This implies that $\langle \pi_{\theta_t}, r\rangle \not\to r(a)$ and hence $\pi_{\theta_t}(a) \not\to 1$ as $t \to \infty$ for all $a \in \{2, 3, \ldots, K\}$. Hence, the only possible scenario is $\pi_{\theta_t}(1) \to 1$ as $t \to \infty$, completing the proof.

□

B.3.3 Additional Lemmas

Lemma 1. *Assuming Assumptions 7 and 8 are satisfied, using Algorithm 3 with the following constant learning rate,*

$$0 < \eta < \frac{4}{9 \left\| r \right\|_\infty \lambda_{\max}(X^\top X)}, \tag{4.5}$$

ensures (i) $\langle \pi_{\theta_{t+1}}, r \rangle > \langle \pi_{\theta_t}, r \rangle$ *for all* $t \geq 1$ *and (ii)* $\pi_{\theta_t}(a) \to 1$ *for an arm* $a \in [K]$ *as* $t \to \infty$.

Proof. According to Lemma 45, we have, for all $t \geq 1$,

$$\left| \left\langle \pi_{\theta_{t+1}} - \pi_{\theta_t}, r \right\rangle - \left\langle \frac{d \left\langle \pi_{\theta_t}, r \right\rangle}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \leq \frac{9}{4} \left\| r \right\|_\infty \lambda_{\max}(X^\top X) \left\| \theta_{t+1} - \theta_t \right\|_2^2,$$

which implies that,

$$\begin{aligned} \left\langle \pi_{\theta_{t+1}}, r \right\rangle - \left\langle \pi_{\theta_t}, r \right\rangle &\geq \left\langle \frac{d\langle \pi_{\theta_t}, r \rangle}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle - \frac{9}{4} \left\| r \right\|_\infty \lambda_{\max}(X^\top X) \left\| \theta_{t+1} - \theta_t \right\|_2^2 \\ &= \left(\eta - \eta^2 \frac{9}{4} \left\| r \right\|_\infty \lambda_{\max}(X^\top X) \right) \left\| \frac{d \left\langle \pi_{\theta_t}, r \right\rangle}{d\theta_t} \right\|_2^2. \end{aligned}$$

Using a constant learning rate,

$$0 < \eta < \frac{4}{9 \left\| r \right\|_\infty \lambda_{\max}(X^\top X)},$$

we have,

$$\left\langle \pi_{\theta_{t+1}}, r \right\rangle - \left\langle \pi_{\theta_t}, r \right\rangle \geq \eta \left(1 - \eta \frac{9 \left\| r \right\|_\infty \lambda_{\max}(X^\top X)}{4} \right) \left\| \frac{d \left\langle \pi_{\theta_t}, r \right\rangle}{d\theta_t} \right\|_2^2 \geq 0. \tag{B.7}$$

Note that $\langle \pi_{\theta_t}, r \rangle \leq r(a^*) < \infty$. According to the monotone convergence, $\langle \pi_{\theta_t}, r \rangle \to c \leq r(a^*)$ as $t \to \infty$. According to Equation (B.7), we have,

$$\lim_{t \to \infty} \left\| \frac{d \left\langle \pi_{\theta_t}, r \right\rangle}{d\theta_t} \right\|_2^2 = 0. \tag{B.8}$$

Next, we prove that there is no stationary points in finite region by contradiction. Suppose there exists $\theta' \in \mathbb{R}^d$ ($\|\theta'\|_2 < \infty$), such that,

$$\frac{d\langle \pi_{\theta'}, r \rangle}{d\theta'} = X^\top \left(\operatorname{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) \, r = \mathbf{0}. \tag{B.9}$$

Taking inner product with $w \in \mathbb{R}^K$ on both sides of Equation (B.9), we have,

$$\begin{aligned} w^\top X^\top \left(\operatorname{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) \, r &= r'^\top \left(\operatorname{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) \, r \qquad (r' \coloneqq Xw) \\ &= w^\top \mathbf{0} = 0. \end{aligned} \tag{B.10}$$

Since $\|\theta'\|_2 < \infty$ and X is bounded ($\max_{i\in[K],\ j\in[d]} |X_{i,j}| \le C$ for some $C < \infty$), we have, for all $i \in [K]$,

$$\pi_{\theta'}(i) = \frac{\exp\{[X\theta'](i)\}}{\sum_{j\in[K]} \exp\{[X\theta'](j)\}} > 0. \tag{B.11}$$

Next, according to Lemma 31, we have,

$$r'^{\top} \left(\mathrm{diag}(\pi_{\theta'}) - \pi_{\theta'}\pi_{\theta'}^{\top}\right)\ r = \sum_{i=1}^{K-1} \pi_{\theta'}(i) \sum_{j=i+1}^{K} \pi_{\theta'}(j)\ \big(r'(i) - r'(j)\big)\ \big(r(i) - r(j)\big). \tag{B.12}$$

Given any non-trivial reward vector, i.e., $r \neq c\,\mathbf{1}$ for any $c \in \mathbb{R}$, since $r' \in \mathbb{R}^K$ preserves the order of $r \in \mathbb{R}^K$, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ iff $r'(i) > r'(j)$, we have, for all $i, j \in [K]$,

$$\big(r'(i) - r'(j)\big)\ \big(r(i) - r(j)\big) > 0. \tag{B.13}$$

On the other hand, since $r \neq c\,\mathbf{1}$, there exists at least one pair of $i \neq j$, such that,

$$\big(r'(i) - r'(j)\big)\ \big(r(i) - r(j)\big) > 0. \tag{B.14}$$

Combining Equations (B.9) to (B.14), we have,

$$\begin{aligned} 0 = w^{\top}\mathbf{1} &= w^{\top}\Big(\frac{d\ \langle \pi_{\theta'}, r\rangle}{d\theta'}\Big) \\ &= w^{\top} X^{\top} \left(\mathrm{diag}(\pi_{\theta'}) - \pi_{\theta'}\pi_{\theta'}^{\top}\right)\ r \\ &= r'^{\top} \left(\mathrm{diag}(\pi_{\theta'}) - \pi_{\theta'}\pi_{\theta'}^{\top}\right)\ r \\ &> 0, \end{aligned}$$

which is a contradiction. Therefore, for any $\theta' \in \mathbb{R}^d$ ($\|\theta'\|_2 < \infty$), θ' is not a stationary point.

Next, we show that $\|\theta_t\|_2 \to \infty$ as $t \to \infty$ also by contradiction. Suppose there exists $C < \infty$, such that for all $t \ge 1$,

$$\theta_t \in S_C \coloneqq \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le C\}.$$

From the above arguments, we have, for all $\theta \in S_C$, $\left\|\frac{d\ \langle \pi_{\theta_t}, r\rangle}{d\theta}\right\|_2 > 0$. Since S_C is compact, we have,

$$\inf_{\theta\in S_C} \left\|\frac{d\ \langle \pi_{\theta_t}, r\rangle}{d\theta}\right\|_2 \ge \varepsilon > 0,$$

for some $\varepsilon > 0$, which implies that, for all $t \ge 1$,

$$\left\|\frac{d\ \langle \pi_{\theta_t}, r\rangle}{d\theta_t}\right\|_2 \ge \varepsilon > 0,$$

contradicting Equation (B.8). Therefore, we have, $\|\theta_t\|_2 \to \infty$ as $t \to \infty$.

Next, we show that $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$. Suppose $\pi_{\theta_t}(i) \not\to 1$ for any action $i \in [K]$, then there exists at least two different actions $j \neq k$ such that $\pi_{\theta_t}(j) \not\to 0$ and $\pi_{\theta_t}(k) \not\to 0$. Using similar calculations in Equation (B.12), we have, $\left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \not\to 0$ as $t \to \infty$, contradicting Equation (B.8). Therefore, $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$, i.e. π_{θ_t} approaches a one-hot policy. □

Lemma 31 (Alternative expression of co-variance)**.** *Given any vectors $x \in \mathbb{R}^K$, $y \in \mathbb{R}^K$, we have, for all policy $\pi \in \Delta(K)$,*

$$\mathrm{Cov}_\pi(x, y) = \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j)\ (x(i) - x(j))\ (y(i) - y(j)).$$

Proof. Note that, $\mathrm{Cov}_\pi(x, y) = \left\langle x, \left(\mathrm{diag}(\pi) - \pi\pi^\top\right) y \right\rangle$. Next, we have,

$$\begin{aligned}
\left\langle x, \left(\mathrm{diag}(\pi) - \pi\pi^\top\right) y \right\rangle &= \sum_{i=1}^{K} \pi(i)\, x(i)\, y(i) - \sum_{i=1}^{K} \pi(i)\, y(i) \sum_{j=1}^{K} \pi(j)\, x(j) \\
&= \sum_{i=1}^{K} \pi(i)\, x(i)\, y(i) - \sum_{i=1}^{K} \pi(i)^2\, x(i)\, y(i) - \sum_{i=1}^{K} \pi(i)\, y(i) \sum_{j \neq i} \pi(j)\, x(j) \\
&= \sum_{i=1}^{K} \pi(i)\, x(i)\, y(i)\ (1 - \pi(i)) - \sum_{i=1}^{K} \pi(i)\, y(i) \sum_{j \neq i} \pi(j)\, x(j) \\
&= \sum_{i=1}^{K} \pi(i)\, x(i)\, y(i) \sum_{j \neq i} \pi(j) - \sum_{i=1}^{K} \pi(i)\, y(i) \sum_{j \neq i} \pi(j)\, x(j) \\
&= \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j)\ (x(i)\, y(i) + x(j)\, y(j)) - \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j)\ (x(j)\, y(i) + x(i)\, y(j)) \\
&= \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j)\ (x(i) - x(j))\ (y(i) - y(j)),
\end{aligned}$$

finishing the proofs. □

Lemma 32. *Let Assumptions 7 and 8 hold for a given reward vector $r \in \mathbb{R}^d$ and a feature matrix $X \in \mathbb{R}^{K \times d}$. Then Algorithm 3 guarantees that $\sum_{t=1}^{\infty} (1 - \pi_{\theta_t}(a)) = \infty$ for all $a \in [K]$.*

Proof. We prove this by contradiction. Without the loss of generality, we assume $r(1) > r(2) > \cdots > r(K)$. Under Assumptions 7 and 8, according to Lemma 1, $\pi_{\theta_t}(a) \to 1$ as $t \to \infty$ for some action $a \in [K]$. For a fixed $a \in [K]$, suppose $\sum_{t \geq 1} (1 - \pi_{\theta_t}(a)) < \infty$. Then

for all $a' \in [K]$, we have,

$$
\begin{aligned}
&|[X\theta_{t+1}](a') - [X\theta_t](a')| \\
&\quad = \eta \,|\sum_{i=1}^{K} \langle x_{a'}, x_i \rangle \, \pi_{\theta_t}(i) \, (r(i) - \langle \pi_{\theta_t}, r \rangle)| \\
&\quad \le C \sum_{i=1}^{K} \pi_{\theta_t}(i) \left| (r(i) - \langle \pi_{\theta_t}, r \rangle) \right| \\
&\qquad\qquad\qquad\qquad \text{(Let } C := \eta \, \max_{i \in [K]} |\langle x_{a'}, x_i \rangle| > 0 \text{ and using triangle inequality)} \\
&\quad \le C \left[\sum_{\substack{i=1 \\ i \neq a}}^{K} \pi_{\theta_t}(i) \left| \underbrace{(r(i) - \langle \pi_{\theta_t}, r \rangle)}_{< r(1) - r(K)} \right| + \underbrace{\pi_{\theta_t}(a)}_{\le 1} \, |(r(a) - \langle \pi_{\theta_t}, r \rangle| \right] \\
&\quad \le C \left((r(1) - r(K)) \sum_{\substack{i=1 \\ i \neq a}}^{K} \pi_{\theta_t}(i) + |r(a) - \langle \pi_{\theta_t}, r \rangle| \right) \\
&\quad = C \left((r(1) - r(K)) \sum_{\substack{i=1 \\ i \neq a}}^{K} \pi_{\theta_t}(i) + |\sum_{\substack{i=1 \\ i \neq a}}^{K} \pi_{\theta_t}(i) \, (r(a) - r(i))| \right) \\
&\quad \le C \left((r(1) - r(K)) \sum_{i=1, i \neq a}^{K} \pi_{\theta_t}(i) + \sum_{i=1, i \neq a}^{K} \pi_{\theta_t}(i) |(r(a) - r(i))| \right) \quad \text{(triangle inequality)} \\
&\quad \le C \left((r(1) - r(K)) \sum_{i=1, i \neq a}^{K} \pi_{\theta_t}(i) + (r(1) - r(K)) \sum_{i=1, i \neq a}^{K} \pi_{\theta_t}(i) \right) \\
&\quad \le 2\, C \, (r(1) - r(K)) \, (1 - \pi_{\theta_t}(a)),
\end{aligned}
$$

which implies that, for all $t > 1$,

$$
|X\theta_t(a') - X\theta_1(a')| \le 2\, C \, (r(1) - r(K)) \sum_{s=1}^{t-1} (1 - \pi_{\theta_s}(a)).
$$

Therefore, if $\sum_{t \ge 1} (1 - \pi_{\theta_t}(a)) < \infty$, then we have,

$$
\sup_{t \ge 1} |X\theta_t(a')| \le \sup_{t \ge 1} |X\theta_t(a') - X\theta_1(a')| + |X\theta_1(a')| < \infty,
$$

Therefore, there exists $\epsilon > 0$, such that, for all $a \in [K]$,

$$
\inf_{t \ge 1} \pi_{\theta_t}(a) = \inf_{t \ge 1} \frac{\exp\{[X\theta_t](a)\}}{\sum_{a' \in [K]} \exp\{[X\theta_t](a')\}} \ge \epsilon > 0,
$$

This implies that the algorithm does not converge to a one-hot policy, which leads to a contradiction. □

B.4 Proofs of Section 4.5

B.4.1 Asymptotic Global Convergence

Theorem 9. Given a reward vector $r \in \mathbb{R}^K$ and a feature matrix $X \in \mathbb{R}^{K \times d}$ such that Assumptions 7 and 10 are satisfied, Algorithm 4 with the constant learning rate as in Equation (4.9), we have, almost surely, $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$.

