TOWARDS A THEORETICAL UNDERSTANDING TO THE GENERALIZATION OF RLHF

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) and its variants have emerged as the dominant approaches for aligning Large Language Models (LLMs) with human intent. While empirically effective, the theoretical generalization properties of these methods in high-dimensional settings remain to be explored. To this end, we build the generalization theory on RLHF of LLMs under the linear reward model, through the framework of algorithmic stability. In contrast to the existing works built upon the consistency of maximum likelihood estimations on reward model, our analysis is presented under an end-to-end learning framework, which is consistent with practice. Concretely, we prove that under a key **feature coverage** condition, the empirical optima of the policy model have a generalization bound of order $\mathcal{O}(n^{-\frac{1}{2}})$. Moreover, the results can be extrapolated to parameters obtained by gradient-based algorithms, i.e., Gradient Ascent (GA) and Stochastic Gradient Ascent (SGA). Thus, we argue that our results provide new theoretical evidence for the empirically observed generalization of LLMs after RLHF.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) and its variants have emerged as the dominant approaches for aligning Large Language Models (LLMs) with human intent(Christiano et al., 2017; Ouyang et al., 2022; Guo et al., 2025; Yang et al., 2025; Comanici et al., 2025). Guided by this paradigm, LLMs have demonstrated remarkable capabilities across diverse domains, such as image generation (Liu et al., 2024; Shen et al., 2025), Embodied AI (Intelligence et al., 2025; Ye et al., 2025) and so on. While empirically effective, the theoretical understanding of the generalization properties of RLHF remains to be explored, particularly in high-dimensional settings.

Currently, the theoretical frameworks of RLHF for LLMs largely rely on the structural insight that the optimal policy for the RLHF objective admits a closed-form posterior Boltzmann distribution depending on the reward model (Rafailov et al., 2023). Building on this, the existing literature (Zhu et al., 2023; Xiong et al., 2023) typically focus on the statistical quality of reward modeling: they analyze the convergence of reward parameters learned via Maximum Likelihood Estimation (MLE) from offline data, and subsequently treat the estimated parameters in the closed-form posterior Boltzmann distribution as the learned policy.

However, this analytical paradigm faces two fundamental limitations. First, it necessitates stringent data coverage assumptions to guarantee the consistency of reward parameters(Song et al., 2024), which is rarely satisfied in high-dimensional regimes (Ethayarajh, 2019; Aghajanyan et al., 2021; Ruppik et al., 2025). Second, and perhaps more critically, it focuses solely on the supervised learning nature of reward modeling. By treating policy optimization as an exact maximization step, such analyses fail to capture the benefits inherent to the reinforcement learning process itself.

To address these limitations, we propose an end-to-end theoretical framework that captures the dynamic interaction between the policy and the environment during practical RLHF training. Our analyses are conducted on the KL-regularized objective (Zhao et al., 2024; Azar et al., 2024) under the standard linear reward assumption (Jin et al., 2020). Within this framework, our analysis focus on the suboptimality gap of the learned policy, isolating the generalization error as a critical component.

Technically, our analyses are grounded in a key **feature coverage assumption**, which intuitively posits that as the dataset grows, the column space of the empirical covariance matrix of model's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

feature progressively expands to cover the feature space. Consequently, the residual component of any feature vector falling outside the covered subspace diminishes. This geometric property allows us to rigorously evaluate the generalization error of model obtained across different optimization regimes, varying from the empirical optima to models obtained by gradient-based methods.

Our theoretical results can be summarized as follows:

- **The Suboptimality Gap on Empirical Optima.** Under a strong no-degenerate assumption (Assumption 3), we prove the empirical optima can match the ground truth parameters when data is sufficient. Moreover, under the key feature coverage assumption (Assumption 5), by analyzing the algorithmic stability (Bousquet & Elisseeff, 2002; Yi et al., 2022), we prove a $\tilde{\mathcal{O}}(n^{-\frac{1}{2}})$ bound of suboptimality gap on empirical optima, when we have n training data.
- **The Suboptimality Gap under Gradient-Based Algorithms.** We extend our algorithmic stability analysis to practical iterative optimization. Under the key feature coverage assumption, we derive explicit suboptimality gap bounds for Gradient Ascent (GA) and Stochastic Gradient Ascent (SGA) of order $\tilde{\mathcal{O}}(T^{-\frac{1}{4}} + n^{-\frac{1}{2}})$ and $\tilde{\mathcal{O}}(T^{-\frac{1}{8}} + n^{-\frac{1}{2}})$, respectively.

The remainder is organized as follows. Sections 2 and 3 review related work and introduce the formal setting, respectively. Section 4 formally defines the suboptimality gap and presents its decomposition. Section 5 analyzes the suboptimality gap of the empirical optimum with respect to the true optimal policy. The analysis based on the algorithmic stability is extended to the gradient-based optimization algorithms (GA and SGA) in Section 6. Section 7 empirically verifies our theoretical results.

2 Related Work

Theoretical Foundations of RLHF. The theoretical foundations of RLHF are deeply rooted in dueling bandits (Dudík et al., 2015; Bengs et al., 2021; Chang et al., 2024) and preference-based reinforcement learning (Wirth et al., 2017; Xu et al., 2020) frameworks. For dueling bandits, the learner aims to minimize regret based on pairwise comparisons between actions. Theoretical analyses have progressively expanded from the tabular setting (Zhan et al., 2023) to linear models (Xiong et al., 2023; Saha et al., 2023; Xu et al., 2024) and general function approximation (Chen et al., 2022). These analyses are extended to RLHF or preference-based optimization (Rafailov et al., 2023) due to their empirical success. Zhu et al. (2023) quantified the estimation error between the reward parameters learned via MLE and the ground truth, and showed that pessimistic MLE guarantees low sub-optimality of the induced policy. Subsequent work has extended this framework to DPO and other related algorithms (Ethayarajh et al., 2024; Bai et al., 2025). Unlike prior works (Xiong et al., 2023; Zhu et al., 2023; Chaudhari et al., 2025) that primarily evaluate the consistency between parameters estimated from offline preference data and the ground truth, we adopt an end-to-end analytical framework and directly analyze the suboptimality gap of the policy obtained by RL algorithm.

Algorithmic Stability and Generalization. Algorithmic stability, rigorously formalized by (Bousquet & Elisseeff, 2002), provides a powerful framework for bounding generalization error. While early works focused on deterministic algorithms, (Hardt et al., 2016) expanded this framework to randomized algorithms, establishing uniform stability guarantees for Stochastic Gradient Descent in both convex and non-convex settings (Bassily et al., 2020; Yi et al., 2022). (Feldman & Vondrak, 2019) and (Bousquet et al., 2020) derived sharper high-probability bounds by improving dependence on sample size and failure probability. Complementing these statistical advancements, recent works have grounded stability in the optimization dynamics, showing that the algorithm's sensitivity is explicitly controlled by the gradient norm (Kuzborskij & Lampert, 2018; Lei & Ying, 2020; Wang & Ma, 2022; Zhu et al., 2025). Despite these advancements in supervised learning, the application of stability analysis to Reinforcement Learning remains limited. Existing works primarily focus on value-based methods in tabular settings (Fan et al., 2020; Agarwal et al., 2020). In this paper, we bridge this gap by applying the uniform stability framework to a KL-regularized RLHF objective, deriving explicit generalization bounds for GA and SGA in the context of RLHF.

3 Formal Setting

This section presents the background material that will be used throughout the paper.

Notation. We use $\oplus$ for direct sum and $\|\cdot\|$ for the Euclidean norm of vectors and the spectral norm for matrices. For a matrix $\boldsymbol{A}$, let $\mathrm{C}(\boldsymbol{A})$ and $\mathrm{N}(\boldsymbol{A})$ be its column and null spaces, while $\sigma_{\max}(\boldsymbol{A})$,

$\sigma_{\min}(\boldsymbol{A})$ and $\sigma^{+}_{\min}(\boldsymbol{A})$ represent its largest, smallest, and smallest positive eigenvalues, respectively. $\mathcal{X}$ and $\mathcal{A}$ denote the context and action spaces. Finally, $\mathcal{O}(\cdot)$ refers to the standard big-O notation, $\tilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors, and $a \lesssim b$ implies $a \leq Cb$ for some constant $C > 0$.

3.1 PRELIMINARIES

Reinforcement Learning from Human Feedback. We situate our work within the RLHF paradigm as formulated in Ziegler et al. (2019), given a dataset of prompts $\boldsymbol{S} = \{x_1, \cdots, x_n\}$ and a fixed reference policy π_{ref}. Unlike the standard two-stage RLHF framework where a reward model $r(x, a)$ is learned from preferences, we operate in a setting where ground-truth rewards on prompts $x \in \mathcal{X}$ and response $a \in \mathcal{A}$ are available. For a given x, denote the objective as:

$$f_\pi(x) = \mathbb{E}_{a\sim\pi(\cdot|x)}\left[r(x, a)\right] - \lambda \mathrm{D}_{\mathrm{KL}}(\pi(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x)). \tag{1}$$

The goal of RLHF is to learn a policy π to maximize the objective $J(\pi)$, where:

$$J(\pi) = \mathbb{E}_{x\in\mathcal{X}}[f_\pi(x)], \tag{2}$$

i.e., we want to maximize the expected reward without deviating too much from the reference policy. W.l.o.g., to simplify following analysis, we set $\lambda = 1$ below. In the case that $\lambda \neq 1$, we divide both sides of (1) by λ, and treat the term $\frac{r(x,a)}{\lambda}$ as a new reward function. This rescaling does not affect the solutions of the original objective function.

Before introducing the parametric model, we first characterize the optimal solution to (2) over the space of all valid probability distributions.

Proposition 1 (Rafailov et al. (2023))**.** *For any reward function $r(x, a)$, the optimal policy π^* that maximizes* (2) *admits the following closed-form expression: $\pi^*(a|x) = \frac{\pi_{\text{ref}}(a|x)\exp(r(x,a))}{\sum_{a'}\pi_{\text{ref}}(a|x)\exp(r(x,a))}$.*

Motivated by this optimal form, we adopt a parametric approach to approximate the reward and the policy. We consider the widely used linear reward assumption (Zhu et al., 2023; Xiong et al., 2023) and the corresponding posterior Boltzmann policy class.

Definition 1 (linear reward function class)**.** *Let $\boldsymbol{\phi}(x, a) \in \mathbb{R}^d$ be a fixed feature mapping. We consider the class of linear reward functions defined as: $\mathcal{F} = \{r_{\boldsymbol{\theta}} | r_{\boldsymbol{\theta}}(x, a) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a)\rangle\}$.*

Correspondingly, motivated by the optimal solution form in Proposition 1, consider policy class as:

Definition 2 (Posterior Boltzmann policy function class)**.**

$$\Pi = \left\{ \pi_{\boldsymbol{\theta}} | \pi_{\boldsymbol{\theta}}(a|x) = \frac{\pi_{\text{ref}}(a|x)\exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a)\rangle\right)}{\sum_{a'\in\mathcal{A}}\pi_{\text{ref}}(a'|x)\exp\left(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x, a)\rangle\right)} \right\}.$$

To facilitate theoretical analysis, we make the following standard regularity assumptions regarding the boundedness of features and the realizability of the true reward.

Assumption 1. *For all $(x, a) \in \mathcal{X} \times \mathcal{A}$, the feature norm is bounded such that $\|\boldsymbol{\phi}(x, a)\| \leq C$.*

Assumption 2. *There exists a ground-truth parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that the true reward function is linear, i.e. $r(x, a) = \langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(x, a)\rangle$, with the parameter bounded by $\|\boldsymbol{\theta}^*\| \leq D$.*

Under Assumption 2 and Proposition 1, it is straightforward to verify that the policy parameterized by ground-truth $\boldsymbol{\theta}^*$, denoted as $\pi_{\boldsymbol{\theta}^*}$, is the global maximizer of the expected objective $J(\pi_{\boldsymbol{\theta}})$, i.e.,

$$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} J(\pi_{\boldsymbol{\theta}}). \tag{3}$$

However, the context space $\mathcal{X}$ is vast, making direct maximization of the expected objective in (2) computationally intractable. In practice, we instead approximate it using a finite prompt dataset $\boldsymbol{S} = \{x_i\}_{i=1}^n$, leading to the following empirical objective:

$$\max_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i), \tag{4}$$

We use $\boldsymbol{\theta}_{\boldsymbol{S}}$ to denote the parameters obtained on $\boldsymbol{S}$. For our analysis, we focus on the empirical stationary points within a bounded region $\Theta_R := \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\| \leq R\}$ in (5), where we set $R = 3D$. This focus is sufficient because the parameters obtained by gradient-based algorithms are naturally confined within this bound (see Lemma 19), eliminating the need to consider the entire space.

$$\boldsymbol{\theta}^*_{\boldsymbol{S}} \in \{\boldsymbol{\theta} \in \Theta_R | \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}) = 0\} \triangleq \Theta_{\boldsymbol{S}}. \tag{5}$$

4 SUBOPTIMALITY DECOMPOSITION

To rigorously quantify the performance of the learned policy on the unknown context space $\mathcal{X}$, we define the sub-optimality gap and decompose it into three distinct components: concentration, optimization, and generalization.

Definition 3 (Suboptimality Gap Decomposition)**.** *Let $\boldsymbol{\theta}_{\boldsymbol{S}}$ be the output of an algorithm on dataset $\boldsymbol{S}$, and $\boldsymbol{\theta}^*_{\boldsymbol{S}} \in \Theta_{\boldsymbol{S}}$ is one of the empirical stationary point. The suboptimality gap is decomposed as:*

$$\begin{aligned}&\mathrm{Subopt}(\boldsymbol{\theta}_{\boldsymbol{S}}) = J(\pi_{\boldsymbol{\theta}^*}) - J(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}}) \\ &\leq \underbrace{|J(\pi_{\boldsymbol{\theta}^*}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})|}_{\text{Concentration error}} + \underbrace{|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}})|}_{\text{Optimization error}} + \underbrace{|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}})|}_{\text{Generalization error}},\end{aligned} \tag{6}$$

Optimization error. The optimization error quantifies the performance gap between the learned parameter $\boldsymbol{\theta}_{\boldsymbol{S}}$ and the empirical stationary point $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ on the training objective $J_{\boldsymbol{S}}$. In the context of non-convex optimization, iterative algorithms typically possess theoretical guarantees of convergence to empirical stationary points. Crucially, based on Lemma 6, we can demonstrate that such optimization error is explicitly controlled by the gradient norm $\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}})\|$.

Corollary 1. *Under Assumption 5, for any $\boldsymbol{\theta}_{\boldsymbol{S}} \in \Theta_R$ and $\delta > 0$, if $\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}})\| \leq \delta$, then*

$$\begin{aligned}|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}})| \leq & 4R\epsilon_n + \frac{2C\delta \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}))} \\ & + 2RC\sqrt{4R\epsilon_n + \frac{2C\delta \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}))}}.\end{aligned}$$

Remark 1. *In this corollary, the optimization error is obtained under the feature coverage Assumption 5, which results in the ϵ_n appeared in the upper bound.*

Moreover, for all empirical stationary points (Proposition 4), their empirical risks are identical, i.e., $J_{\boldsymbol{S}}(\boldsymbol{\theta})$ is invariant over the set $\{\boldsymbol{\theta} \mid \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}) = 0\}$. Thus, all local optima are global optima. [1]

Concentration error. The concentration error quantifies the statistical deviation of the empirical optimal objective value from the expected optimal objective value. Since the parameter $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ in $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})$ depends on the dataset $\boldsymbol{S}$, the standard concentration inequality does not work.

Fortunately, as discussed above, we can prove that $\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^*) = \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$, so that $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) = J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*})$ and $\mathbb{E}_{\boldsymbol{S}}[J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})] = J(\pi_{\boldsymbol{\theta}^*})$. Therefore, the standard concentration can be applied:

Lemma 1. (Concentration inequality) Under Assumptions 1, 2 and 5, for any $n \geq 1$ and $\rho \in (0, 1)$, the following bound holds with probability at least $1 - \rho$ over the random draw of the dataset $\boldsymbol{S}$,

$$|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})| \leq 3RC\sqrt{\frac{2\log(2/\rho)}{n}}. \tag{7}$$

Generalization error. The generalization error measures the discrepancy between the true expected performance of the learned policy and its empirical performance. This gap arises because the policy $\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}}$ is optimized based on the training set $\boldsymbol{S}$, rendering the empirical objective a potentially biased estimate of the true objective.