Proof. Without the loss of generality, we assume $r(1) > r(2) > \cdots > r(K)$ as ties between distinct actions do not occur under Assumption 7. In this setting, $a^* = \arg\max_a \pi^*(a) = 1$. According to Lemma 3, we know that there exists an action $a \in [K]$, such that, almost surely, $\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle = r(a)$. We will prove that almost surely, $\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle = r(1)$. We will prove this by contradiction. For this, assume $\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle = r(k)$ where $k > 1$.
We define $N_\infty(a)$ as the number of times action a has been sampled as $t \to \infty$. Define $\mathcal{A}_\infty$ as the set of actions that are sampled infinitely many times as $t \to \infty$, i.e.,

$$\mathcal{A}_\infty := \{a \in [K] \mid N_\infty(a) = \infty\}.$$

According to Lemma 38, there exists a finite large enough τ_{i_1} such that for all $t \geq \tau_{i_1}$,

$$\langle \pi_{\theta_t}, r \rangle > r(i_1),$$

where $i_1 := \arg\min_{a \in \mathcal{A}_\infty} r(a)$. By combining the above inequality with Lemma 33, we can conclude that

$$\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle > r(i_1).$$

This implies that for all $a \geq i_1$, $\pi_{\theta_t}(a) \not\to 1$ as $t \to \infty$. Hence, we only need to consider the arms $k < i_1$ in the subsequent proof. Under the assumption that $\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle = r(k)$, we know that there exists a $\tau > \tau_{i_1}$ such that for all large enough finite $t \geq \tau$,

$$r(k) > \langle \pi_{\theta_t}, r \rangle > r(k+1) \geq r(i_1). \tag{B.15}$$

Next, we will prove that $\lim_{t\geq 1} \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \to \infty$ for any arm $a > k$. Defining $z_t(a) := [X\theta_t](a)$ as the logit corresponding to arm a, we can express the ratio as,

$$\frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} = \exp([X\theta_t](a^*) - [X\theta_t](a)) = \exp(z_t(a^*) - z_t(a)). \tag{B.16}$$

Using the decomposition of the stochastic process in Section 4.5.1, setting $a_1 = a^*$ and $a_2 = a$ and recursing Equation (4.7) until $t = \tau$, we have

$$z_t(a^*) - z_t(a) = z_\tau(a^*) - z_\tau(a) + \underbrace{\sum_{s=\tau}^{t-1} [P_s(a^*) - P_s(a)]}_{\text{(i)}} + \underbrace{\sum_{s=\tau}^{t} [W_{s+1}(a^*) - W_{s+1}(a)]}_{\text{(ii)}}. \tag{B.17}$$

In the following proof, we will show that Term (i) dominates Term (ii). We first investigate Term (i), the cumulative progress and bound it similar to deterministic setting in Theorem 7.

$$
\begin{aligned}
&P_s(a^*) - P_s(a) = \mathbb{E}_s[z_{s+1}(a^*)] - z_s(a^*) - (\mathbb{E}_s[z_{s+1}(a)] - z_s(a)) \\
&= \mathbb{E}_s\left[[X\theta_{s+1}](a^*) - [X\theta_s](a^*)\right] - \mathbb{E}_s\left[[X\theta_{s+1}](a) - [X\theta_s](a)\right] && (z_s(a) := [X\theta_s](a)) \\
&= \eta \left\langle x_{a^*}, \mathbb{E}_s\left[\frac{d\langle \pi_{\theta_s}, \hat{r}_s\rangle}{d\theta_s}\right]\right\rangle - \eta \left\langle x_a, \mathbb{E}_s\left[\frac{d\langle \pi_{\theta_s}, \hat{r}_s\rangle}{d\theta_s}\right]\right\rangle && \text{(By the update in Algorithm 4)} \\
&= \eta \left\langle x_{a^*} - x_a, \frac{d\langle \pi_{\theta_s}, r\rangle}{d\theta_s}\right\rangle && \text{(By Lemma 44)} \\
&= \eta \sum_{i\in[K]} \langle x_i, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left(r(i) - \langle \pi_{\theta_s}, r\rangle\right) \\
& && \text{(Using the definition of the deterministic gradient)} \\
&= \eta \sum_{\substack{i\in[K] \\ i\neq k}} \langle x_i - x_k, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left(r(i) - \langle \pi_{\theta_s}, r\rangle\right) \\
& && \text{(Since } \textstyle\sum_{i\in[K]} \langle x_k, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left(r(i) - \langle \pi_{\theta_s}, r\rangle\right) = 0) \\
&= \eta \left[\sum_{i=1}^{k-1} \langle x_i - x_k, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left(r(i) - \langle \pi_{\theta_s}, r\rangle\right) + \sum_{i=k+1}^{K} \langle x_k - x_i, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left(\langle \pi_{\theta_s}, r\rangle - r(i)\right)\right] \\
&= \eta \Bigg[\sum_{i=1}^{k-1} \underbrace{\langle x_i - x_k, x_{a^*} - x_a\rangle}_{\substack{\geq 0 \text{ due to Assumption 10} \\ (\text{since } i<k<a)}} \pi_{\theta_s}(i) \underbrace{\left(r(i) - \langle \pi_{\theta_s}, r\rangle\right)}_{>0 \ (\text{since } \langle \pi_{\theta_s}, r\rangle < r(k) < r(i))} \\
&\quad + \sum_{i=k+1}^{K} \langle x_k - x_i, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left\langle \pi_{\theta_s}, r\right\rangle - r(i))\Bigg] \\
&> \eta \Bigg[\sum_{i=1}^{k-1} \underbrace{\langle x_i - x_k, x_{a^*} - x_a\rangle}_{\substack{\geq 0 \text{ due to Assumption 10} \\ (\text{since } i<k<a)}} \pi_{\theta_s}(i) \underbrace{\left(r(i) - r(k)\right)}_{>0 \ (\text{since } i<k)} \\
&\quad + \sum_{i=k+1}^{K} \underbrace{\langle x_k - x_i, x_{a^*} - x_a\rangle}_{\substack{\geq 0 \text{ due to Assumption 10} \\ (\text{since } a>k, i>k)}} \pi_{\theta_s}(i) \underbrace{\left\langle \pi_{\theta_s}, r\right\rangle - r(i))}_{>0 \ (\text{since } \langle \pi_{\theta_s}, r\rangle > r(k+1) \geq r(i))}\Bigg] \quad (\text{Since } \langle \pi_{\theta_s}, r\rangle < r(k)) \\
&> \eta \Bigg[\sum_{i=1}^{k-1} \langle x_i - x_k, x_{a^*} - x_a\rangle \, \pi_{\theta_s}(i) \left(r(i) - r(k)\right) \\
&\quad + \sum_{i=k+1}^{K} \langle x_k - x_i, x_{a^*} - x_a\rangle \, \pi_{\theta_t}(i) \left(\langle \pi_{\theta_\tau}, r\rangle - r(i)\right)\Bigg]
\end{aligned}
$$

According to Assumption 10, not all feature weights are strictly positive. Therefore, we define the set to represent the arms that contribute to the progress as:

$$\mathcal{X}_a(x_k) := \{i \in [K] \mid |\langle x_i - x_k, x_{a^*} - x_a\rangle| > 0\}.$$

Note that $\mathcal{X}_a(x_k)$ is non-empty since $\langle x_{a^*} - x_k, x_{a^*} - x_a\rangle > 0$ and $a^* \in \mathcal{X}_a(x_k)$. Additionally $k \notin \mathcal{X}_a(x_k)$ since $\langle x_k - x_k, x_{a^*} - x_a\rangle = 0$ We can then define that $C_3^\tau := \min\{C_1, C_2^\tau\} > 0$ where

$$C_1 := \min_{\substack{1\le i\le k-1\\ i\in\mathcal{X}_a(x_k)}} \langle x_i - x_k, x_{a^*} - x_k\rangle \left(r(i) - r(k)\right) > 0,$$

$$C_2^\tau := \min_{\substack{k+1\le i\le K\\ i\in\mathcal{X}_a(x_k)}} \langle x_k - x_i, x_{a^*} - x_k\rangle \left(\inf_{t>\tau} \langle \pi_{\theta_t}, r\rangle - r(i)\right) > 0.$$

Then, we have

$$\begin{aligned} P_s(a^*) - P_s(a) &> \eta \left[C_1 \sum_{\substack{i\le k-1\\ i\in\mathcal{X}_a(x_k)}} \pi_{\theta_s}(i) + C_2^\tau \sum_{\substack{i\ge k+1\\ i\in\mathcal{X}_a(x_k)}} \pi_{\theta_s}(i)\right] \\ &> \eta\, C_3^\tau \underbrace{\sum_{i\in\mathcal{X}_a(x_k)} \pi_{\theta_s}(i)}_{:=\Gamma_s}. \end{aligned} \tag{B.18}$$

By summing Equation (B.18) from τ to $t-1$, we get,

$$\sum_{s=\tau}^{t-1} \left[P_s(a^*) - P_s(a)\right] > \eta \sum_{s=\tau}^{t-1} C_3^\tau \Gamma_s. \tag{B.19}$$

Next, we bound Term (ii), the cumulative noise. We will first prove some useful properties of $W_s(a)$ which will be used to bound Term (ii). According to Corollary 5, we know that for a^* and a, $\mathbb{E}_s[W_{s+1}(a^*) - W_{s+1}(a)] = 0$, for all $s \ge 1$ and is bounded by

$$|W_{s+1}(a^*) - W_{s+1}(a)| \le 4\,\eta\, R_{\max} \|y_{a^*,a}\|_1 \le 4\,\eta\, R_{\max}\, C_4, \quad (\text{Let } C_4 := \max_a \|y_{a^*,a}\|_1 > 0)$$

where $y_{a^*,a} := (X - \mathbf{1}{x_k}^\top)(x_{a^*} - x_a)$.

Therefore, $\{|W_{s+1}(a^*) - W_{s+1}(a)|\}_{s\ge1}$ is a martingale difference sequence with respect to filtration $\{\mathcal{F}\}_{s\ge1}$ that can be normalized to be in the range of $[0, 1/2]$ since $W_{s+1}(a)$ is bounded. For this, define $\widetilde{W}_{s+1}(a^*, a) := \frac{|W_{s+1}(a^*) - W_{s+1}(a)|}{8\,\eta\, R_{\max}\, C_4}$. Additionally, we have

$$\begin{aligned} \mathrm{Var}[\widetilde{W}_{s+1}(a^*, a)] &= \frac{\mathrm{Var}[|W_{s+1}(a^*) - W_{s+1}(a)|]}{(8\,\eta\, R_{\max}\, C_4)^2} \\ &\le \frac{2\eta^2\, R_{\max}^2}{(8\,\eta\, R_{\max}\, C_4)^2} \sum_{\substack{j\in[K]\\ j\ne k}} \left(\langle x_j - x_k, x_{a^*} - x_a\rangle\right)^2 \pi_{\theta_s}(j)\left(1 - \pi_{\theta_s}(j)\right) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{By Corollary 5}) \\ &\le \frac{2\eta^2\, R_{\max}^2}{(8\,\eta\, R_{\max}\, C_4)^2} \sum_{\substack{j\in[K]\\ j\ne k}} \left(\langle x_j - x_k, x_{a^*} - x_a\rangle\right)^2 \pi_{\theta_s}(j) \qquad (1 - \pi_{\theta_s}(j) \le 1) \end{aligned}$$

Recall that $\mathcal{X}_a(x_k) := \{i \in [K] \mid |\langle x_i - x_k, x_{a^*} - x_a\rangle| > 0\}$

$$\leq \frac{2\eta^2 R_{\max}^2 C_5}{(8\,\eta\, R_{\max}\, C_4)^2} \sum_{j \in \mathcal{X}_a(x_k)} \pi_{\theta_s}(j)$$

(Let $C_5 := \max_{j \in \mathcal{X}_a(x_k)}(\langle x_j - x_k, x_{a^*} - x_a\rangle)^2$)

$$\leq \frac{C_5}{32\, C_4^2} \sum_{j \in \mathcal{X}_a(x_k)} \pi_{\theta_s}(j)$$

Recall that $\Gamma_s := \sum_{j \in \mathcal{X}_a(x_k)} \pi_{\theta_s}(j)$

$$= C_6\, \Gamma_s, \qquad \text{(Let } C_6 := \tfrac{C_5}{32\, C_4^2} > 0\text{)}$$

Using the above equation in combination with Lemma 43, for all $\delta \in (0, 1]$ and $t \geq \tau$, with probability $1 - \delta$,

$$|\widetilde{W}_{s+1}(a^*, a)| \leq 6 \sqrt{\left(C_6 \sum_{s=\tau}^{t} \Gamma_s + \frac{4}{3}\right) \log\left(\frac{C_6 \sum_{s=\tau}^{t} \Gamma_s + 1}{\delta}\right)} + 2 \log\left(\frac{1}{\delta}\right) + \frac{4}{3}\log(3).$$

Recall that $\widetilde{W}_{s+1}(a^*, a) := \frac{|W_{s+1}(a^*) - W_{s+1}(a)|}{8\,\eta\, R_{\max}\, C_4}$. Set $C_7 := 8\,\eta\, R_{\max}\, C_4$. Then, we have

$$\sum_{s=\tau}^{t} |W_{s+1}(a^*) - W_{s+1}(a)| \leq 6\, C_7 \sqrt{\left(C_6 \sum_{s=\tau}^{t} \Gamma_s + \frac{4}{3}\right) \log\left(\frac{C_6 \sum_{s=\tau}^{t} \Gamma_s + 1}{\delta}\right)} + 2C_7 \log\left(\frac{1}{\delta}\right) + \frac{4C_7}{3}\log(3). \tag{B.20}$$

Using the above results and combining it with Equation (B.17), we have

$$\begin{aligned} & z_t(a^*) - z_t(a) \\ &= z_\tau(a^*) - z_\tau(a) + \sum_{s=\tau}^{t-1} [P_s(a^*) - P_s(a)] + \sum_{s=\tau}^{t} [W_{s+1}(a^*) - W_{s+1}(a)] \\ &\geq z_\tau(a^*) - z_\tau(a) + \sum_{s=\tau}^{t-1} [P_s(a^*) - P_s(a)] - \sum_{s=\tau}^{t} |W_{s+1}(a^*) - W_{s+1}(a)| \end{aligned}$$

$(\forall u, v \in \mathbb{R},\, u - v \geq -|u - v|)$

Using Equation (B.19) to lower-bound the progress term,

$$\geq z_\tau(a^*) - z_\tau(a) + \eta\, C_3^\tau \sum_{s=\tau}^{t-1} \Gamma_s - \sum_{s=\tau}^{t} |W_{s+1}(a^*) - W_{s+1}(a)|$$

Using Equation (B.20) to lower-bound the cumulative noise term,

$$\begin{aligned}
&\geq z_\tau(a^*) - z_\tau(a) + \eta\, C_3^\tau \sum_{s=\tau}^{t-1} \Gamma_s \\
&\quad - 12\, C_7 \sqrt{\left(C_6 \sum_{s=\tau}^{t} \Gamma_s + \frac{4}{3}\right) \log\left(\frac{C_6 \sum_{s=\tau}^{t} \Gamma_s + 1}{\delta}\right)} \\
&\quad - 4C_7 \log\left(\frac{1}{\delta}\right) - \frac{8C_7}{3} \log(3)
\end{aligned} \tag{B.21}$$

Define,

$$\begin{aligned}
\mathcal{P}(n) &:= 12\, C_7 \sqrt{\left(C_6\, n + \frac{4}{3}\right) \log\left(\frac{C_6\, n + 1}{\delta}\right)} \\
\mathcal{Q}(n) &:= \eta\, C_3^\tau\, n
\end{aligned}$$

Let us characterize the order complexity of the above expressions in terms on n,

$$\begin{aligned}
\mathcal{P}(n) &\in \Theta(\sqrt{\log(n)\, n}), \\
\mathcal{Q}(n) &\in \Theta(n).
\end{aligned}$$

Additionally, we know that,

$$\lim_{n\to\infty} \frac{\mathcal{P}(n)}{\mathcal{Q}(n)} = \frac{\sqrt{\ln(n)\, n}}{n} = 0 \implies \mathcal{P}(n) \in o(\mathcal{Q}(n)).$$

This implies $\mathcal{Q}(n)$ dominates $\mathcal{P}(n)$ as $n \to \infty$. Additionally, note that

$$\begin{aligned}
\sum_{s=\tau}^{\infty} \Gamma_s &= \sum_{s=\tau}^{\infty} \sum_{i \in \mathcal{X}_a(x_k)} \pi_{\theta_s}(i) \\
&\geq \sum_{s=\tau}^{\infty} \pi_{\theta_s}(a^*) && \text{(Since } a^* \in \mathcal{X}_a(x_k)\text{)} \\
&= \infty. && \text{(Since } a^* \in \mathcal{A}_\infty \text{ and Lemma 42)}
\end{aligned}$$

Using these results with Equation (B.21) with $n = \sum_{s=\tau}^{\infty} \Gamma_s$ implies that $z_t(a^*) - z_t(a) \to \infty$ as $t \to \infty$. Using Equation (B.16), we conclude that for all arms $a > k$, almost surely,

$$\lim_{t\to\infty} \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \to \infty \implies \lim_{t\to\infty} \frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(1)} \to 0. \tag{B.22}$$

Hence, for all $k > 1$,

$$\begin{aligned}
r(k) - \langle \pi_{\theta_t}, r \rangle &= \sum_{i=1}^{K} \pi_{\theta_t}(i)\,(r(k) - r(i)) \\
&= \sum_{i=1}^{k-1} \pi_{\theta_t}(i)\,\underbrace{(r(k) - r(i))}_{<0} + \sum_{i=k+1}^{K} \pi_{\theta_t}(i)\,(r(k) - r(i)) \\
&< \pi_{\theta_t}(1)\,(r(k) - r(1)) + \sum_{i=k+1}^{K} \pi_{\theta_t}(i)\,(r(k) - r(i)) \\
&= \pi_{\theta_t}(1)\,\underbrace{(r(1) - r(k))}_{>0} \left[\sum_{i=k+1}^{K} \underbrace{\frac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(1)}}_{\to 0}\,\underbrace{\frac{r(k) - r(i)}{r(1) - r(k)}}_{>0} - 1 \right] \\
&< 0 \qquad \text{(For large enough } t \geq \tau\text{)}
\end{aligned}$$

This contradicts with the assumption that $\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle = r(k)$ where $k > 1$. Hence, almost surely, for all $k \neq 1$, $\pi_{\theta_t}(k) \not\to 1$, implying that $\pi_{\theta_t}(1) \to 1$ as $t \to \infty$.

□

B.4.2 Rate of Convergence

Theorem 10. Given a reward vector $r \in \mathbb{R}^K$ and a feature matrix $X \in \mathbb{R}^{K\times d}$ such that Assumptions 7 and 10 are satisfied, Algorithm 4 with the constant learning as in Equation (4.9) results in the following sub-linear convergence rate:

$$\mathbb{E}[\langle \pi^*, r \rangle - \langle \pi_{\theta_1}, r \rangle] \leq \frac{6\,\rho\,\kappa^2}{\mu\,T},$$

where $\rho \coloneqq \frac{8\,R_{\max}^3 K^{3/2}}{\Delta^2}$, $\kappa \coloneqq \frac{\lambda_{\max}[X^\top X]}{\lambda_{\min}[X^\top X]}$, and $\mu \coloneqq [\mathbb{E}[\inf_{t\geq 1}[\pi_{\theta_t}(a^*)]^{-2}]]^{-1} > 0$.

Proof. For the following proof, define:

$$\begin{aligned}
f(\theta) &\coloneqq \langle \pi_\theta, r \rangle \\
\nabla f(\theta) &\coloneqq X^\top (\operatorname{diag}(\pi_\theta) - \pi_\theta\,\pi_\theta^\top)\,r. \\
\widetilde{\nabla} f(\theta) &\coloneqq X^\top (\operatorname{diag}(\pi_\theta) - \pi_\theta\,\pi_\theta^\top)\,\hat{r}.
\end{aligned}$$

Additionally, for $z \in \{X\theta \mid \theta \in \mathbb{R}^d\} \subseteq \mathbb{R}^K$ define,

$$\begin{aligned}
\bar{\pi}_z &\coloneqq \operatorname{softmax}(z) \\
J(z) &\coloneqq \langle \bar{\pi}_z, r \rangle \\
\nabla J(z) &\coloneqq (\operatorname{diag}(\bar{\pi}_z) - \bar{\pi}_z\,\bar{\pi}_z^\top)\,r \\
\nabla \tilde{J}(z) &\coloneqq (\operatorname{diag}(\bar{\pi}_z) - \bar{\pi}_z\,\bar{\pi}_z^\top)\,\hat{r}.
\end{aligned}$$

Since $z = X\theta$, we have $f(\theta) = J(z)$.
According to Lemma 46, f is $3\,\lambda_{\max}[X^\top X]\,\|\nabla J(z)\|$ non-uniform smooth and by Lemma 35, the stochastic gradients are bounded by $\sqrt{2\,\lambda_{\max}[X^\top X]\,R_{\max}}$. Let

$$L_1 \coloneqq 3\,\lambda_{\max}[X^\top X] \quad B \coloneqq \sqrt{2\,\lambda_{\max}[X^\top X]\,R_{\max}}$$

Using Algorithm 4 with $\eta_t \in \left(0, \frac{1}{L_1\,B}\right)$, Lemma 34 implies,

$$\begin{aligned}
|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle| &\leq \frac{1}{2}\,\frac{L_1\,\|\nabla J(z_t)\|}{1 - L_1\,B\,\eta_t}\,\|\theta_{t+1} - \theta_t\|_2^2 \\
&\leq \frac{2\,L_1\,\|\nabla J(z_t)\|}{2}\,\|\theta_{t+1} - \theta_t\|_2^2 \\
&\qquad \text{(Since } \eta_t \leq \tfrac{1}{6\,(\lambda_{\max}[X^\top X])^{3/2}\,\sqrt{2\,R_{\max}}} = \tfrac{1}{2\,L_1\,B},\ 1 - L_1\,B\,\eta_t \geq \tfrac{1}{2}\text{)} \\
\implies f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle &\geq -\frac{2\,L_1\,\|\nabla J(z_t)\|}{2}\,\|\theta_{t+1} - \theta_t\|_2^2
\end{aligned}$$

$$\begin{aligned}
f(\theta_{t+1}) - f(\theta_t) - \eta_t\left\langle \nabla f(\theta_t), \widetilde{\nabla} f(\theta_t)\right\rangle &\geq -\frac{2\,\eta_t{}^2\,L_1\,\|\nabla J(z_t)\|}{2}\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2 \\
&\qquad \text{(By the update in Algorithm 4, } \theta_{t+1} = \theta_t + \eta_t\,\widetilde{\nabla} f(\theta_t)\text{)} \\
\implies f(\theta_{t+1}) \geq f(\theta_t) + \eta_t\left\langle \nabla f(\theta_t), \widetilde{\nabla} f(\theta_t)\right\rangle - \frac{2\,\eta_t{}^2\,L_1\|\nabla J(z_t)\|}{2}\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2 & \\
J(z_{t+1}) \geq J(z_t) + \eta_t\left\langle \nabla f(\theta_t), \widetilde{\nabla} f(\theta_t)\right\rangle - \frac{2\,\eta_t{}^2\,L_1\,\|\nabla J(z_t)\|}{2}\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2 & . \\
&\qquad (f(\theta) = J(z) \text{ for } z = X\theta)
\end{aligned}$$

$$J^* - J(z_{t+1}) \leq J^* - J(z_t) - \eta_t\left\langle \nabla f(\theta_t), \widetilde{\nabla} f(\theta_t)\right\rangle + \frac{2\,\eta_t{}^2\,L_1\,\|\nabla J(z_t)\|}{2}\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2$$

$$\begin{aligned}
&\text{(Multiplying both sides by } -1 \text{ and adding } J^* \coloneqq \sup_{\theta\in\mathbb{R}^d}\langle \pi_\theta, r\rangle\text{)} \\
\mathbb{E}_t[J(z^*) - J(z_{t+1})] \leq \mathbb{E}_t[J(z^*) - J(z_{t+1})] - \eta_t\left\langle \nabla f(\theta_t), \mathbb{E}_t[\widetilde{\nabla} f(\theta_t)]\right\rangle + \frac{2\,\eta_t{}^2\,L_1\,\|\nabla J(z_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right] & \\
&\text{(Taking expectation with respect to the randomness in iteration } t \text{ on both sides)} \\
\mathbb{E}_t[J(z^*) - J(z_{t+1})] = \mathbb{E}_t[J(z^*) - J(z_{t+1})] - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{2\,\eta_t{}^2\,L_1\,\|\nabla J(z_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right] & \\
&\text{(By Lemma 44 the stochastic gradients are unbiased)} \\
\implies \delta(z_{t+1}) \leq \delta(z_t) - \eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{2\,\eta_t{}^2\,L_1\,\|\nabla J(z_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right] & \\
&\text{(Let } \delta(z) \coloneqq \mathbb{E}_t[J(z^*) - J(z)]\text{)}
\end{aligned}$$

Simplifying the third term on the RHS,

$$\mathbb{E}_t\left[\left\|\nabla \tilde{f}(\theta_t)\right\|_2^2\right] = \mathbb{E}_t\left[\left\|X^\top \nabla \tilde{J}(z_t)\right\|_2^2\right] = \mathbb{E}_t\left[\nabla \tilde{J}(z_t)^\top X X^\top \nabla \tilde{J}(z_t)\right] \leq \lambda_{\max}[X^\top X]\, \mathbb{E}_t\left[\left\|\nabla \tilde{J}(z_t)\right\|_2^2\right]$$

$$\implies \delta(z_{t+1}) \leq \delta(z_t) - \eta_t \left\|\nabla f(\theta_t)\right\|_2^2 + \frac{2\,{\eta_t}^2\, \lambda_{\max}[X^\top X]\, L_1 \left\|\nabla J(z_t)\right\|}{2}\, \mathbb{E}_t\left[\left\|\nabla \tilde{J}(z_t)\right\|_2^2\right]$$

By Lemma 36, J satisfies the strong growth condition, $\mathbb{E}_t\left\|\nabla \tilde{J}(z)\right\|_2^2 \leq \rho \left\|\nabla J(z)\right\|$ with $\rho \coloneqq \frac{8\, R_{\max}^3\, K^{3/2}}{\Delta^2}$ where $\Delta \coloneqq \min_{i \neq j} |r(i) - r(j)|$,

$$\leq \delta(z_t) - \eta_t \left\|\nabla f(\theta_t)\right\|_2^2 + \frac{2\,{\eta_t}^2\, \rho\, \lambda_{\max}[X^\top X]\, L_1}{2} \left\|\nabla J(z_t)\right\|_2^2$$

Simplifying the second term on the RHS in the above equation,

$$\left\|\nabla f(\theta_t)\right\|_2^2 = \left\|X^\top \nabla J(z_t)\right\|_2^2 = \nabla J(z_t)^\top X X^\top \nabla J(z_t) \geq \lambda_{\min}[X^\top X] \left\|\nabla J(z_t)\right\|_2^2$$
$$(X^\top \nabla J(z_t) \neq 0, \text{ since } \nabla J(z_t) \text{ does not lie in the null space of } X^\top)$$

$$\implies \delta(z_{t+1}) \leq \delta(z_t) - \eta_t\, \lambda_{\min}[X^\top X] \left\|\nabla J(z_t)\right\|_2^2 + \frac{2\,{\eta_t}^2\, \rho\, \lambda_{\max}[X^\top X]\, L_1}{2} \left\|\nabla J(z_t)\right\|_2^2$$

$$= \delta(z_t) - \eta_t\, \lambda_{\min}[X^\top X] \left\|\nabla J(z_t)\right\|_2^2 + \frac{6\,{\eta_t}^2\, \rho\, [\lambda_{\max}[X^\top X]]^2}{2} \left\|\nabla J(z_t)\right\|_2^2$$
$$(\text{Since } L_1 \coloneqq 3\, \lambda_{\max}[X^\top X])$$