To rigorously bound this error, we adopt the framework of *algorithmic stability*. The central premise is that if a learning algorithm is insensitive to small perturbations in the dataset, it prevents overfitting and ensures that the empirical objective is a reliable proxy for the expected objective.

For algorithm $\mathcal{H} : \boldsymbol{S} \to \mathbb{R}^d$, we first formally define the generalization error and the notion of uniform stability used in our analysis.

Definition 4 (Generalization Error)**.** *For learned parameters $\mathcal{H}(\boldsymbol{S}) = \boldsymbol{\theta}_{\boldsymbol{S}}$, the generalization error is the gap between its empirical/population risks, which is formalized as: $\left|J(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}})\right|$.*

Definition 5 (Uniform Stability Bousquet & Elisseeff (2002))**.** *An algorithm $\mathcal{H}$ is $\mathcal{E}_{\mathrm{stab}}$-uniformly stable if for any two datasets $\boldsymbol{S}, \boldsymbol{S}' \in \mathcal{X}^n$ that differ by at most one example, the uniform stability is:*

$$\mathcal{E}_{\mathrm{stab}}(\mathcal{H}(\boldsymbol{S})) = \sup_{x \in \mathcal{X}, \boldsymbol{S}, \boldsymbol{S}'} \left|f_{\mathcal{H}(\boldsymbol{S})}(x) - f_{\mathcal{H}(\boldsymbol{S}')}(x)\right|. \tag{8}$$

[1] This does not imply PL-inequality (Karimi et al., 2016); this only serve as a necessary condition for it.

It has been proven that uniform stability implies generalization, which is the theoretical foundation for our subsequent analysis in Sections 4 and 5.

Proposition 2 (Bousquet et al. (2020)). *If the algorithm $\mathcal{H}$ satisfies uniform stability and $|f_{\boldsymbol{\theta}}(\cdot)| \leq 3RC$, we have that for any $\rho \in (0,1)$, with probability at least $1-\rho$,*

$$|J(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}})| \lesssim \mathcal{E}_{\text{stab}} \log(n) \log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}}. \tag{9}$$

5 Suboptimality Gap of Empirical Stationary Points

In this section, we first analyze the performance of the empirical stationary points, i.e., $\boldsymbol{\theta}^*_{\boldsymbol{S}} \in \Theta_{\boldsymbol{S}}$. According to the discussion in Section 4, the critical point is to bound the generalization error.

To this end, we analyze the error in two distinct regimes: the ideal regime with sufficient prompt coverage (Assumption 3), and a more realistic regime where prompts are insufficient to cover the entire feature space (Assumption 5).

5.1 RL with Sufficient Prompts

Standard machine learning theory typically relies on the assumption that the training data provides sufficient coverage of feature patterns, thereby ensuring that the learned model can accurately extrapolate to the entire feature space. In linear settings, this requirement is formally characterized by the positive definiteness of the feature covariance matrix.

Assumption 3. *For any $\boldsymbol{\theta}$, x and some σ, it holds $\mathbb{E}_x\left[\mathsf{Var}_{\pi_{\boldsymbol{\theta}}(a|x)}[\boldsymbol{\phi}(x,a)]\right] = \mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})] \succeq \sigma \boldsymbol{I}$.*

As can be seen, this Assumption implies the expected covariance matrix is positive definite, meaning its column space covers the whole Euclidean space, and hence the feature space. Within this coverage assumption to the feature space, we have the following theorem, which states that with sufficient prompts, the empirical stationary point is exactly the ground-truth parameters with high probability.

Theorem 1. *Under Assumption 3, when*

$$n > \frac{128C^4}{\sigma^2}\left(\log\left(\frac{1}{\rho}\right) + d\log\left(\frac{432RC^5}{\sigma^2}\right)\right), \tag{10}$$

the ground truth $\boldsymbol{\theta}^*$ is the unique empirical stationary point in Θ_R with probability at least $1-\rho$.

Our result indicates that with sufficient data coverage, the empirical stationary point achieves exact identification of the ground truth $\boldsymbol{\theta}^*$ with high probability.

Notably, this result relies on the assumption that feature vectors span the entire Euclidean space. However, in practice, due to limited prompt diversity, feature vectors typically reside on a low-dimensional manifold. This motivates our analysis of the insufficient coverage regime in the following.

5.2 RL without Sufficient Prompts

As discussed at the end of Section 5.1, the strict feature coverage assumption is often not met in practical scenarios. In regimes where the feature vectors fail to span the entire space, the exact identification of the ground truth parameter becomes mathematically impossible. Thus, instead of pursuing the exact identification of the ground truth parameter, we shift our focus to analyzing the **suboptimality gap** of the policy induced by the learned parameters.

We begin by establishing a standard regularity condition on the feature representations within the action space. To ensure that the responses associated with a given prompt provide distinct feature signals, we assume the following:

Assumption 4. *For a given x and its corresponding response $\{a_i\}_{i=1}^{|\mathcal{A}|}$, $|\mathcal{A}| \leq d$[2], the set of feature vectors $\{\boldsymbol{\phi}(x,a_i) | i = 1, 2, \cdots, |\mathcal{A}|\}$ are linearly independent.*

Next, we introduce the core assumption. We characterize the feature vector $\boldsymbol{\phi}(x,a)$ by a direct sum decomposition, where part of it is from the column space of empirical covariance matrix of feature.

[2]We can relax the assumption on $\mathcal{A}$ to $\mathcal{A}_x$ depends on x, which does not influence our conclusions in sequel.

Assumption 5 (Feature Coverage). *For any prompt x, action a, and training set $\boldsymbol{S}$, the feature vector $\boldsymbol{\phi}(x,a)$ admits the following decomposition:*

$$\boldsymbol{\phi}(x,a) = \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{b}_{\boldsymbol{S}}(x,a) + \boldsymbol{r}_{\boldsymbol{S}}(x,a),$$

where the residual $\boldsymbol{r}_{\boldsymbol{S}}(x,a)$ satisfies $\sup_{x,a}\|\boldsymbol{r}_{\boldsymbol{S}}(x,a)\| \le \epsilon_n$ with $\epsilon_n \le \mathcal{O}(n^{-1})$[3] when $n \to \infty$.

Due to the fundamental theorem of linear algebra (Strang, 2012), the $\boldsymbol{r}_{\boldsymbol{S}}(x,a)$ can be located in $\mathrm{N}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))$ since $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{b}_{\boldsymbol{S}}(x,a) \in \mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))$ and $\mathbb{R}^d = \mathrm{N}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})) \oplus \mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))$.

This assumption implies that although the empirical covariance matrix $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$ fails to cover the entire feature space, its column space can approximately cover the feature space $\{\boldsymbol{\phi}(x,a) \mid (x,a) \in \mathcal{X}\times\mathcal{A}\}$ as the diversity of collected prompts expands with n, while the magnitude of the residual component is decreasing with sample size n. In fact, due to Lemma 18, $\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))$ is invariant over $\boldsymbol{\theta}$. Thus, it can be viewed as the feature accessed during the training process, which is required to approximately cover feature space $\{\boldsymbol{\phi}(x,a) | (x,a) \in \mathcal{X}\times\mathcal{A}\}$.

Building on the Assumption 5, we analyze the algorithmic stability $\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$ of the empirical stationary point $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ to get the generalization bound of $\boldsymbol{\theta}^*_{\boldsymbol{S}}$.

Theorem 2. *Under Assumptions 1, 2 and 5, we have*

$$\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}) \le \left(\frac{\Gamma_1}{n} + 2RC\sqrt{\frac{\Gamma_1}{n}}\right), \tag{11}$$

where $\Gamma_1 = \sup_{\boldsymbol{S},\boldsymbol{S}'}\left\{4nR\epsilon_n + \frac{8RC^3\cdot\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}))}\right\}$.

Then, for any $\rho \in (0,1)$, with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$,

$$\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| \le \mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\log(n)\log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right). \tag{12}$$

As detailed in Appendix E, we observe that under the key feature coverage Assumption 5, the generalization error bound is of order $\tilde{\mathcal{O}}(n^{-\frac{1}{2}})$ and the dominated constant factor Γ_1 depends on the "conditional number" $\frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}$ and the magnitude of the residual term $\boldsymbol{r}_{\boldsymbol{S}}$.