Since $\eta_t \leq \frac{\lambda_{\min}[X^\top X]}{6\, \rho\, [\lambda_{\max}[X^\top X]]^2}$ and let $\kappa \coloneqq \frac{\lambda_{\max}[X^\top X]}{\lambda_{\min}[X^\top X]}$

$$\implies \delta(z_{t+1}) \leq \delta(z_t) - \frac{1}{6\, \rho\, \kappa^2} \left\|\nabla J(z_t)\right\|_2^2 \tag{B.23}$$

By Lemma 37, J satisfies the non-uniform Łojasiewciz condition with $\xi = 0$ and $C(z) = \bar{\pi}_z(a^*)$,

$$\leq \delta(z_t) - \frac{1}{6\, \rho\, \kappa^2} \left[\bar{\pi}_{z_t}(a^*)\right]^2 \left[\delta(z_t)\right]^2$$

$$\leq \delta(z_t) - \frac{1}{6\, \rho\, \kappa^2}\, m\, \left[\delta(z_t)\right]^2 \quad (\text{By Theorem 9, } m \coloneqq \inf_{t \geq 1} [\bar{\pi}_{z_t}(a^*)]^2 > 0)$$

Taking expectation with respect to all previous iterations $t \geq 1$ on both sides,

$$\implies \mathbb{E}[\delta(z_{t+1})] \leq \mathbb{E}[\delta(z_t)] - \frac{1}{6\, \rho\, \kappa^2}\, \mathbb{E}[m\, [\delta(z_t)]^2]$$

To lower bound $\mathbb{E}[m\,[\delta(z_t)]^2]$,

$$\begin{aligned}
\mathbb{E}[\delta(z_t)] &= \mathbb{E}\left[\frac{1}{\sqrt{m}}\sqrt{m}\delta(z_t)\right] \\
&\leq \sqrt{\mathbb{E}\left[\frac{1}{m}\right]}\sqrt{\mathbb{E}[m\,[\delta(z_t)]^2]} \\
&\qquad\qquad \text{(Using Cauchy-Schwarz since } m > 0 \text{ and } \delta(z_t) > 0) \\
\implies \underbrace{\left[\frac{1}{m}\right]^{-1}}_{:=\mu} &(\mathbb{E}[\delta(z_t)])^2 \leq \mathbb{E}[m\,[\delta(z_t)]^2]
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}[\delta(z_{t+1})] &\leq \mathbb{E}[\delta(z_t)] - \frac{\mu}{6\,\rho\,\kappa^2}\,(\mathbb{E}[\delta(z_t)])^2 \\
&= \mathbb{E}[\delta(z_t)] - \frac{1}{\alpha}\,(\mathbb{E}[\delta(z_t)])^2 \qquad (\tfrac{1}{\alpha} \coloneqq \tfrac{\mu}{6\,\rho\,\kappa^2})
\end{aligned}$$

Dividing each side by $\mathbb{E}[\delta(z_t)]\,\mathbb{E}[\delta(z_{t+1})]$,

$$\frac{1}{\mathbb{E}[\delta(z_t)]} \leq \frac{1}{\mathbb{E}[\delta(z_{t+1})]} - \frac{1}{\alpha}\,\frac{\mathbb{E}[\delta(z_t)]}{\mathbb{E}[\delta(z_{t+1})]}.$$

Using the above inequality and recursing from iteration $t = 1$ to T,

$$\begin{aligned}
\frac{1}{\mathbb{E}[\delta(z_1)]} &\leq \frac{1}{\mathbb{E}[\delta(z_{T+1})]} - \frac{1}{\alpha}\sum_{t=1}^{T}\frac{\mathbb{E}[\delta(z_t)]}{\mathbb{E}[\delta(z_{t+1})]} \\
&\leq \frac{1}{\mathbb{E}[\delta(z_{T+1})]} - \frac{T}{\alpha} \qquad (\mathbb{E}[\delta(z_t)] \geq \mathbb{E}[\delta(z_{t+1})]) \\
\implies \frac{T}{\alpha} &\leq \frac{1}{\mathbb{E}[\delta(z_{T+1})]}.
\end{aligned}$$

Therefore,

$$\mathbb{E}[J(z^*) - J(z_{T+1})] \leq \frac{6\,\rho\,\kappa^2}{\mu\,T}.$$

□

B.4.3 Additional Lemmas

Lemma 33. *Using Algorithm 4, if there exist a $\tau \geq 1$ such that $\langle \pi_{\theta_\tau}, r\rangle \geq r(a)$, we have, almost surely,*

$$\lim_{t\to\infty}\langle \pi_{\theta_t}, r\rangle > r(a).$$

Proof. According to Equation (B.24), we have, for all finite $t \geq 1$, $\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r\rangle] > \langle \pi_{\theta_t}, r\rangle$, where $\mathbb{E}_t$ takes expectation w.r.t. the randomness in iteration t. Therefore, we have, for all

finite $t > \tau$,

$$\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] > \langle \pi_{\theta_\tau}, r \rangle > r(a).$$

According to Equation (B.24), we also have

$$\begin{aligned} &\lim_{t\to\infty} (\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] - \langle \pi_{\theta_t}, r \rangle) = 0 \\ \implies &\lim_{t\to\infty} \langle \pi_{\theta_t}, r \rangle = \lim_{t\to\infty} \mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] > \langle \pi_{\theta_\tau}, r \rangle \geq r(a). \end{aligned}$$

□

Lemma 2. *We set the constant learning rate as:*

$$\eta = \min\left\{ \frac{1}{6\,(\lambda_{\max}[X^\top X])^{3/2}\,\sqrt{2\,R_{\max}}}, \frac{\lambda_{\min}[X^\top X]}{6\,\rho\,[\lambda_{\max}[X^\top X]]^2} \right\}, \tag{4.9}$$

where $\rho := \frac{8\,R_{\max}^3 K^{3/2}}{\Delta^2}$, $\Delta := \min_{i\neq j}|r(i) - r(j)|$, $\kappa := \frac{\lambda_{\max}[X^\top X]}{\lambda_{\min}[X^\top X]}$ *is the condition number of* $X^\top X$, *and* $\mu := [\mathbb{E}[\inf_{t\geq 1}[\pi_\theta(a^*)]^{-2}]]^{-1} > 0$. *Algorithm 4 with the above learning rate assures that, for all* $t \geq 1$,

$$\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] - \langle \pi_{\theta_t}, r \rangle \geq \frac{1}{6\,\rho\,\kappa^2}\,\|J(z_t)\|_2^2.$$

Proof. Following the initial steps in Theorem 9, we have

$$J(z_t) - J(z_{t+1}) \leq -\eta_t \left\langle \nabla f(\theta_t), \widetilde{\nabla} f(\theta_t) \right\rangle + \frac{{\eta_t}^2\,6\,\lambda_{\max}[X^\top X]\,\|\nabla J(\theta_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right]$$

Taking expectation with respect to $a_t \sim \pi_{\theta_t}(\cdot)$ and $R_t(a_t) \sim P_{a_t}$,

$$= -\eta_t \left\langle \nabla f(\theta_t), \mathbb{E}_t[\widetilde{\nabla} f(\theta_t)] \right\rangle + \frac{{\eta_t}^2\,6\,\lambda_{\max}[X^\top X]\,\|\nabla J(\theta_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right]$$

By Lemma 44, the gradient is unbiased,

$$= -\eta_t\,\|\nabla f(\theta_t)\|_2^2 + \frac{{\eta_t}^2\,6\,\lambda_{\max}[X^\top X]\,\|\nabla J(\theta_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right]$$

Simplifying the second term on the RHS in the above equation,

$$\begin{aligned} \|\nabla f(\theta_t)\|_2^2 = \left\|X^\top \nabla J(z_t)\right\|_2^2 &= \nabla J(z_t)^\top X X^\top \nabla J(z_t) \geq \lambda_{\min}[X^\top X]\,\|\nabla J(z_t)\|_2^2 \\ &\qquad (X^\top \nabla J(z_t) \neq 0, \text{ since } \nabla J(z_t) \text{ does not lie in the null space of } X^\top) \\ &\leq -\eta_t\,\lambda_{\min}[X^\top X]\|\nabla J(\theta_t)\|_2^2 + \frac{{\eta_t}^2\,6\,\lambda_{\max}[X^\top X]\,\|\nabla J(\theta_t)\|}{2}\,\mathbb{E}_t\left[\left\|\widetilde{\nabla} f(\theta_t)\right\|_2^2\right] \end{aligned}$$

Simplifying the third term on the RHS,

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla \tilde{f}(\theta_t)\right\|_2^2\right] &= \mathbb{E}\left[\left\|X^\top \nabla \tilde{J}(z_t)\right\|_2^2\right] = \mathbb{E}\left[\nabla \tilde{J}(z_t)^\top X X^\top \nabla \tilde{J}(z_t)\right] \leq \lambda_{\max}[X^\top X]\, \mathbb{E}\left[\left\|\nabla \tilde{J}(z_t)\right\|_2^2\right] \\ &\leq -\eta_t\, \lambda_{\min}[X^\top X] \|\nabla J(\theta_t)\|_2^2 + \frac{{\eta_t}^2\, 6\, [\lambda_{\max}[X^\top X]]^2\, \|\nabla J(\theta_t)\|}{2}\, \mathbb{E}_t\left[\left\|\nabla \tilde{J}(\theta_t)\right\|_2^2\right]\end{aligned}$$

By Lemma 36, J satisfies the strong growth condition, $\mathbb{E}\left[\left\|\nabla \tilde{J}(z)\right\|_2^2\right] \leq \rho\, \|\nabla J(z)\|$ with $\rho := \frac{8\, R_{\max}^3\, K^{3/2}}{\Delta^2}$ where $\Delta := \min_{i \neq j} |r(i) - r(j)|$,

$$\begin{aligned}&\leq -\eta_t\, \lambda_{\min}[X^\top X] \|\nabla J(\theta_t)\|_2^2 + \frac{{\eta_t}^2\, \rho\, 6\, \lambda_{\max}[X^\top X]^2\, \|\nabla J(\theta_t)\|_2^2}{2} \\ &\leq \frac{1}{6\, \rho\, \kappa^2}\, \|J(z_t)\|_2^2 \qquad \left(\eta_t \leq \tfrac{\lambda_{\min}[X^\top X]}{6\, \rho\, [\lambda_{\max}[X^\top X]]^2}\right)\end{aligned}$$

□

Lemma 3. *Using Algorithm 4 with a constant step-size as in Equation* (4.9) *will converge to a one-hot policy (i.e. there exists an (possibly random) arm $k \in [K]$ such that $\pi_{\theta_t}(k) \to 1$ as $t \to \infty$) almost surely.*

Proof. According to Lemma 2, we have $\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] \geq \langle \pi_{\theta_t}, r \rangle$ for all $t \geq 1$. The sequence $\{\langle \pi_{\theta_t}, r \rangle\}_{t \geq 1}$ satisfies the condition of sub-martingale. Therefore, according to Corollary 3 in Mei et al. (2022a), we have almost surely,

$$\begin{aligned}&\lim_{t \to \infty} \mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] - \langle \pi_{\theta_{t+1}}, r \rangle = 0 \\ \implies &\lim_{t \to \infty} \mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] - \langle \pi_{\theta_t}, r \rangle = 0\end{aligned}$$

According to Lemma 2, we have

$$\begin{aligned}\mathbb{E}_t[\langle \pi_{\theta_{t+1}}, r \rangle] - \langle \pi_{\theta_t}, r \rangle &\geq \frac{\Delta^2\, [\lambda_{\min}[X^\top X]]^2}{48\, R_{\max}^3\, K^{3/2}\, [\lambda_{\max}[X^\top X]]^2}\, \|J(z_t)\|_2^2 \\ &\geq \frac{\Delta^2\, [\lambda_{\min}[X^\top X]]^2}{48\, R_{\max}^3\, K^{3/2}\, [\lambda_{\max}[X^\top X]]^2} \sum_{i=1}^K \pi_{\theta_t}(i)^2\, (r(i) - \langle \pi_{\theta_t}, r \rangle)^2 > 0.\end{aligned} \tag{B.24}$$

Therefore, we have

$$\lim_{t \to \infty} \sum_{i=1}^K \pi_{\theta_t}(i)^2\, (r(i) - \langle \pi_{\theta_t}, r \rangle)^2 = 0,$$

which implies for all arms $i \in [K]$,

$$\lim_{t \to \infty} \pi_{\theta_t}(i)^2\, (r(i) - \langle \pi_{\theta_t}, r \rangle)^2 = 0.$$

We can assume that $\lim_{t\to\infty}(r(i) - \langle \pi_{\theta_t}, r\rangle)^2 > 0$ for all arms $i \in [K]$. Combined with above equation, we have $\lim_{t\to\infty} \pi_{\theta_t}(i) = 0$ for all arms $i \in [K]$, which contradicts with the fact that $\sum_{i=1}^{K} \pi_{\theta_t}(i) = 1$. Therefore, there exist at least one arm $k \in [K]$ such that

$$\begin{aligned}
&\lim_{t\to\infty}(r(k) - \langle \pi_{\theta_t}, r\rangle)^2 = 0 \\
\implies &\lim_{t\to\infty} \langle \pi_{\theta_t}, r\rangle = r(k) \\
\implies &\lim_{t\to\infty} \pi_{\theta_t}(k) = 1, \qquad \text{(By Assumption 7)}
\end{aligned}$$

which completes the proof. □

Lemma 34 (Lemma 5 in Lu et al. (2024))**.** *Assuming that f is L_1-non-uniform smooth and the stochastic gradient is bounded, i.e. $\|\widetilde{\nabla} f(\theta_t)\| \le B$, using Algorithm 4 with $\eta_t \in (0, \frac{1}{L_1 B})$ we have,*

$$|f(\theta_{t+1}) - f(\theta_t) - \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle| \le \frac{1}{2}\,\frac{L_1\,\|\nabla f(\theta_t)\|}{1 - L_1\,B\,\eta_t}\,\|\theta_{t+1} - \theta_t\|_2^2. \tag{B.25}$$

Lemma 35.

$$\|\frac{d\langle \pi_\theta, \hat{r}\rangle}{d\theta}\| \le \sqrt{2\,\lambda_{\max}[X^\top X]\,R_{\max}} \tag{B.26}$$

Proof.

$$\begin{aligned}
\left\|\frac{d\langle \pi_{\theta_t}, \hat{r}\rangle}{d\theta_t}\right\|_2^2 &= \left\|X^\top \nabla \tilde{J}(z_t)\right\|_2^2 \\
&= \left\|\tilde{J}(z_t)^\top X\, X^\top \nabla \tilde{J}(z_t)\right\|_2^2 \\
&\le \lambda_{\max}[X^\top X]\left\|\nabla \tilde{J}(z)\right\|_2^2 \\
&= \lambda_{\max}[X^\top X] \sum_{a\in[K]} \left(\frac{d\langle \bar{\pi}_z, \hat{r}\rangle}{dz(a)}\right) \\
&= \lambda_{\max}[X^\top X] \sum_{a\in[K]} \left(\mathbb{1}\{a' = a\} - \bar{\pi}_z(a)\right)^2 (R(a))^2 \\
&\le \lambda_{\max}[X^\top X]\, R_{\max}^2 \sum_{a\in[K]} \left(\mathbb{1}\{a' = a\} - \bar{\pi}_z(a)\right)^2 \\
&\le \lambda_{\max}[X^\top X]\, R_{\max}^2 \left[(1 - \bar{\pi}_z(a))^2 + \sum_{a\neq a'} \bar{\pi}_z(a)^2\right] \\
&\le \lambda_{\max}[X^\top X]\, R_{\max}^2 \left[(1 - \bar{\pi}_z(a))^2 + \left(\sum_{a\neq a'} \bar{\pi}_z(a)\right)^2\right] \qquad (\|\cdot\|_2 \le \|\cdot\|_1) \\
&= 2\,\lambda_{\max}[X^\top X]\,, R_{\max}^2 (1 - \bar{\pi}_z(a))^2 \\
&\le 2\,\lambda_{\max}[X^\top X]\, R_{\max}^2
\end{aligned}$$

□

Lemma 36 (Lemma 4.3 in Mei et al. (2023)). *Using Algorithm 4, we have*

$$\mathbb{E}\left[\left\|\frac{d\langle\bar{\pi}_z,\hat{r}\rangle}{dz}\right\|_2^2\right] \leq \frac{8\,R_{\max}^3\,K^{3/2}}{\Delta^2}\,\|\frac{\langle\bar{\pi}_z,r\rangle}{dz}\|$$

where $\Delta := \min_{i\neq j}|r(i)-r(j)|$ *and* $\bar{\pi}_z := \mathrm{softmax}(z)$ *for* $z\in\mathbb{R}^K$.

Lemma 37 (Lemma 3 of Mei et al. (2020b)). *Let Assumption 7 hold and let* $\pi^* := \arg\max_{\pi\in\Delta_K}\langle\pi,r\rangle$. *Then*

$$\|\frac{d\langle\bar{\pi}_z,r\rangle}{dz}\| \geq \bar{\pi}_z(a^*)\,\langle\pi^*-\bar{\pi}_z,r\rangle$$

where $\bar{\pi}_z := \mathrm{softmax}(z)$ *for* $z\in\mathbb{R}^K$.

Lemma 38. *Using Algorithm 4 with any constant* $\eta\in\Omega(1)$, *we have, for all large enough* $t\geq 1$,*almost surely,*

$$r(i_2) > \langle\pi_{\theta_t},r\rangle > r(i_1)$$

, *where* $i_1 := \arg\min_{a\in\mathcal{A}_\infty} r(a)$ *and* $i_{|\mathcal{A}_\infty|} := \arg\max_{a\in\mathcal{A}_\infty} r(a)$.

Proof. **Part I**: $\langle\pi_{\theta_t},r\rangle > r(i_1)$.

According to Lemma 40, we have at least another arm $i_{|\mathcal{A}_\infty|}$ such that $r(i_{|\mathcal{A}_\infty|}) > r(i_1)$ and $N_\infty(i_{|\mathcal{A}_\infty|}) = \infty$. Define that

$$\mathcal{A}^+(i_1) := \left\{a^+\in[K] : r(a^+) > r(i_1)\right\},\quad \mathcal{A}^-(i_1) := \left\{a^-\in[K] : r(a^-) < r(i_1)\right\}.$$

Then, we have, for all large enough t,

$$\begin{aligned}
\langle\pi_{\theta_t},r\rangle - r(i_1) &= \sum_{a\in\mathcal{A}^+(i_1)}\pi_{\theta_t}(a)\,(r(a)-r(i_1)) - \sum_{a\in\mathcal{A}^-(i_1)}\pi_{\theta_t}(a)\,(r(i_1)-r(a)) \\
&> \pi_{\theta_t}(i_{|\mathcal{A}_\infty|})\,(r(i_{|\mathcal{A}_\infty|})-r(i_1)) - \sum_{a\in\mathcal{A}^-(i_1)}\pi_{\theta_t}(a)\,(r(i_1)-r(a)) \\
&= \pi_{\theta_t}(i_{|\mathcal{A}_\infty|})\left[\underbrace{r(i_{|\mathcal{A}_\infty|})-r(i_1)}_{>0} - \sum_{a\in\mathcal{A}^-(i_1)}\frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(i_{|\mathcal{A}_\infty|})}\underbrace{(r(i_1)-r(a))}_{>0}\right]
\end{aligned}$$

Since $N_\infty(a) < \infty$ for all $a\in\mathcal{A}^-(i_1)$, according to Lemma 39, we have $\sup_{t\geq 1}\frac{\pi_{\theta_t}(i_{|\mathcal{A}_\infty|})}{\pi_{\theta_t}(a)} = \infty$. Therefore, for all large enough t, $\langle\pi_{\theta_t},r\rangle > r(i_1)$.

Part II: $r(i_{|\mathcal{A}_\infty|}) > \langle \pi_{\theta_t}, r \rangle$. Similarly, we have

$$\begin{aligned} r(i_{|\mathcal{A}_\infty|}) - \langle \pi_{\theta_t}, r \rangle &= \sum_{a \in \mathcal{A}^-(i_{|\mathcal{A}_\infty|})} \pi_{\theta_t}(a) \left(r(i_{|\mathcal{A}_\infty|}) - r(a) \right) - \sum_{a \in \mathcal{A}^+(i_{|\mathcal{A}_\infty|})} \pi_{\theta_t}(a) \left(r(a) - r(i_{|\mathcal{A}_\infty|}) \right) \\ &> \pi_{\theta_t}(i_1) \left(r(i_{|\mathcal{A}_\infty|}) - r(i_1) \right) - \sum_{a \in \mathcal{A}^+(i_{|\mathcal{A}_\infty|})} \pi_{\theta_t}(a) \left(r(a) - r(i_{|\mathcal{A}_\infty|}) \right) \\ &= \pi_{\theta_t}(i_1) \left[\underbrace{r(i_{|\mathcal{A}_\infty|}) - r(i_1)}_{>0} - \sum_{a \in \mathcal{A}^+(i_1)} \frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(i_1)} \underbrace{\left(r(a) - r(i_{|\mathcal{A}_\infty|}) \right)}_{>0} \right] \end{aligned}$$

Since $N_\infty(a) < \infty$ for all $a \in \mathcal{A}^+(i_{|\mathcal{A}_\infty|})$, according to Lemma 39, we have $\sup_{t \geq 1} \frac{\pi_{\theta_t}(i_1)}{\pi_{\theta_t}(a)} = \infty$. Therefore, for all large enough t, $r(i_{|\mathcal{A}_\infty|}) > \langle \pi_{\theta_t}, r \rangle$. □

Lemma 39. *Using Algorithm 4, for any two different actions $i, j \in [K]$ with $i \neq j$, if $N_\infty(i) = \infty$ and $N_\infty(j) < \infty$, then we have, almost surely,*

$$\sup_{t \geq 1} \frac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(j)} = \infty.$$

Proof. We will prove this by contradiction. Assume that $\sup_{t \geq 1} \frac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(j)} = C < \infty$ for some $C > 0$. According to the extended Borel-Cantelli (Lemma 42), since $N_\infty(i) = \infty$, we have $\sum_{t=0}^{\infty} \pi_{\theta_t}(i) = \infty$. Similarly, since $N_\infty(j) < \infty$, we have $\sum_{t=0}^{\infty} \pi_{\theta_t}(j) < \infty$. Therefore,

$$\begin{aligned} \sum_{t=1}^{\infty} \pi_{\theta_t}(i) &= \sum_{t=1}^{\infty} \pi_{\theta_t}(j) \frac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(j)} \\ &< C \sum_{t=1}^{\infty} \pi_{\theta_t}(j) < \infty, \qquad (C = \sup_{t \geq 1} \tfrac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(j)}) \end{aligned}$$

which contradicts the fact that $\sum_{t=1}^{\infty} \pi_{\theta_t}(i) = \infty$. Therefore, we have $\sup_{t \geq 1} \frac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(j)} = \infty$. □

Lemma 40. *Using Algorithm 4 with any fixed learning rate $\eta > 0$, there exist at least a pair of two distinct actions i, $j \in [K]$ and $i \neq j$, such that, almost surely,*

$$N_\infty(i) = \infty, \text{ and } N_\infty(j) = \infty.$$

Proof. By pigeonhole principle, there exists at least one action $i \in [K]$, such that, almost surely,

$$N_\infty(i) := \lim_{t \to \infty} N_t(i) = \infty.$$

We argue the existence of another action by contradiction. Suppose for all the other actions $j \in [K]$ and $j \neq i$, we have $N_\infty(j) < \infty$. According to Lemma 42, for all $j \neq i$, we have,

almost surely,

$$\sum_{t=1}^{\infty} \pi_{\theta_t}(j) := \lim_{t\to\infty} \sum_{s=1}^{t} \pi_{\theta_s}(j) < \infty.$$

Recall we have the following update:

$$\theta_{t+1} = \theta_t + \eta\, X^\top H_t \hat{r}_t \implies z_{t+1} = z_t + \eta\, X X^\top H_t \hat{r}_t.$$

Then for any arm $\tilde{a} \in [K]$,

$$\begin{aligned}
z_{t+1}(\tilde{a}) &= z_t(\tilde{a}) + \eta \sum_{a=1}^{K} \langle x_{\tilde{a}}, x_a \rangle\, \pi_{\theta_t}(a)\, [\hat{r}(a) - \langle \pi_{\theta_t}, \hat{r} \rangle] \\
&= z_t(\tilde{a}) + \eta \left[\sum_{a=1}^{K} I_t(a) \left(\langle x_{\tilde{a}}, x_a \rangle\, (1 - \pi_{\theta_t}(a))\, R_t - \sum_{j \neq a} \langle x_{\tilde{a}}, x_j \rangle\, \pi_{\theta_t}(j)\, R_t \right) \right] \\
&= z_t(\tilde{a}) + \eta \left[I_t(i) \left(\langle x_{\tilde{a}}, x_i \rangle\, (1 - \pi_{\theta_t}(i))\, R_t - \sum_{j \neq i} \langle x_{\tilde{a}}, x_j \rangle\, \pi_{\theta_t}(j)\, R_t \right) \right. \\
&\qquad \left. + \sum_{\substack{a=1 \\ a \neq i}}^{K} I_t(a) \left(\langle x_{\tilde{a}}, x_a \rangle\, (1 - \pi_{\theta_t}(a))\, R_t - \sum_{j \neq a} \langle x_{\tilde{a}}, x_j \rangle\, \pi_{\theta_t}(j)\, R_t \right) \right] \qquad \text{(B.27)}
\end{aligned}$$

Using Equation (B.27), recursing from 1 to $t-1$, and using the triangle inequality, we have

$$\begin{aligned}
|z_t(\tilde{a}) - z_1(\tilde{a})| \leq \eta \sum_{s=1}^{t-1} & \left| I_t(i) \left(\langle x_{\tilde{a}}, x_i \rangle\, (1 - \pi_{\theta_t}(i))\, R_t - \sum_{j \neq i} \langle x_{\tilde{a}}, x_j \rangle\, \pi_{\theta_t}(j)\, R_t \right) \right| \\
& + \eta \sum_{s=1}^{t-1} \left| \sum_{\substack{a=1 \\ a \neq i}}^{K} I_t(a) \left(\langle x_{\tilde{a}}, x_a \rangle\, (1 - \pi_{\theta_t}(a))\, R_t - \sum_{j \neq a} \langle x_{\tilde{a}}, x_j \rangle\, \pi_{\theta_t}(j)\, R_t \right) \right|
\end{aligned}$$

Let $C := \max_{a,a'} |\langle x_a, x_{a'} \rangle|$. Since $|R_t| \leq R_{\max}$ and using triangle inequality,

$$\begin{aligned}
&\leq \eta\, R_{\max}\, C \sum_{s=1}^{t-1} \Bigg[I_s(i) \Bigg((1 - \pi_{\theta_s}(i)) + \sum_{j \neq i} \pi_{\theta_s}(j) \Bigg) \\
&\qquad\qquad + \sum_{\substack{a=1 \\ a \neq i}}^{K} I_s(a) \Bigg((1 - \pi_{\theta_s}(a)) + \sum_{j \neq a} \pi_{\theta_s}(j) \Bigg) \Bigg] \\
&= 2\, \eta\, R_{\max}\, C \sum_{s=1}^{t-1} \Bigg[I_s(i) \sum_{j \neq i} \pi_{\theta_s}(j) + \sum_{\substack{a=1 \\ a \neq i}}^{K} I_s(a) \sum_{j \neq a} \pi_{\theta_s}(a) \Bigg] \\
&\leq 2\eta\, R_{\max}\, C \sum_{s=1}^{t-1} \Bigg[\sum_{j \neq i} \pi_{\theta_s}(j) + (K-1) \sum_{\substack{a=1 \\ a \neq i}}^{K} I_s(a) \Bigg] \\
&= 2\, \eta\, R_{\max}\, C \Bigg[\sum_{j \neq i} \sum_{s=1}^{t-1} \pi_{\theta_s}(j) + (K-1) \sum_{\substack{a=1 \\ a \neq i}}^{K} \sum_{s=1}^{t-1} I_s(a) \Bigg] \\
&= 2\, \eta\, R_{\max}\, C \Bigg[\sum_{j \neq i} \sum_{s=1}^{t-1} \pi_{\theta_s}(j) + (K-1) \sum_{\substack{a=1 \\ a \neq i}}^{K} N_{t-1}(a) \Bigg]
\end{aligned}$$