The estimation of the algorithmic stability coefficient hinges on Lemma 6, which explicitly bridges the pointwise objective difference with the empirical gradient norms. The core intuition is that both empirical stationary points, $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ and $\boldsymbol{\theta}^*_{\boldsymbol{S}'}$, exhibit negligible gradient norms on $J_{\boldsymbol{S}}$. Consequently, this smallness in gradient norms directly translates into a tight stability coefficient via Lemma 6.

Building upon the generalization error bound for $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ established in Theorem 2, we can now proceed to bound the suboptimality gap with respect to the ground truth $\boldsymbol{\theta}^*$.

Theorem 3. *Under Assumptions 1, 2, 5, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$,*

$$\begin{aligned}&|J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})| \\ &\le \left(\frac{\Gamma_1}{n} + 2RC\sqrt{\frac{\Gamma_1}{n}}\right)\log(n)\log\left(\frac{2}{\rho}\right) + 6RC\sqrt{\frac{\log(2/\rho)}{n}} + 3RC\sqrt{\frac{2\log(4/\rho)}{n}} = \tilde{\mathcal{O}}\left(n^{-\frac{1}{2}}\right).\end{aligned} \tag{13}$$

This result demonstrates that under the posterior Boltzmann policy structure, we achieve a dimension-free suboptimality gap of order $\tilde{\mathcal{O}}(n^{-\frac{1}{2}})$, even when the training data provides limited coverage of the feature space. The coefficient Γ_1 is defined as in the generalization-error analysis in Theorem 2.

The intuitive explanation for this dimension-independent bound is that the feature patterns learned by the model are determined by the subspace spanned by the empirical covariance matrix $\boldsymbol{V}_{\boldsymbol{S}}$. Under the feature coverage assumption, as n increases, this subspace progressively expands. Consequently, the feature patterns of unseen data are effectively covered by the patterns already captured in the training set, implying that the dimensions orthogonal to the covered subspace make negligible contributions to the generalization error, thereby decoupling the suboptimality gap from the dimension d.

[3]The exact $\mathcal{O}(n^{-1})$ decay is not essential to our analysis; a slower decay yields a correspondingly slower bound. We use the $\mathcal{O}(n^{-1})$ rate below for a cleaner theoretical result. See Appendix J.2 for details.

6 SUBOPTIMALITY GAP UNDER GRADIENT-BASED ALGORITHMS

In the previous section, we analyzed the suboptimality gap of the empirical stationary point $\boldsymbol{\theta}^*_{\boldsymbol{S}}$. However, in practice, we can only approximate $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ instead of exactly find it. The approximation process relies on the iterative gradient-based algorithms, such as Gradient Ascent (GA) or Stochastic Gradient Ascent (SGA). In this section, we extend our analysis to characterize the suboptimality gap of the policy learned by these gradient-based algorithms.

As clarified in Section 4, the suboptimality gap depends on generalization, optimization, and concentration error, the last of which is handled in Lemma 1. Thus it remains to bound the first two.

6.1 THE SUBOPTIMALITY GAP UNDER GA

We first analyze GA with the following update rule for each iteration t and the learning rate $\eta > 0$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \tag{14}$$

Definition 6 (GA algorithm output)**.** *Let T be the number of iterations, $\Theta_T = \{\boldsymbol{\theta}_t\}_{t=1}^T$ be the set of iterations obtained by* (14)*. We define $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ as: $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T} = \arg\min_{\boldsymbol{\theta}\in\Theta_T}\{\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\|\}$.*

We start with the optimization error. The standard result in non-convex optimization theory (e.g., Theorem 3.7 in Bubeck et al. (2015)) shows that under the L-smoothness condition proved in Lemma 12, $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ converges to the stationary point as in the following lemma.

Lemma 2. *By taking $\eta = \frac{1}{L_f}$ (L_f is defined in Lemma 12), the parameter $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ satisfies*

$$\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\|^2 \leq \frac{12 L_f R C}{T}. \tag{15}$$

Lemma 2 shows that GA approximates the empirical stationary point. As discussed in Section 4, small empirical gradient norm lead to small optimization error, as formalized by the following lemma.

Lemma 3. *Under Assumptions 1, 2 and 5, any $\boldsymbol{\theta}^*_{\boldsymbol{S}} \in \Theta_{\boldsymbol{S}}$ and $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ from Definition 6 satisfy:*

$$\left| J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) \right| = \mathcal{O}\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right). \tag{16}$$

The proof of Lemma 3 is provided in Section 4. This lemma shows that the optimization error of the parameter $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ obtained via GA is of order $\mathcal{O}(T^{-\frac{1}{4}} + n^{-\frac{1}{2}})$ and the constant factor depends on "conditional number" of $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$ and the magnitude of the residual term $\boldsymbol{r}_{\boldsymbol{S}}$, respectively.

Having quantified the optimization error for $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$, we now turn to its generalization error by adapting the algorithmic stability framework established for empirical stationary points in Section 5.2:

Theorem 4. *Under Assumptions 1, 2 and 5, for the GA output $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ defined in Definition 6, we have*

$$\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{GA}_{\boldsymbol{S},T}) \lesssim \sqrt{\epsilon_n} + \left(\sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}}} + \sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right)\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right),$$

where $\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'} = \sup_{\boldsymbol{S},\boldsymbol{S}'} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}))}$ *and* $\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}))}$.

Then, for any $\rho \in (0,1)$, with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$, it holds that:

$$\left| J(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) \right| \leq \mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{GA}_{\boldsymbol{S},T}) \log(n) \log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} = \tilde{\mathcal{O}}\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right). \tag{17}$$

The proof of Theorem 4 is in Appendix F. As can be seen, the generalization error of $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ is of order $\tilde{\mathcal{O}}(T^{-\frac{1}{4}} + n^{-\frac{1}{2}})$ with constant dependence similar to Theorem 2.

Interestingly, standard results in (Hardt et al., 2016) suggest that in the non-convex setting, more iterations leads to worse algorithmic stability, thereby worse generalization. In contrast, our analysis shows the algorithmic stability of $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ decreases with T. This phenomenon occurs because, for RLHF problem we explored, the critical Lemma 6 links the gradient norm $\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T})\|$ with algorithmic stability, implying that the small gradient norm of $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ leads to small algorithmic stability.

Thus, as discussed in Section 4, combining the generalization error bound (17), the optimization error bound (16), and the concentration error bound (7) yields the following suboptimality gap for $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$.

Theorem 5. *Under Assumptions 1, 2, and 5, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$:* $\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J(\pi_{\boldsymbol{\theta}^*})\right| = \tilde{\mathcal{O}}\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right)$.

By taking $T = n^2$, i.e., conducting the GA for n^2 steps, the suboptimality gap of $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ has the same order of empirical stationary points proved in Section 5.2.

6.2 The Suboptimality Gap under SGA

While GA provides theoretical clarity, computing full-batch gradients is computationally prohibitive in practice. A more realistic algorithm is SGA (Harvey et al., 2019), which has less computational complexity. The update rule of SGA is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(x_{i_t}), \tag{18}$$

where the index i_t is uniformly sampled from $\{1,\dots,n\}$ at iteration t, and η_t is the learning rate.

Definition 7 (SGA Algorithm Output)**.** *Let T be the number of iterations, $\Theta_T = \{\boldsymbol{\theta}_t\}_{t=1}^T$ be the set of iterations obtained by* (18)*. Similar to GA, we define $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$ as: $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T} = \arg\min_{\boldsymbol{\theta}\in\Theta_T}\{\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\|\}$.*

Remark 2. *Identifying $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$ as in Definition 7 requires evaluating the full gradient at each step, which incurs additional computational cost. However, it is standard in theoretical analysis of non-convex stochastic optimization to ensure optimal convergence rates (Ghadimi & Lan, 2013). While the last iterate is often used in practice, theoretical guarantees for it typically require stronger assumptions (e.g., the Polyak-Lojasiewicz condition), which are beyond the scope of this work.*

Similar to Lemma 2, the following lemma characterizes the convergence rate of SGA.

Lemma 4. *For any $\rho' \in (0,1)$, let $\eta_t = \frac{1}{2L_f\sqrt{t}}$, with L_f defined in Lemma 12, and let $\lambda > 0$ satisfies $\frac{e^\lambda - \lambda - 1}{\lambda} \le \frac{L_f}{8R^2C^4}$. Then, with probability at least $1-\rho'$, we have*

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}})\right\|^2 \le \frac{6L_f}{\sqrt{T}}\left(\left(6RC + \frac{2R^2C^4}{L_f}\right) + \frac{2R^2C^4}{L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{1}{\rho'}\right)\right). \tag{19}$$

SGA updates rely on gradient estimates with stochastic noise. Thus, unlike the deterministic convergence of GA in Section 6.1, Lemma 4 establishes a high-probability bound for $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$. Similar to Lemma 3, Lemma 4 implies the optimization error. Detailed proofs are in Appendix G.