From the assumption that $N_\infty(j) < \infty$, for any arm $\tilde{a} \in [K]$, almost surely,

$$\sup_{t \geq 1} |z_t(\tilde{a})| \leq \sup_{t \geq 1} |z_t(\tilde{a}) - z_1(\tilde{a})| + |z_1(\tilde{a})| < \infty.$$

Since for all arms $\tilde{a} \in [K]$, the logit is always finite, there exists a finite constant $c_{\tilde{a}} \geq 0$, such that,

$$\begin{aligned}
&\inf_{t \geq 1} \pi_{\theta_t}(\tilde{a}) = \inf_{t \geq 1} \frac{\exp(z_t(\tilde{a}))}{\sum_{a' \in [K]} \exp(z_t(a'))} \geq c_{\tilde{a}} > 0. \\
\implies &\sum_{t=1}^{\infty} \pi_{\theta_t}(\tilde{a}) = \lim_{t \to \infty} \sum_{s=1}^{t} \pi_{\theta_s}(a) \geq \lim_{t \to \infty} t\, c_{\tilde{a}} = \infty.
\end{aligned}$$

According to Lemma 42, we have, almost surely, for all $\tilde{a} \in [K]$,

$$N_\infty(\tilde{a}) = \infty$$

which is a contradiction with the assumption that $N_\infty(j) < \infty$ for all $j \neq i$. Therefore, there exists another action $j \neq i$ such that $N_\infty(j) = \infty$. □

B.5 Additional Lemmas

Lemma 41. *For an arbitrary action a', $\mathbb{E}_t[W_{t+1}(a')] = 0$, $|W_{t+1}(a')| \le 4\,\eta\,R_{\max}\,\|y_{a'}\|_1$ where $y_{a'} := Xx_{a'}$, and*

$$Var[W_{t+1}(a')] \le 2\eta^2\,R_{\max}^2 \sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\,\pi_{\theta_t}(j)\,(1-\pi_{\theta_t}(j))$$

Proof.

$$\begin{aligned} W_{t+1}(a') &= z_{t+1}(a') - \mathbb{E}_t[z_{t+1}(a')] = [X\theta_{t+1}](a') - \mathbb{E}_t[[X\theta_{t+1}](a')] \\ &= \langle x_{a'}, \eta X^\top H_t\,(\hat{r}_t - r)\rangle = \eta\,[Xx_{a'}]^\top H_t\,(\hat{r}_t - r) \\ &= \eta\,y_{a'}^\top H_t\,(\hat{r}_t - r) && (y_{a'} := Xx_{a'}) \end{aligned}$$

We consider a centered version of the rewards formed by subtracting $r(i)$ from all the rewards. Specifically, we consider bounding the term,

$$\eta\,y_{a'}^\top H_t\,[(\hat{r}_t - r) - (\hat{r}_t(i) - r(i))\mathbf{1}] = \eta\,y_{a'}^\top H_t\,(\hat{r}_t - r) = W_{t+1}(a') \qquad (\text{Since } H_t\mathbf{1} = 0)$$

For convenience, we will overload the notation and subsequently use $\hat{r}_t - r$ to refer to the centered rewards. This implies that $(\hat{r}_t - r)(i) = 0$. With this in mind, we will show that $\mathbb{E}[W_{t+1}(a')] = 0$, $W_{t+1}(a')$ is bounded and upper-bound $\mathrm{Var}[W_{t+1}(a')]$. Since $y_{a'}$ and H_t are independent of the randomness and the importance-weighted reward estimate is unbiased, we have

$$\mathbb{E}[W_{t+1}(a')] = \eta y_{a'}^\top H_t\,\mathbb{E}[\hat{r}_t - r] = 0.$$

Then, we have

$$\begin{aligned} |W_{t+1}(a')| &\le \eta\,\|y_{a'}\|_1\,\|H_t\,(\hat{r}_t - r)\|_\infty && (\text{Holder's inequality}) \\ &= \eta\,\|y_{a'}\|_1\,\max_a\{|I_t(a) - \pi_{\theta_t}(a)|\,R_t(a_t) - \pi_{\theta_t}(a)\,[r(a) - \langle\pi_{\theta_t}, r\rangle]\} \\ &\le \eta\,\|y_{a'}\|_1\,4R_{\max} \end{aligned}$$

Since all entries of X are bounded, $y_{a'}$ is bounded

$$\implies |W_{t+1}(a')| \le 4\,\eta\,R_{\max}\,\|y_{a'}\|_1 \text{ is bounded.}$$

Next, we will bound the variance of $W_{t+1}(a')$:

$$\begin{aligned}
\mathrm{Var}[W_{t+1}(a')] &= \eta^2 \mathbb{E}\left[[y_{a'}^\top H_t\,(\hat{r}_t - r)]^2\right] \\
&\leq \eta^2\,\mathbb{E}\left[[y_{a'}^\top H_t\,\hat{r}_t]^2\right] && \text{(Since } \mathbb{E}[\hat{r}_t] = r\text{)} \\
&= \eta^2\,\mathbb{E}[(y_{a'}^\top H_t\,\hat{r}_t)^\top\,(y_{a'}^\top H_t\,\hat{r}_t)] \\
&= \eta^2\,\mathbb{E}[\hat{r}_t^\top\,H_t\,y_{a'}y_{a'}^\top H_t\,\hat{r}_t] && (H_t \text{ is symmetric}) \\
&= \eta^2\,\mathbb{E}\left[\mathrm{Tr}[\hat{r}_t^\top\,H_t\,y_{a'}y_{a'}^\top H_t\,\hat{r}_t]\right] && \text{(Trace of a scalar is equal to the scalar)} \\
&= \eta^2\,\mathbb{E}\left[\mathrm{Tr}[[y_{a'}y_{a'}^\top]\,[H_t\hat{r}_t]\,[H_t\hat{r}_t]^\top]\right] && \text{(Cyclic property of trace)} \\
&= \eta^2\ \mathrm{Tr}\left[\underbrace{[y_{a'}y_{a'}^\top]}_{:=Y}\ \mathbb{E}\left[\underbrace{[H_t\hat{r}_t]\,[H_t\hat{r}_t]^\top}_{:=X}\right]\right] \\
&\quad \text{(Trace is a linear operator and } y_{a'} \text{ does not depend on the randomness)} \\
&= \eta^2\ \mathrm{Tr}[Y^\top\mathbb{E}[X]] && (Y \text{ is symmetric}) \\
&= \eta^2\ \sum_{\substack{j=1\\j\neq i}}^{K}\sum_{\substack{k=1\\k\neq i}}^{K} Y_{j,k}\,\mathbb{E}[X_{j,k}] \\
&\quad \text{(Definition of trace and since } \hat{r}_t(i) = 0 \text{ because of the centering)}
\end{aligned}$$

$$\implies \mathrm{Var}[W_{t+1}(a')] \leq \eta^2\ \sum_{\substack{j=1\\j\neq i}}^{K} Y_{j,j}^2\,\mathbb{E}[X_{j,j}^2] + \eta^2\ \sum_{\substack{j=1\\j\neq i}}^{K}\sum_{\substack{k=1\\k\neq i\\k\neq j}}^{K} Y_{j,k}\,\mathbb{E}[X_{j,k}] \tag{B.28}$$

We need to upper-bound each entry in $\mathbb{E}[X]$ and we do this next.

$$\begin{aligned}
\mathbb{E}[X_{j,j}^2] &= \mathbb{E}[(I_t(j) - \pi_{\theta_t}(j))^2\,R_t^2(a_t)] && \text{(Using the definition of } H_t\hat{r}_t\text{)} \\
&\leq \pi_{\theta_t}(j)\ \left[[1-\pi_{\theta_t}(j)]^2\,r^2(j)\right] + \sum_{b\neq j}\pi_{\theta_t}(b)\ \left[(\pi_{\theta_t}(j))^2\ r^2(b)\right] \\
&\leq R_{\max}^2\ \left[\pi_{\theta_t}(j)\,(1-\pi_{\theta_t}(j))^2 + (1-\pi_{\theta_t}(j))\,(\pi_{\theta_t}(j))^2\right] \\
\implies \mathbb{E}[X_{j,j}^2] &\leq 2\,R_{\max}^2\,\pi_{\theta_t}(j)\,(1-\pi_{\theta_t}(j))
\end{aligned}$$

For $j \neq k$,

$$\begin{aligned}
\mathbb{E}[X_{j,k}] &= \mathbb{E}[(I_t(j) - \pi_{\theta_t}(j))\,(I_t(k) - \pi_{\theta_t}(k))\,R_t^2(a_t)] \qquad \text{(Using the definition of } H_t\hat{r}_t\text{)} \\
&= \pi_{\theta_t}(j)\left[(1-\pi_{\theta_t}(j))\,(-\pi_{\theta_t}(k))\,r^2(j)\right] + \pi_{\theta_t}(k)\left[(1-\pi_{\theta_t}(k))\,(-\pi_{\theta_t}(j))\,r^2(k)\right] \\
&+ \sum_{\substack{b\neq j\\ b\neq k}} \pi_{\theta_t}(b)\left[(-\pi_{\theta_t}(k))\,(-\pi_{\theta_t}(j))\,r^2(b)\right] \\
&\leq \sum_{\substack{b\neq j\\ b\neq k}} \pi_{\theta_t}(b)\left[(-\pi_{\theta_t}(k))\,(-\pi_{\theta_t}(j))\,r^2(b)\right] \qquad \text{(First two terms are negative)} \\
&\leq R_{\max}^2\,(1-\pi_{\theta_t}(j)-\pi_{\theta_t}(k))\,\pi_{\theta_t}(j)\,\pi_{\theta_t}(k) \\
&\leq R_{\max}^2\,\pi_{\theta_t}(j)\,\pi_{\theta_t}(k) \qquad \text{(Bounding the negative terms by zero)}
\end{aligned}$$

Additionally,

$$\begin{aligned}
\mathbb{E}[X_{j,k}] &\geq \pi_{\theta_t}(j)\left[(1-\pi_{\theta_t}(j))\,(-\pi_{\theta_t}(k))\,r^2(j)\right] + \pi_{\theta_t}(k)\left[(1-\pi_{\theta_t}(k))\,(-\pi_{\theta_t}(j))\,r^2(k)\right] \\
&\geq -R_{\max}^2\left[\pi_{\theta_t}(j)\,(1-\pi_{\theta_t}(j))\,\pi_{\theta_t}(k) + \pi_{\theta_t}(k)\,(1-\pi_{\theta_t}(k))\,\pi_{\theta_t}(j)\right] \\
&\geq -2R_{\max}^2\pi_{\theta_t}(j)\,\pi_{\theta_t}(k) \qquad (1-\pi_{\theta_t}(a)) \leq 1) \\
\implies |\mathbb{E}[X_{j,k}]| &\leq 2R_{\max}^2\pi_{\theta_t}(j)\,\pi_{\theta_t}(k)
\end{aligned}$$

Combining the above relations with Equation (B.28),

$$\begin{aligned}
\mathrm{Var}[W_{t+1}(a')] &\leq \eta^2 \sum_{\substack{j=1\\ j\neq i}}^{K} Y_{j,j}^2\,\mathbb{E}[X_{j,j}^2] + \eta^2 \sum_{\substack{j=1\\ j\neq i}}^{K}\sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} Y_{j,k}\,\mathbb{E}[X_{j,k}] \\
&\leq \eta^2 \Big|\sum_{\substack{j=1\\ j\neq i}}^{K} Y_{j,j}^2\,\mathbb{E}[X_{j,j}^2] + \sum_{\substack{j=1\\ j\neq i}}^{K}\sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} Y_{j,k}\,\mathbb{E}[X_{j,k}]\Big| \\
&\leq \eta^2 \sum_{\substack{j=1\\ j\neq i}}^{K} Y_{j,j}^2\,\mathbb{E}[X_{j,j}^2] + \Big|\sum_{\substack{j=1\\ j\neq i}}^{K}\sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} Y_{j,k}\,\mathbb{E}[X_{j,k}]\Big| \qquad \text{(Using triangle inequality)} \\
&\leq \eta^2 \sum_{\substack{j=1\\ j\neq i}}^{K} Y_{j,j}^2\,\mathbb{E}[X_{j,j}^2] + \sum_{\substack{j=1\\ j\neq i}}^{K}\sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} |Y_{j,k}|\,|\mathbb{E}[X_{j,k}]| \\
&\leq \eta^2\,R_{\max}^2\left[\sum_{\substack{j=1\\ j\neq i}}^{K} Y_{j,j}^2\;\pi_{\theta_t}(j)\,(1-\pi_{\theta_t}(j)) + \sum_{\substack{j=1\\ j\neq i}}^{K}\sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} |Y_{j,k}|\,\pi_{\theta_t}(j)\,\pi_{\theta_t}(k)\right]
\end{aligned}$$

In order to simplify the second term, without loss of generality, assume that the terms are ordered such that $|y_{a'}(1)| \geq |y_{a'}(2)| \ldots \geq |y_{a'}(K)|$, and recall that $Y_{j,k} = y_{a'}(j)\, y_{a'}(k)$. Hence,

$$\begin{aligned}
\sum_{\substack{j=1\\ j\neq i}}^{K} \sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} |Y_{j,k}|\, \pi_{\theta_t}(j)\, \pi_{\theta_t}(k) &= \sum_{\substack{j=1\\ j\neq i}}^{K} \sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} |y_{a'}(j)|\, |y_{a'}(k)|\, \pi_{\theta_t}(j)\, \pi_{\theta_t}(k) \\
&= 2 \sum_{\substack{j=1\\ j\neq i}}^{K-1} |y_{a'}(j)|\, \pi_{\theta_t}(j) \sum_{\substack{k=j+1\\ k\neq i}}^{K} |y_{a'}(k)|\, \pi_{\theta_t}(k) \\
&\leq 2 \sum_{\substack{j=1\\ j\neq i}}^{K-1} y_{a'}^2(j)\, \pi_{\theta_t}(j) \sum_{\substack{k=j+1\\ k\neq i}}^{K} \pi_{\theta_t}(k) \\
&\qquad\qquad \text{(Since } |y_{a'}(k)| \leq |y_{a'}(j)| \text{ for } k > j\text{)} \\
&= \sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\, \pi_{\theta_t}(j) \sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} \pi_{\theta_t}(k) \leq \sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\, \pi_{\theta_t}(j) \sum_{\substack{k=1\\ k\neq j}}^{K} \pi_{\theta_t}(k) \\
\implies \sum_{\substack{j=1\\ j\neq i}}^{K} \sum_{\substack{k=1\\ k\neq i\\ k\neq j}}^{K} |Y_{j,k}|\, \pi_{\theta_t}(j)\, \pi_{\theta_t}(k) &\leq \sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\, \pi_{\theta_t}(j)\, (1 - \pi_{\theta_t}(j))
\end{aligned}$$

Putting everything together,

$$\begin{aligned}
\mathrm{Var}[W_{t+1}(a')] &\leq \eta^2\, R_{\max}^2 \left[\sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\, \pi_{\theta_t}(j)\, (1 - \pi_{\theta_t}(j)) + \sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\, \pi_{\theta_t}(j)\, (1 - \pi_{\theta_t}(j)) \right] \\
&\leq 2\eta^2\, R_{\max}^2 \sum_{\substack{j=1\\ j\neq i}}^{K} y_{a'}^2(j)\, \pi_{\theta_t}(j)\, (1 - \pi_{\theta_t}(j))
\end{aligned}$$

□

Corollary 5. Define $y_{a,a'} := (X - \mathbf{1}v^\top)(x_a - x_{a'})$ where $v \in \mathbb{R}^d$. For an arbitrary action a and a', $\mathbb{E}[W_{t+1}(a')] = 0$, $|W_{t+1}(a) - W_{t+1}(a')| \leq 4\,\eta\, R_{\max}\, \|y_{a,a'}\|_1$, and

$$\mathrm{Var}[|W_{t+1}(a) - W_{t+1}(a')|] \leq 2\eta^2\, R_{\max}^2 \sum_{\substack{j=1\\ j\neq i}}^{K} (y_{a,a'}(j))^2\, \pi_{\theta_t}(j)\, (1 - \pi_{\theta_t}(j)).$$

Proof. Define that $\widetilde{W}_{t+1}(a,a') := |W_{t+1}(a) - W_{t+1}(a')|$.

$$\begin{aligned}
\widetilde{W}_{s+1}(a,a') &= |z_{t+1}(a) + z_{t+1}(a') - \mathbb{E}[z_{t+1}(a)] - \mathbb{E}[z_{t+1}(a')]| \\
&= [X\theta_{t+1}](a) + [X\theta_{t+1}](a') - \mathbb{E}[[X\theta_{t+1}](a)] - \mathbb{E}[[X\theta_{t+1}](a')] \\
&\qquad\qquad (z_t(a) := [X\theta_t](a)) \\
&= \langle x_a - x_{a'}, \eta X^\top H_t\,(\hat{r}_t - r)\rangle \\
&= \langle x_a - x_{a'}, \eta (X - \mathbf{1}v^\top)^\top H_t\,(\hat{r}_t - r)\rangle \qquad (\text{Since } v\mathbf{1}^\top H_t = 0) \\
&= \eta\,[(X - \mathbf{1}v^\top)(x_a - x_{a'})]^\top H_t\,(\hat{r}_t - r) \\
&= \eta\, y_{a,a'}^\top\, H_t\,(\hat{r}_t - r) \qquad (y_{a,a'} := (X - \mathbf{1}x_i^\top)(x_a - x_{a'}))
\end{aligned}$$

The proof follows from Lemma 41, with $W_{t+1}(a') = \widetilde{W}_{t+1}(a,a')$. □

For completeness, we append external lemmas here.

Lemma 42 (Extended Borel-Cantelli)**.** *Let $(\mathcal{F}_n)_{n\geq 1}$ be a filtration, $A_n \in \mathcal{F}_n$. Then, almost surely,*

$$\{\omega \,:\, \omega \in A_n \text{ infinitely often }\} = \left\{\omega \,:\, \sum_{n=1}^{\infty} \Pr(A_n \mid \mathcal{F}_n)\right\}.$$

Lemma 43 (Theorem C.3 of Mei et al. (2023))**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables, such that for all finite $t \geq 1$, $|X_t| \leq \frac{1}{2}$. Define that*

$$S_n := |\sum_{t=1}^{n} \mathbb{E}[X_t \mid X_1, \ldots, X_{t-1}] - X_t| \text{ and } V_n := \sum_{t=1}^{n} Var[X_t | X_1, \ldots, X_{t-1}].$$

Then, for all $\delta > 0$,

$$\Pr\left(\exists n : S_n \geq 6\sqrt{\left(V_n + \frac{4}{3}\right)\log\left(\frac{V_n + 1}{\delta}\right)} + 2\log(\frac{1}{\delta}) + \frac{4}{3}\log 3\right) \leq \delta.$$

Lemma 44 (Unbiased Gradient)**.** *Using Algorithm 4, we have, for all finite $t \geq 1$,*

$$\mathbb{E}_t\left[\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right] = \frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t}.$$

Proof. First, we show that $\mathbb{E}_t\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t}\right] = \frac{d\langle \pi_{z_t}, r\rangle}{dz_t}$ where $z_t := X\theta_t$.

For the sampled arm a_t, we have,

$$\begin{aligned}
\mathbb{E}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t(a_t)}\right] &= \mathbb{E}_{R_t(a_t)\sim P_{a_t}}\Big[\,(1 - \pi_{z_t}(a_t))\; R_t(a_t)\Big] \\
&= (1 - \pi_{z_t}(a_t))\; \mathbb{E}_{R_t(a_t)\sim P_{a_t}}\Big[R_t(a_t)\Big] \\
&= (1 - \pi_{z_t}(a_t))\; r(a_t).
\end{aligned}$$

For any other arms $a \neq a_t$ that are not sampled, we have,

$$\begin{aligned}\mathbb{E}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t(a)}\right] &= \mathbb{E}_{R_t(a_t)\sim P_{a_t}}\Big[-\pi_{z_t}(a)\, R_t(a_t)\Big] \\ &= -\pi_{z_t}(a)\, \mathbb{E}_{R_t(a_t)\sim P_{a_t}}\Big[R_t(a_t)\Big] \\ &= -\pi_{z_t}(a)\, r(a_t).\end{aligned}$$

Combing the above two equations, we have, for all $a \in [K]$,

$$\mathbb{E}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t(a)}\right] = (\mathbb{I}\{a_t = a\} - \pi_{z_t}(a))\; r(a_t).$$

Taking expectation over $a_t \sim \pi_{z_t}(\,)$, we have,

$$\begin{aligned}\mathbb{E}_t\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t(a)}\right] &= \Pr\left(a_t = a\right) \mathbb{E}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t(a)} \;\Big|\; a_t = a\right] \\ &\quad + \Pr\left(a_t \neq a\right) \mathbb{E}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\langle \pi_{z_t}, \hat{r}_t\rangle}{dz_t(a)} \;\Big|\; a_t \neq a\right] \\ &= \pi_{z_t}(a)\; (1 - \pi_{z_t}(a))\; r(a) + \sum_{a' \neq a} \pi_{z_t}(a')\; (-\pi_{z_t}(a))\; r(a') \\ &= \pi_{z_t}(a) \sum_{a' \neq a} \pi_{z_t}(a')\; (r(a) - r(a')) \\ &= \pi_{z_t}(a)\; (r(a) - \langle \pi_{z_t}, r\rangle) \\ &= \frac{d\langle \pi_{z_t}, r\rangle}{dz_t(a)}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}_t\left[\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{d\theta_t}\right] = X^\top\, \mathbb{E}_t\left[\frac{d\langle \pi_{\theta_t}, \hat{r}_t\rangle}{dz_t}\right] = X^\top\, \frac{d\langle \pi_{z_t}, r\rangle}{dz_t(a)} = \frac{d\langle \pi_{\theta_t}, r\rangle}{d\theta_t}.$$

□

Lemma 45 (Smoothness)**.** *Given any reward vector $r \in \mathbb{R}^K$ and feature matrix $X \in \mathbb{R}^{K\times d}$. The expected reward function $\theta \mapsto \langle \pi_\theta, r\rangle$ with $\pi_\theta = \mathrm{softmax}(X\theta)$ is L-smooth with*

$$L = \frac{9}{2}\, \|r\|_\infty\, \lambda_{\max}(X^\top X), \tag{B.29}$$

Proof. Let $S \coloneqq S(X, r, \theta) \in \mathbb{R}^{d\times d}$ be the second-order derivative of the value map $\theta \mapsto \langle \pi_\theta, r\rangle$. By Taylor's theorem, it suffices to show that the spectral radius of S (regardless of θ) is bounded by L. Now, by its definition we have

$$\begin{aligned}S &= \frac{d}{d\theta}\left\{\frac{d\langle \pi_\theta, r\rangle}{d\theta}\right\} \\ &= \frac{d}{d\theta}\left\{X^\top(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top)\; r\right\}. \qquad \text{(by Equation (4.4))}\end{aligned}$$

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

$$S_{i,j} = \frac{d\{\sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, (r(a) - \pi_\theta^\top r)\}}{d\theta(j)} \tag{B.30}$$
$$= \sum_{a=1}^{K} X_{a,i}\, \frac{d\pi_\theta(a)}{d\theta(j)}\, \Big(r(a) - \pi_\theta^\top r\Big) - \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, \sum_{a'=1}^{K} \frac{d\pi_\theta(a')}{d\theta(j)}\, r(a').$$

We have, for all $a \in [K]$ and $j \in [d]$,

$$\frac{d\pi_\theta(a)}{d\theta(j)} = \frac{d}{d\theta(j)} \Big\{ \frac{\exp\{[X\theta](a)\}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}} \Big\} \tag{B.31}$$
$$= \frac{\frac{d\ \exp\{[X\theta](a)\}}{d\theta(j)}\ \sum_{a' \in [K]} \exp\{[X\theta](a')\} - \exp\{[X\theta](a)\}\ \frac{d\ \sum_{a' \in [K]} \exp\{[X\theta](a')\}}{d\theta(j)}}{\big(\sum_{a' \in [K]} \exp\{[X\theta](a')\}\big)^2}$$
$$= \frac{\exp\{[X\theta](a)\}\, X_{a,j}\ \sum_{a' \in [K]} \exp\{[X\theta](a')\} - \exp\{[X\theta](a)\}\ \sum_{a' \in [K]} \exp\{[X\theta](a')\}\, X_{a',j}}{\big(\sum_{a' \in [K]} \exp\{[X\theta](a')\}\big)^2}$$
$$= \frac{\exp\{[X\theta](a)\}\, X_{a,j} - \exp\{[X\theta](a)\}\ \sum_{a' \in [K]} \pi_\theta(a')\, X_{a',j}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}}$$
$$= \pi_\theta(a) \Big(X_{a,j} - \sum_{a' \in [K]} \pi_\theta(a')\, X_{a',j}\Big).$$

Combining Equations (B.30) and (B.31), we have,

$$S_{i,j} = \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, (r(a) - \pi_\theta^\top r)\, X_{a,j} - \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, (r(a) - \pi_\theta^\top r)\, \sum_{a'=1}^{K} \pi_\theta(a')\, X_{a',j}$$
$$- \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, \sum_{a'=1}^{K} \pi_\theta(a') \Big(X_{a',j} - \sum_{a''=1}^{K} \pi_\theta(a'')\, X_{a'',j}\Big)\, r(a').$$

To show the bound on the spectral radius of S, pick $y \in \mathbb{R}^d$. Then,

$$\Big|y^\top S y\Big| = \Big|\sum_{i=1}^{d} \sum_{j=1}^{d} S_{i,j}\, y(i)\, y(j)\Big|$$
$$= \Big| \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{K} y(i)\, X_{a,i}\, \pi_\theta(a)\, (r(a) - \pi_\theta^\top r)\, X_{a,j}\, y(j)$$
$$- \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{K} y(i)\, X_{a,i}\, \pi_\theta(a)\, (r(a) - \pi_\theta^\top r)\, \sum_{a'=1}^{K} \pi_\theta(a')\, X_{a',j}\, y(j)$$
$$- \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{K} y(i)\, X_{a,i}\, \pi_\theta(a)\, \sum_{a'=1}^{K} \pi_\theta(a') \Big(X_{a',j} - \sum_{a''=1}^{K} \pi_\theta(a'')\, X_{a'',j}\Big)\, r(a')\, y(j)\Big|,$$

which is equal to,

$$\begin{aligned}\left|y^\top S y\right| = \Big| &\sum_{a=1}^{K} [Xy](a)\,\pi_\theta(a)\,(r(a) - \pi_\theta^\top r)\,[Xy](a) \\ &- \sum_{a=1}^{K} [Xy](a)\,\pi_\theta(a)\,(r(a) - \pi_\theta^\top r) \sum_{a'=1}^{K} \pi_\theta(a')\,[Xy](a') \\ &- \sum_{a=1}^{K} [Xy](a)\,\pi_\theta(a) \sum_{a'=1}^{K} \pi_\theta(a')\,r(a') \Big([Xy](a') - \sum_{a''=1}^{K} \pi_\theta(a'')\,[Xy](a'')\Big)\Big|.\end{aligned}$$

Denote

$$H(\pi_\theta) \coloneqq \mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \in \mathbb{R}^{K \times K}.$$

We have,

$$\begin{aligned}\left|y^\top S y\right| &= \left|\left(H(\pi_\theta)\; r\right)^\top (Xy \odot Xy) - \left(H(\pi_\theta)\; r\right)^\top (Xy)\,(\pi_\theta^\top Xy) - (\pi_\theta^\top Xy)\,\left(H(\pi_\theta) Xy\right)^\top r\right| \\ &= \left|\left(H(\pi_\theta)\; r\right)^\top (Xy \odot Xy) - 2\left(H(\pi_\theta)\; r\right)^\top (Xy)\,(\pi_\theta^\top Xy)\right|,\end{aligned}$$

where $\odot$ is Hadamard (component-wise) product. According to the triangle inequality and Hölder's inequality, we have,

$$\begin{aligned}\left|y^\top S y\right| &\le \left|\left(H(\pi_\theta)\; r\right)^\top (Xy \odot Xy)\right| + 2\left|\left(H(\pi_\theta)\; r\right)^\top (Xy)\right| |\pi_\theta^\top Xy| \\ &\le \|H(\pi_\theta) r\|_\infty \,\|Xy \odot Xy\|_1 + 2\,\|H(\pi_\theta) r\|_1\,\|Xy\|_\infty\,\|\pi_\theta\|_1\,\|Xy\|_\infty \\ &= \|H(\pi_\theta) r\|_\infty \,\|Xy\|_2^2 + 2\,\|H(\pi_\theta) r\|_1\,\|Xy\|_\infty^2 \qquad \left(\|Xy \odot Xy\|_1 = \|Xy\|_2^2,\ \|\pi_\theta\|_1 = 1\right) \\ &\le \|H(\pi_\theta) r\|_\infty \,\|Xy\|_2^2 + 2\,\|H(\pi_\theta) r\|_1\,\|Xy\|_2^2. \qquad (\|Xy\|_\infty \le \|Xy\|_2)\end{aligned}$$