Lemma 5. *Under Assumptions 1, 2 and 5, for any $\rho' \in (0,1)$, with probability at least $1 - \frac{\rho'}{2}$,*

$$\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| = \tilde{\mathcal{O}}\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right). \tag{20}$$

Next, we turn to the analysis of generalization error by applying algorithmic stability.

Theorem 6. *Under Assumptions 1, 2 and 5, for any $\rho' \in (0,1)$ and parameter $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$, we have*

$$\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}) \lesssim \sqrt{\epsilon_n} + \left(\sqrt{\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}}} + \sqrt{\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right)\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right),$$

where $\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}))}$ and $\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'} = \sup_{\boldsymbol{S},\boldsymbol{S}'} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}))}$.

Then, for any $\rho \in (0,1)$, with probability at least $1-\rho-\rho'$ over the randomness of $\boldsymbol{S}$ and i_t in SGA:

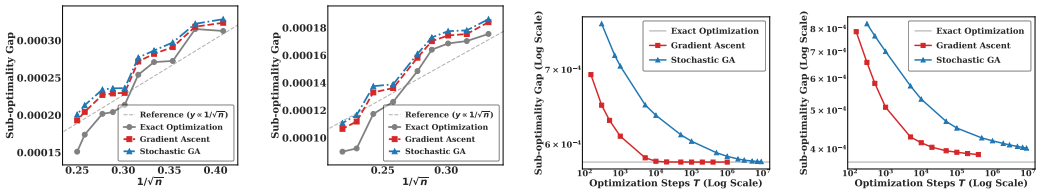
$$\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}})\right| \le \mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T})\log(n)\log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} = \tilde{\mathcal{O}}\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right). \tag{21}$$

The proof of Theorem 6 is provided in Appendix G. The proof strategy aligns with the analysis of GA in Section 6.1, with the difference in the bound's order stemming from the slower convergence rate of the stochastic algorithm. Finally, we prove the suboptimality gap of $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$.

Theorem 7. *Under Assumptions 1, 2 and 5, for any $\rho, \rho' \in (0,1)$, the following bound holds with probability at least $1-\rho-\rho'$ over the randomness of $\boldsymbol{S}$:* $\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J(\pi_{\boldsymbol{\theta}^*})\right| = \tilde{\mathcal{O}}\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right)$.

The proof proceeds according to the sub-optimality decomposition in Section 4. By taking $T = n^4$, the suboptimality gap of $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$ has the same order of empirical stationary points in Section 5.2.

Remark 3. *Notably, we adopt the high-probability algorithmic stability (Yuan & Li, 2023) mentioned in Appendix G to analyze SGA, since SGA is proved to be stable with high probability.*



(a) $d = 24$ (b) $d = 32$ (c) $n = 7$ (d) $n = 10$

Figure 1: The suboptimality gap under the feature coverage assumption for empirical stationary points and gradient-based iterates. **Left panels (1a-1b):** The suboptimality gap plotted against $n^{-\frac{1}{2}}$. **Right panels (1c-1d):** The suboptimality gap w.r.t. optimization steps T on a log-log scale.

7 Experiment

In this section, we present simulation experiments to empirically verify our theoretical findings.

7.1 Results for Positive Definite Covariance Matrix

We first verify the results under Assumption 3, i.e., the expected covariance matrix of feature is positive-definite. Our Theorem 1 indicates that under this assumption, the empirical stationary point is the ground-truth parameters.

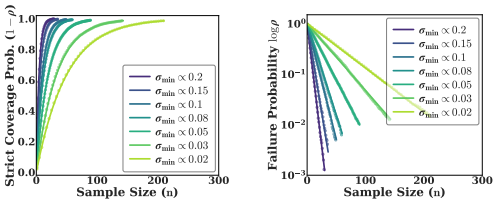


Figure 2: Number of samples required to recover ground-truth via empirical stationary points.

Lemma 11 admits that $\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) = \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta})$. Thus, when $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$ is positive-definite matrix for all $\boldsymbol{\theta}$, the empirical stationary point becomes the ground truth parameters. Thus, verifying the positive-definite property of $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$ is enough. The setup of our example refers to Appendix H.1.

Results: In Figure 2, we report the percentage of positive-definite matrices under varied $\sigma_{\min} = \min_{\boldsymbol{\theta}} \sigma_{\min}(\mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})])$. According to our conclusion in Theorem 1, the percentage of non-positive definite matrices $\rho \propto e^{-n}/\sigma_{\min}$. The relationship is verified in Figure 2.

7.2 The Results under Feature Coverage Assumption

Next, we empirically verify the results under our key feature coverage Assumption 5. Theorems 3, 5 and 7 show that under this assumption, the suboptimality gap is guaranteed to be small for both empirical stationary points and gradient-based iterates. Experimental details are in Appendix H.2.

Suboptimality Gap Scaling with sample size n. We verify the proved dimension-independent generalization bound of order $\mathcal{O}(n^{-\frac{1}{2}})$ in Theorem 2, 4 and 6. For problems in dimensions $d = 24, d = 32$, the suboptimality gap in Figure 1 verify the order of $\mathcal{O}(n^{-\frac{1}{2}})$.

Suboptimality Gap vs. steps T. On the other hand, our conclusions in Theorems 4 and 6 show that the generalization error decreases with the optimization steps, as verified in Figure 1.

8 Concluding Discussion

In this paper, we have established an end-to-end theoretical framework for RLHF based on algorithmic stability. Departing from existing paradigms grounded in the consistency of reward parameter estimation, we prove that under a key **feature coverage** condition, the empirical optima of the policy model have a suboptimality gap of order $\mathcal{O}(n^{-\frac{1}{2}})$. Furthermore, we extend our analysis to gradient-based algorithms, specifically Gradient Ascent and Stochastic Gradient Ascent, deriving explicit suboptimality bounds for the policies obtained by these algorithms. These results theoretically confirm that RLHF processes can learn generalizable policies solely from limited feedback, providing new theoretical evidence for the empirically observed success of RLHF.

AI USE STATEMENT

We used generative AI tools to assist with broad literature search the identification of potentially relevant prior work. We also used generative AI tools for language polishing of parts of the appendix to improve clarity and readability. All AI-assisted literature search results and identified references were manually reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Alekh Agarwal, Mikael Henaff, et al. PC-PG: Policy cover directed exploration for provable policy gradient learning. *Advances in Neural Information Processing Systems*, 2020.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Chenjia Bai, Yang Zhang, Shuang Qiu, Qiaosheng Zhang, Kang Xu, and Xuelong Li. Online preference alignment for language models via count-based exploration. Preprint arXiv:2501.12735, 2025.

Raef Bassily, Vitaly Feldman, et al. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 2020.

Viktor Bengs, Róbert Busa-Fekete, et al. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 2021.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, 2020.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

Jonathan D. Chang, Wenhao Zhan, et al. Dataset reset policy optimization for RLHF. Preprint arXiv:2404.08495, 2024.

Shreyas Chaudhari, Pranjal Aggarwal, et al. RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. *ACM Computing Surveys*, 2025.

Xiaoyu Chen, Han Zhong, et al. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, 2022.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.

Gheorghe Comanici, Eric Bieber, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint arXiv:2507.06261, 2025.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, 2015.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. Preprint arXiv:1909.00512, 2019.

Kawin Ethayarajh, Winnie Xu, et al. KTO: Model alignment as prospect theoretic optimization. Preprint arXiv:2402.01306, 2024.

Jianqing Fan, Zhaoran Wang, et al. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, 2020.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, 2019.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 1975.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 2013.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Preprint arXiv:2501.12948, 2025.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, 2019.

Physical Intelligence, Ali Amin, et al. $\pi^*_{0.6}$: a VLA that learns from experience. Preprint arXiv:2511.14759, 2025.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2018.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Yixin Liu, Kai Zhang, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. Preprint arXiv:2402.17177, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2023.