For $a \in [K]$, denote by $H_{a,:}(\pi_\theta)$ the a-th row of $H(\pi_\theta)$ as a row vector. Then,

$$\begin{aligned}\|H_{a,:}(\pi_\theta)\|_1 &= \pi_\theta(a) - \pi_\theta(a)^2 + \pi_\theta(a) \sum_{a' \ne a} \pi_\theta(a') \\ &= \pi_\theta(a) - \pi_\theta(a)^2 + \pi_\theta(a)\,(1 - \pi_\theta(a)) \\ &= 2\,\pi_\theta(a)\,(1 - \pi_\theta(a)) \\ &\le \frac{1}{2}. \qquad (\text{using } x\,(1-x) \le 1/4 \text{ for all } x \in [0,1])\end{aligned}$$

On the other hand,

$$\begin{aligned}\|H(\pi_\theta) r\|_1 &= \sum_{a \in [K]} \pi_\theta(a)\,\left|r(a) - \pi_\theta^\top r\right| \\ &\le \max_{a \in [K]} \left|r(a) - \pi_\theta^\top r\right| \\ &\le 2\,\|r\|_\infty. \qquad \left(\text{using } r \in \left[-\|r\|_\infty, \|r\|_\infty\right]^K\right)\end{aligned}$$

Therefore, we have,

$$\begin{aligned}
\left|y^\top S(X,r,\theta)\ y\right| &\le \|H(\pi_\theta)r\|_\infty \ \|Xy\|_2^2 + 2\ \|H(\pi_\theta)r\|_1\ \|Xy\|_2^2 \\
&= \max_{a\in[K]} \left|(H_{a,:}(\pi_\theta))^\top r\right| \ \|Xy\|_2^2 + 2\ \|H(\pi_\theta)r\|_1\ \|Xy\|_2^2 \\
&\le \max_{a\in[K]} \|H_{a,:}(\pi_\theta)\|_1\ \|r\|_\infty\ \|Xy\|_2^2 + 4\,\|r\|_\infty\ \|Xy\|_2^2 \\
&\le \Big(\frac{1}{2}+4\Big)\ \|r\|_\infty\ \|Xy\|_2^2 \\
&\le \frac{9}{2}\ \|r\|_\infty\ \|X\|_{\text{op}}^2\ \|y\|_2^2 \\
&= \frac{9}{2}\ \|r\|_\infty\, \lambda_{\max}(X^\top X)\ \|y\|_2^2\,,
\end{aligned}$$

where $\|X\|_{\text{op}}$ is the operator norm of $X \in \mathbb{R}^{K\times d}$ (squared root of largest eigenvalue of $X^\top X$),

$$\|X\|_{\text{op}} = \sup\left\{\|Xv\|_2 : \|v\|_2 \le 1,\ v \in \mathbb{R}^d\right\}.$$

According to Taylor's theorem, for all $\theta,\ \theta' \in \mathbb{R}^d$, there exists $\theta_\zeta := \zeta\,\theta + (1-\zeta)\ \theta'$ with $\zeta \in [0,1]$, such that,

$$\begin{aligned}
\left|(\pi_{\theta'} - \pi_\theta)^\top r - \Big\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta\Big\rangle\right| &= \frac{1}{2}\ \left|\left(\theta'-\theta\right)^\top S(X,r,\theta_\zeta)\ \left(\theta'-\theta\right)\right| \\
&\le \frac{9}{4}\ \|r\|_\infty\, \lambda_{\max}(X^\top X)\ \|\theta'-\theta\|_2^2.
\end{aligned}$$

□

Lemma 46 (Non-uniform smoothness)**.** *For all $\theta \in \mathbb{R}^d$, the spectral radius of Hessian matrix $\frac{d^2\{\langle \pi_\theta, r\rangle\}}{d^2\theta^2} \in \mathbb{R}^{d\times d}$ is upper bounded by $3\,\lambda_{\max}[X^\top X]\|\frac{d\langle \bar\pi_z, r\rangle}{dz}\|$, i.e. for all $y \in \mathbb{R}^d$,*

$$|y^\top \frac{d^2\{\langle \pi_\theta, r\rangle\}}{d^2\theta^2} y| \le 3\,\lambda_{\max}[X^\top X]\|\frac{d\langle \bar\pi_z, r\rangle}{dz}\|\ \|y\|_2^2.$$

Proof. Following the initial proof of Lemma 45, let $S := S(r,\theta) \in \mathbb{R}^{d\times d}$ be the second derivative of the map $\theta \to \langle \pi_\theta, r\rangle$,

$$\begin{aligned}
S &= \frac{d}{d\theta}\left\{\frac{d\langle \pi_\theta, r\rangle}{d\theta}\right\} \\
&= \frac{d}{d\theta}\left\{X^\top H(\pi_\theta)\, r\right\}.
\end{aligned}$$

For fixed $i, j \in [d]$,

$$\begin{aligned}
S_{i,j} &= \frac{d\,[X^\top H(\pi_\theta)\, r](i)}{d\theta(j)} \\
&= \frac{d\,[\sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle)}{d\theta(j)} \\
&= \sum_{a=1}^{K} X_{a,i}\, \frac{\pi_\theta(a)}{d\theta(j)}\, (r(a) - \langle \pi_\theta, r \rangle) - \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a) \sum_{a'=1}^{K} \frac{d\pi_\theta(a')}{d\theta(j)}\, r(a').
\end{aligned}$$

For all $a \in [K]$ and $j \in [d]$

$$\begin{aligned}
\frac{d\pi_\theta(a)}{d\theta(j)} &= \frac{d}{d\theta(j)} \left\{ \frac{\exp([X\theta](a))}{\sum_{a' \in [K]} \exp([X\theta](a'))} \right\} \\
&= \frac{\frac{d \exp([X\theta](a))}{d\theta(j)} \sum_{a' \in [K]} \exp([X\theta](a')) - \exp([X\theta](a)) \frac{d \sum_{a' \in [K]} \exp([X\theta](a'))}{d\theta(j)}}{(\sum_{a' \in [K]} \exp([X\theta](a'))^2} \\
&= \frac{\exp([X\theta](a))\, X_{a,j} \sum_{a' \in [K]} \exp([X\theta](a')) - \exp([X\theta](a)) \sum_{a' \in [K]} \exp([X\theta](a'))\, X_{a',j}}{(\sum_{a' \in [K]} \exp([X\theta](a'))^2} \\
&= \frac{\exp([X\theta](a))\, X_{a,j} - \exp([X\theta](a)) \sum_{a' \in [K]} \pi_\theta(a')\, X_{a',j}}{\sum_{a' \in [K]} \exp([X\theta](a'))} \\
&= \pi_\theta(a) \left(X_{a,j} - \sum_{a' \in [K]} \pi_\theta(a')\, X_{a',j} \right)
\end{aligned}$$

Combining the above inequalities,

$$\begin{aligned}
S_{i,j} = &\sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle)\, X_{a,j} - \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle) \sum_{a'=1}^{K} \pi_\theta(a')\, X_{a',j} \\
&- \sum_{a=1}^{K} X_{a,i}\, \pi_\theta(a) \sum_{a'=1}^{K} \pi_\theta(a') \left(X_{a',j} - \sum_{a''=1}^{K} \pi_\theta(a'')\, X_{a'',j} \right) r(a').
\end{aligned}$$

To show the bound on the spectral radius of S, pick $y \in \mathbb{R}^d$. Then,

$$\begin{aligned}
|y^\top S y| &= |\sum_{i=1}^{d} \sum_{j=1}^{d} S_{i,j}\, y(i)\, y(j)| \\
&= \Bigg| \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{K} y(i)\, X_{a,i}\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle)\, X_{a,j}\, y(j) \\
&\quad - \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{K} y(i)\, X_{a,i}\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle) \sum_{a'=1}^{K} \pi_\theta(a')\, X_{a',j}\, y(j) \\
&\quad - \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{a=1}^{K} y(i)\, X_{a,i}\, \pi_\theta(a) \sum_{a'=1}^{K} \pi_\theta(a') \left(X_{a',j} - \sum_{a''=1}^{K} \pi_\theta(a'')\, X_{a'',j} \right) r(a')\, y(j) \Bigg|
\end{aligned}$$

which is equal to,

$$
\begin{aligned}
|y^\top S y| = \Bigg| & \sum_{a=1}^{K} [Xy](a)\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle)\, [Xy](a) \\
& - \sum_{a=1}^{K} [Xy](a)\, \pi_\theta(a)\, (r(a) - \langle \pi_\theta, r \rangle) \sum_{a'=1}^{K} \pi_\theta(a')\, [Xy](a') \\
& - \sum_{a=1}^{K} [Xy](a)\, \pi_\theta(a) \sum_{a'=1}^{K} \pi_\theta(a')\, r(a') \left([Xy](a') - \sum_{a''=1}^{K} \pi_\theta(a'')\, [Xy] a'' \right) \Bigg|.
\end{aligned}
$$

Denote

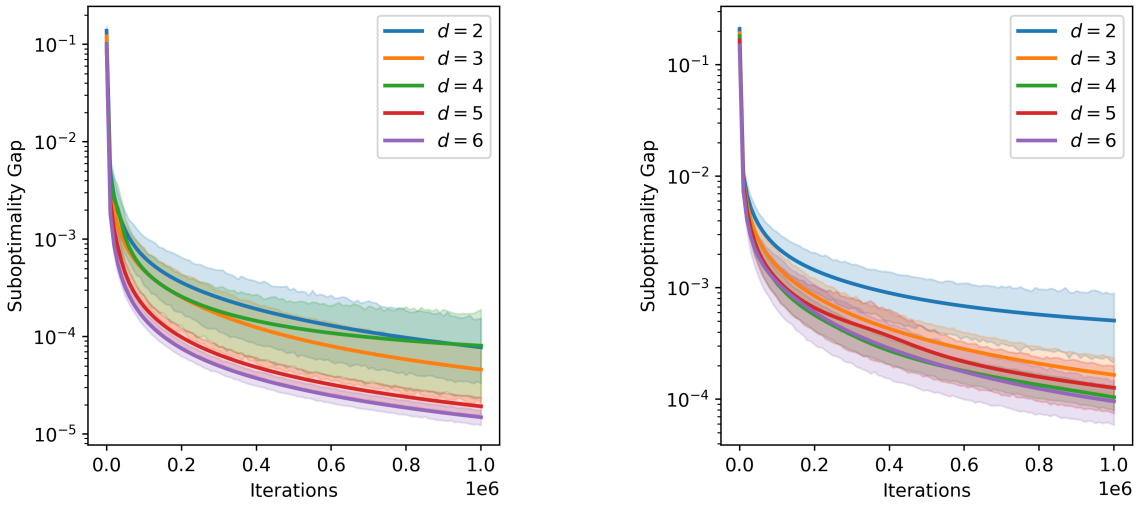
$$
H(\pi_\theta) = \mathrm{diag}(\pi_\theta) - \pi_\theta\, \pi_\theta^\top \in \mathbb{R}^{K \times K}.
$$

We then have,

$$
\begin{aligned}
|y^\top S y| &= |(H(\pi_\theta)\, r)^\top (Xy \odot Xy) - (H(\pi_\theta)\, r)^\top (Xy)\, (\pi_\theta^\top Xy) - \left(\pi_\theta^\top Xy \right) (H(\pi_\theta)\, Xy)^\top r| \\
& \qquad\qquad (\odot \text{ is the Hadamard (component-wise) product}) \\
&= |(H(\pi_\theta)\, r)^\top (Xy \odot Xy) - 2\, (H(\pi_\theta)\, r)^\top (Xy)\, (\pi_\theta^\top Xy)| \\
&\leq |(H(\pi_\theta)\, r)^\top (Xy \odot Xy)| + 2\, |(H(\pi_\theta)\, r)^\top (Xy)\, (\pi_\theta^\top Xy)| \qquad \text{(Triangle Inequality)} \\
&\leq \|H(\pi_\theta)\, r\|_\infty \|Xy \odot Xy\|_1 + 2\, \|H(\pi_\theta)\, r\|\, \|Xy\|\, \|\pi_\theta\|_1 \|Xy\|_\infty \quad \text{(Hölder's inequality)} \\
&\leq 3 \|H(\pi_\theta)\, r\|\, \|Xy\|_2^2 \qquad\qquad (\|\,\|_\infty \leq \|\,\|,\ \|Xy \odot Xy\|_1 = \|Xy\|_2^2,\ \|\pi_\theta\|_1 \leq 1) \\
&\leq 3\, \lambda_{\max}[X^\top X]\, \|H(\pi_\theta)\, r\|\, \|y\|_2^2 \\
&= 3\, \lambda_{\max}[X^\top X]\, \| \frac{d \langle \bar{\pi}_z, r \rangle}{dz} \|\, \|y\|_2^2
\end{aligned}
$$

□

B.6 Experiments



(a) $K = 3$ (b) $K = 6$

Figure B.1: Softmax PG on exact linear bandits. The learning rate is set by Equation (4.5). Each experiment is run on 50 randomly generated environments for 10^6 iterations. For each environment, the features X and the reward vector r are randomly generated such that Assumption 8 is satisfied, and the features satisfy Assumption 9 when (a) $K = 3$ and satisfy Assumption 10 when (b) $K = 6$. Softmax PG converges to the optimal policy for different feature dimensions d, confirming the results of Theorems 4 and 7.