Benjamin Matthias Ruppik et al. Less is more: Local intrinsic dimensions of contextual language models. Preprint arXiv:2506.01034, 2025.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, et al. VLM-R1: A stable and generalizable R1-style large vision-language model. Preprint arXiv:2504.07615, 2025.

Yuda Song, Gokul Swamy, Aarti Singh, J Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. *Advances in Neural Information Processing Systems*, 2024.

Gilbert Strang. *Linear algebra and its applications*. 2012.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. 2019.

Mingze Wang and Chao Ma. Generalization error bounds for deep neural networks trained by SGD. Preprint arXiv:2206.03299, 2022.

Christian Wirth, Riad Akrour, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 2017.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. Preprint arXiv:2312.11456, 2023.

Shusheng Xu, Wei Fu, et al. Is DPO superior to PPO for LLM alignment? a comprehensive study. Preprint arXiv:2404.10719, 2024.

Yichong Xu, Ruosong Wang, et al. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 2020.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

An Yang, Anfeng Li, et al. Qwen3 technical report. Preprint arXiv:2505.09388, 2025.

Angen Ye, Zeyu Zhang, et al. VLA-R1: Enhancing reasoning in vision-language-action models. Preprint arXiv:2510.01623, 2025.

Mingyang Yi, Ruoyu Wang, and Zhi-Ming Ma. Characterization of excess risk for locally strongly convex population risk. *Advances in Neural Information Processing Systems*, 2022.

Xiaotong Yuan and Ping Li. l_2-uniform stability of randomized learning algorithms: Sharper generalization bounds and confidence boosting. *Advances in Neural Information Processing Systems*, 2023.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. Preprint arXiv:2305.14816, 2023.

Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: $O(1/n)$-and $O(1/n^2)$-type of risk bounds. In *Conference on Learning Theory*, 2017.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for KL-regularized contextual bandits and RLHF. Preprint arXiv:2411.04625, 2024.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, 2023.

Bowei Zhu, Shaojie Li, Mingyang Yi, and Yong Liu. Stability and sharper risk bounds with convergence rate $\tilde{O}(1/n^2)$. *Advances in Neural Information Processing Systems*, 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. Preprint arXiv:1909.08593, 2019.

Appendix

A One Critical Lemma **14**

B Some Regularity Properties on $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$, $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$ **17**

C Some Technical Lemmas **23**

D Proofs for Section 5.1 **25**

E Proofs for Section 5.2 **26**

F Proofs for Section 6.1 **27**

G Proofs for Section 6.2 **28**

H Detailed Experiments Setup **30**

H.1 The Results under Positive Definite Covariance Matrix 30

H.2 Generalization under Insufficient Coverage 31

I Experiments Extended to Real post-training **32**

I.1 Output Head-Only Post-Training 32

I.2 Post-Training with Different Data Coverage 33

J Empirical Validation of Assumptions 4 and 5 **34**

J.1 Discussion on Assumption 4 34

J.2 Discussion on Assumption 5 34

K Some Lemmas on Algebra. **34**

A One Critical Lemma

Firstly, we illustrate the critical lemma in this paper.

Lemma 6. *Under Assumptions 5, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_R$ and positive constants δ_1, δ_2, if we have*

$$\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_i})\| \leq \delta_i,\ i \in \{1, 2\},$$

then for any $x \in \mathcal{A}$, it holds that

$$\begin{aligned}|f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)| \leq & 4R\epsilon_n + \frac{2C\delta_1 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_1))} + \frac{2C\delta_2 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_2))} \\ & + 2CR\sqrt{4R\epsilon_n + \frac{2C\delta_1 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_1))} + \frac{2C\delta_2 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_2))}}.\end{aligned} \tag{22}$$

Before proving this lemma, we need the following lemmas.

Lemma 7. *Let $Z_{\boldsymbol{\theta}}(x) = \log \mathbb{E}_{\pi_{\text{ref}}(a|x)}[\exp(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(a,x)\rangle)]$, we have*

$$Z_{\boldsymbol{\theta}_1}(x) - Z_{\boldsymbol{\theta}_2}(x) = \left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}(a|x)}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle, \tag{23}$$

for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, with $\boldsymbol{\theta}_t = t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2$.

Proof. Define the function $h(t) = \mathbb{E}_{\pi_{\text{ref}}(a|x)}\left[\exp(\langle \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1), \boldsymbol{\phi}(a,x)\rangle)\right]$, which satisfies $\log h(0) = Z_{\boldsymbol{\theta}_1}(x)$ and $\log h(1) = Z_{\boldsymbol{\theta}_2}(x)$. Then we see

$$\begin{aligned}
\log h(1) - \log h(0) &= \int_0^1 \frac{h'(t)}{h(t)} dt \\
&= \int_0^1 \frac{\mathbb{E}_{\pi_{\text{ref}}(a|x)}\left[\exp(\langle \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1), \boldsymbol{\phi}(a,x)\rangle)\langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \boldsymbol{\phi}(a,x)\rangle\right]}{\mathbb{E}_{\pi_{\text{ref}}(a|x)}\left[\exp(\langle \boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1), \boldsymbol{\phi}(a,x)\rangle)\right]} \\
&= \left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}(a|x)}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle,
\end{aligned} \tag{24}$$

which proves our conclusion. □

Lemma 8. *Under Assumption 1, for a given prompt x and any two parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_R$. We have*

$$\begin{aligned}
|f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)| \leq & 2CR\sqrt{\left\langle \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}} \boldsymbol{\phi}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle + \left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle} \\
& + \left| \left\langle \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}} \boldsymbol{\phi}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \right| + \left| \left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle \right|.
\end{aligned} \tag{25}$$

Proof. According to the definition of the function $f_{\boldsymbol{\theta}}(x)$ in (1) and the statement below, we have

$$\begin{aligned}
|f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)| = & \left| \sum_{a \in \mathcal{A}} r(a,x) \left(\pi_{\boldsymbol{\theta}_1}(a \mid x) - \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) \right. \\
& \left. -\lambda \sum_{a \in \mathcal{A}} \left(\pi_{\boldsymbol{\theta}_1}(a \mid x) \log \frac{\pi_{\boldsymbol{\theta}_1}(a \mid x)}{\pi_{\text{ref}}(a \mid x)} - \pi_{\boldsymbol{\theta}_2}(a \mid x) \log \frac{\pi_{\boldsymbol{\theta}_2}(a \mid x)}{\pi_{\text{ref}}(a \mid x)} \right) \right|.
\end{aligned} \tag{26}$$

where the hyper-parameter $\lambda = 1$. Let us tackle the two terms respectively. For the first, by Pinsker's inequality, we have

$$\begin{aligned}
\left| \sum_{a \in \mathcal{A}} r(a,x) \left(\pi_{\boldsymbol{\theta}_1}(a \mid x) - \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) \right| &\leq \max_{a \in \mathcal{A}} r(a,x) \left| \sum_{a \in \mathcal{A}} \left(\pi_{\boldsymbol{\theta}_1}(a \mid x) - \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) \right| \\
&= \max_{a \in \mathcal{A}} \langle \boldsymbol{\phi}(x,a), \boldsymbol{\theta}^* \rangle \left| \sum_{a \in \mathcal{A}} \left(\pi_{\boldsymbol{\theta}_1}(a \mid x) - \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) \right| \\
&\leq \|\boldsymbol{\theta}^*\| \max_{a \in \mathcal{A}} \|\boldsymbol{\phi}(x,a)\| \left| \sum_{a \in \mathcal{A}} \left(\pi_{\boldsymbol{\theta}_1}(a \mid x) - \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) \right| \\
&\leq CR \cdot \mathsf{TV}\left(\pi_{\boldsymbol{\theta}_1}(a \mid x), \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) \\
&\leq CR\sqrt{\frac{1}{2} D_{KL}(\pi_{\boldsymbol{\theta}_1}(a \mid x) \,\|\, \pi_{\boldsymbol{\theta}_2}(a \mid x))}.
\end{aligned} \tag{27}$$

For the KL divergence term, we have

$$\begin{aligned}
D_{KL}\left(\pi_{\boldsymbol{\theta}_1}(a \mid x) \,\|\, \pi_{\boldsymbol{\theta}_2}(a \mid x)\right) &= \sum_{a \in \mathcal{A}} \left(\langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \boldsymbol{\phi}(x,a)\rangle\right) \pi_{\boldsymbol{\theta}_1}(a \mid x) \\
&\quad + \log \mathbb{E}_{\pi_{\text{ref}}(a|x)}[\exp\left(\langle \boldsymbol{\theta}_1, \boldsymbol{\phi}(a,x)\rangle\right)] - \log \mathbb{E}_{\pi_{\text{ref}}(a|x)}[\exp\left(\langle \boldsymbol{\theta}_2, \boldsymbol{\phi}(a,x)\rangle\right)] \\
&= \left\langle \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)}[\boldsymbol{\phi}(x,a)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle + Z_{\boldsymbol{\theta}_1}(x) - Z_{\boldsymbol{\theta}_2}(x) \\
&= \left\langle \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)}[\boldsymbol{\phi}(x,a)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle + \left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}(a|x)}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \right\rangle.
\end{aligned} \tag{28}$$

where the last equality is from Lemma 7, and $\boldsymbol{\theta}_t$ is an interpolation between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ as in Lemma 7. On the other hand, let us upper bound the second term in (26). By basic algebra, we have:

$$\begin{aligned}
&\left|\sum_{a\in\mathcal{A}}\left(\pi_{\boldsymbol{\theta}_1}(a\mid x)\log\frac{\pi_{\boldsymbol{\theta}_1}(a\mid x)}{\pi_{\mathrm{ref}}(a\mid x)}-\pi_{\boldsymbol{\theta}_2}(a\mid x)\log\frac{\pi_{\boldsymbol{\theta}_2}(a\mid x)}{\pi_{\mathrm{ref}}(a\mid x)}\right)\right| \\
&\le\left|\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)}[\langle\boldsymbol{\theta}_1,\boldsymbol{\phi}(x,a)\rangle]-\mathbb{E}_{\pi_{\boldsymbol{\theta}_2}(a|x)}[\langle\boldsymbol{\theta}_2,\boldsymbol{\phi}(x,a)\rangle]\right|+|Z_{\boldsymbol{\theta}_2}(x)-Z_{\boldsymbol{\theta}_1}(x)| \\
&\le\left|\langle\boldsymbol{\theta}_1-\boldsymbol{\theta}_2,\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}}[\boldsymbol{\phi}(x,a)]\rangle\right|+\left|\left\langle\boldsymbol{\theta}_2,\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}}[\boldsymbol{\phi}(x,a)]-\mathbb{E}_{\pi_{\boldsymbol{\theta}_2}}[\boldsymbol{\phi}(x,a)]\right\rangle\right|+\left|\left\langle\int_0^1\mathbb{E}_{\pi_{\boldsymbol{\theta}_t}(a|x)}[\boldsymbol{\phi}(a,x)],\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\rangle\right| \\
&\le\left|\left\langle\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)}\boldsymbol{\phi}(x,a),\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\rangle\right|+\left|\left\langle\int_0^1\mathbb{E}_{\pi_{\boldsymbol{\theta}_t}(a|x)}[\boldsymbol{\phi}(a,x)],\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\rangle\right|+\left|\left\langle\boldsymbol{\theta}_2,\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}}[\boldsymbol{\phi}(x,a)]-\mathbb{E}_{\pi_{\boldsymbol{\theta}_2}}[\boldsymbol{\phi}(x,a)]\right\rangle\right|.
\end{aligned}\tag{29}$$

For the last term above, by definition, we have

$$\begin{aligned}
\left|\left\langle\boldsymbol{\theta}_2,\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}}[\boldsymbol{\phi}(x,a)]-\mathbb{E}_{\pi_{\boldsymbol{\theta}_2}}[\boldsymbol{\phi}(x,a)]\right\rangle\right| &\le\|\boldsymbol{\theta}_2\|\max_{a\in\mathcal{A}}\|\boldsymbol{\phi}(x,a)\|\mathsf{TV}\left(\pi_{\boldsymbol{\theta}_1}(a\mid x),\pi_{\boldsymbol{\theta}_2}(a\mid x)\right) \\
&\le CR\sqrt{\frac{1}{2}D_{KL}(\pi_{\boldsymbol{\theta}_1}(a\mid x)\parallel\pi_{\boldsymbol{\theta}_2}(a\mid x))} \\
&\le CR\sqrt{\left\langle\mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)}[\boldsymbol{\phi}(x,a)],\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\rangle+\left\langle\int_0^1\mathbb{E}_{\pi_{\boldsymbol{\theta}_t}(a|x)}[\boldsymbol{\phi}(a,x)],\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\right\rangle}.
\end{aligned}\tag{30}$$

Then by combining (26)-(30), we obtain our conclusion. □

Lemma 9. *For a given $\boldsymbol{\theta}$ and positive constant δ, if*

$$\|\nabla_{\boldsymbol{\theta}}J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\|\le\delta,$$

then we have

$$\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*)(\boldsymbol{\theta}_{\boldsymbol{S}}^*-\boldsymbol{\theta})\|=\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*)(\boldsymbol{\theta}^*-\boldsymbol{\theta})\|\le\frac{\delta\cdot\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*))}{\sigma_{\min}^+(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))}.\tag{31}$$

Proof. Due to the definition of the parameter $\boldsymbol{\theta}_{\boldsymbol{S}}^*$ in (5), we have

$$0=\nabla_{\boldsymbol{\theta}}J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S}}^*})=\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*)(\boldsymbol{\theta}^*-\boldsymbol{\theta}_{\boldsymbol{S}}^*),$$

it is straightforward to verify the first equality holds. According to the fundamental theorem of linear algebra, it holds that

$$\mathbb{R}^d=\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))+\mathrm{N}(\boldsymbol{V}_{\boldsymbol{S}}^\top(\boldsymbol{\theta})),$$

then we decompose the vector $\boldsymbol{\theta}^*-\boldsymbol{\theta}$ into a sum of orthogonal components,

$$\boldsymbol{\theta}^*-\boldsymbol{\theta}=\boldsymbol{\theta}^C+\boldsymbol{\theta}^N,$$

where $\boldsymbol{\theta}^C\in\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*))$ and $\boldsymbol{\theta}^N\in\mathrm{N}(\boldsymbol{V}_{\boldsymbol{S}}^\top(\boldsymbol{\theta}_{\boldsymbol{S}}^*))$. According to Lemma 18, we have

$$\nabla_{\boldsymbol{\theta}}J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})=\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})(\boldsymbol{\theta}^*-\boldsymbol{\theta})=\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})\cdot\boldsymbol{\theta}^C,$$

Firstly, We decompose the vector $\boldsymbol{\theta}^C$ as a linear combination $\boldsymbol{\theta}^C=\sum_{i=1}^m c_i\boldsymbol{\xi}_i$, where $\{\boldsymbol{\xi}_i\}_{i=1}^m$ is the set of eigenvectors of the matrix $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$ corresponding to its positive eigenvalues $\{\sigma_i\}_{i=1}^m$ ($m\le d$). Then, it holds that:

$$\left\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})\cdot\boldsymbol{\theta}^C\right\|=\left\|\sum_{i=1}^m c_i\sigma_i\boldsymbol{\xi}_i\right\|=\sqrt{\sum_{i=1}^m c_i^2\sigma_i^2}\ge\sqrt{\min_{\sigma\in\{\sigma_j\}_{j=1}^m}\sum_{i=1}^m c_i^2\sigma^2}=\min_{1\le i\le m}\sigma_i(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))\|\boldsymbol{\theta}^C\|.$$

Let $\sigma_{\min}^+(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))$ be the smallest positive eigenvalue of the matrix $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$, then we can bound the $\|\boldsymbol{\theta}^C\|$ as follows:

$$\|\boldsymbol{\theta}^C\|\le\frac{\|\nabla_{\boldsymbol{\theta}}J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\|}{\sigma_{\min}^+(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))}\le\frac{\delta}{\sigma_{\min}^+(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))},$$

according to the lemma 18, it holds that $\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*))=\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))$, therefore we can calculate that

$$\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*)(\boldsymbol{\theta}^*-\boldsymbol{\theta})\|=\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*)\boldsymbol{\theta}^C\|\le\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*))\cdot\|\boldsymbol{\theta}^C\|\le\frac{\delta\cdot\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*))}{\sigma_{\min}^+(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}))}.\tag{32}$$

□

Lemma 10. *Under Assumption 1 and 5, for any feature vector $\boldsymbol{\phi}(x,a)$, there exists a proposed $\boldsymbol{b}_{\boldsymbol{S}}(x,a)$ in Assumption 5 satisfies $\|\boldsymbol{b}_{\boldsymbol{S}}(x,a)\| \leq \max_{\sigma_i(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))>0} \frac{C}{\sigma_i(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}$, where $\{\sigma_i(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\}$ are eigenvalues of $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$.*

Proof. The Assumption 5 supposes that $\boldsymbol{b}_{\boldsymbol{S}}(x,a) \in \mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))$. And $\boldsymbol{V}_{\boldsymbol{S}}$ is a real symmetric matrix, which indicates that $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{P} = \boldsymbol{P}\Lambda$, where the column vectors of $\boldsymbol{P}$ are eigenvectors and $\Lambda = \mathrm{diag}\{\sigma_1, \cdots, \sigma_d\}$ are eigenvalues. Since

$$\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{x} = \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{P}\boldsymbol{P}^{-1}\boldsymbol{x} = \boldsymbol{P}(\Lambda\boldsymbol{P}^{-1}\boldsymbol{x}),$$

we know that $\boldsymbol{b}_{\boldsymbol{S}}(x,a)$ can be expressed as $\boldsymbol{b}_{\boldsymbol{S}}(x,a) = \sum_{i=1}^m c_i(x,a)\boldsymbol{\xi}_i$, with $\{\boldsymbol{\xi}\}_{i=1}^m$ are eigenvectors of $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})$ corresponded to non-negative eigenvalues σ_i. Thus, from one-side, we have

$$\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{b}_{\boldsymbol{S}}(x,a)\| = \|\boldsymbol{\phi}(x,a) - \boldsymbol{r}_{\boldsymbol{S}}(x,a)\| \leq C, \tag{33}$$

since the residual term $\boldsymbol{r}_{\boldsymbol{S}}(x,a)$ is orthogonal to $\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))$. On the other hand,

$$\begin{aligned}
\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\boldsymbol{b}_{\boldsymbol{S}}(x,a)\| &= \left\|\sum_{i=1}^m c_i(x,a)\sigma_i\boldsymbol{\xi}_i\right\| \\
&= \sqrt{\sum_{i=1}^m c_i^2(x,a)\sigma_i^2(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))} \\
&\geq \min_{1\leq i\leq m}\sigma_i(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\|\boldsymbol{c}(x,a)\| \\
&= \min_{1\leq i\leq m}\sigma_i(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\|\boldsymbol{b}_{\boldsymbol{S}}\|.
\end{aligned} \tag{34}$$

Combining it with the above inequality proves our conclusion. □

Now we are ready to prove Lemma 6.

Proof. According to Lemma 8, we have

$$\begin{aligned}
|f_{\boldsymbol{\theta}_1}(x) - f_{\boldsymbol{\theta}_2}(x)| \leq & 2C\sqrt{\left\langle \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}}\boldsymbol{\phi}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\rangle + \left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\rangle} \\
& + \left|\left\langle \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}}\boldsymbol{\phi}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\rangle\right| + \left|\left\langle \int_0^1 \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}}[\boldsymbol{\phi}(a,x)], \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\rangle\right|.
\end{aligned} \tag{35}$$

Under the assumption 5, by using Lemma 9, 10, it holds that

$$\begin{aligned}
\|\langle\boldsymbol{\phi}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle\| &= \|\langle\boldsymbol{b}_{\boldsymbol{S}}(x,a), \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\rangle + \langle\boldsymbol{r}_{\boldsymbol{S}}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle\| \\
&= \|\langle\boldsymbol{b}_{\boldsymbol{S}}(x,a), \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)\rangle + \langle\boldsymbol{b}_{\boldsymbol{S}}(x,a), \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})(\boldsymbol{\theta}^* - \boldsymbol{\theta}_2)\rangle + \langle\boldsymbol{r}_{\boldsymbol{S}}(x,a), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle\| \\
&\leq 2R\epsilon_n + \|\boldsymbol{b}_{\boldsymbol{S}}(x,a)\| \cdot (\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*)\| + \|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})(\boldsymbol{\theta}_2 - \boldsymbol{\theta}^*)\|) \\
&\leq 2R\epsilon_n + \frac{C\delta_1 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_1))} + \frac{C\delta_2 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_2))}.
\end{aligned} \tag{36}$$

Combining (36) with (35), we complete the proof. □

B Some Regularity Properties on $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$, $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$

Proposition 3. *When $\pi_{\boldsymbol{\theta}}$ is from Definition 2, we have*

$$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}} J(\pi_{\boldsymbol{\theta}}), \tag{37}$$

and

$$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}). \tag{38}$$

Proof.

$$\begin{aligned} J(\pi_{\boldsymbol{\theta}}) =& \mathbb{E}_x \left[\mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|x)}[r(x,a)] - D_{KL}\left(\pi_{\boldsymbol{\theta}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)\right)\right] \\ =& \mathbb{E}_x \left[\mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|x)} \left[r(x,a) - \log \frac{\pi_{\boldsymbol{\theta}}(a|x)}{\pi_{\mathrm{ref}}(a|x)} \right] \right] \\ =& -\mathbb{E}_x \left[\mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|x)} \left[\log \frac{\pi_{\boldsymbol{\theta}}(a|x)}{\pi_{\mathrm{ref}}(a|x)\exp(r(x,a))} \right] \right] \\ =& -\mathbb{E}_x \left[\mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|x)} \left[\log \frac{\pi_{\boldsymbol{\theta}}(a|x)\cdot Z(x)}{\pi_{\mathrm{ref}}(a|x)\exp(r(x,a))} \right] - \log Z(x) \right] \\ =& \mathbb{E}_x[\log Z(x)] - \mathbb{E}_x[D_{KL}(\pi_{\boldsymbol{\theta}}(\cdot|x)\|\pi^*(\cdot|x))] \\ \leq& \mathbb{E}_x[\log Z(x)], \end{aligned} \tag{39}$$

where $Z(x) = \mathbb{E}_{a\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(r(x,a))]$ and $\pi^*(a|x) = \frac{\pi_{\mathrm{ref}}(a|x)\cdot\exp(r(x,a))}{Z(x)}$, for any $(x,a) \in \mathcal{X}\times\mathcal{A}$. The last inequality holds for $D_{KL}(\pi_{\boldsymbol{\theta}}(\cdot|x)\|\pi^*(\cdot|x)) \geq 0$, and the equality holds if and only if $\pi_{\boldsymbol{\theta}}(a|x) \equiv \pi^*(a|x)$. The definition of $Z(x)$ indicates that $\mathbb{E}_x[\log Z(x)]$ is a constant independent of $\boldsymbol{\theta}$.

According to the definition of $\boldsymbol{\theta}^*$, it holds that $\pi_{\boldsymbol{\theta}^*}(a|x) \equiv \pi^*(a|x)$, thus we prove the (37). Similarly, we can prove the conclusion in (38). □

This proposition shows that the ground truth parameter $\boldsymbol{\theta}^*$ locates in the optimal solutions of $J(\pi_{\boldsymbol{\theta}})$ and $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$.

Lemma 11. *Under Assumption 2, it holds that*

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) = \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})(\boldsymbol{\theta}^* - \boldsymbol{\theta}), \tag{40}$$

where $\boldsymbol{V}_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \mathsf{Var}_{\pi_{\boldsymbol{\theta}}(a|x_i)}[\boldsymbol{\phi}(x_i,a)]$.

Proof. Firstly, let us check the gradient of loss function $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$, which has the following formulation

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\pi_{\boldsymbol{\theta}}(a|x_i)} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a \mid x_i) \left(r(x_i,a) - \langle \boldsymbol{\theta}, \boldsymbol{\phi}(x_i,a)\rangle \right)\right]. \tag{41}$$

Due to the formulation of $\pi_{\boldsymbol{\theta}}$, we know that

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a \mid x) = \boldsymbol{\phi}(x,a) - \mathbb{E}_{\pi_{\boldsymbol{\theta}}(a|x)}[\boldsymbol{\phi}(x,a)]. \tag{42}$$

Then, due to Assumption 2 and combining the above equations, we obtain our conclusion. □

Proposition 4. *Under the defined linear reward parameterization and Posterior Boltzmann policy class* Π*, any parameter* $\boldsymbol{\theta}$ *satisfying the first-order stationarity condition* $\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) = 0$ *is a global maximizer of the empirical objective* $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$*. In other words, for any* $\boldsymbol{\theta}' \in \mathbb{R}^d$*, we have*

$$\boldsymbol{\theta}' \in \arg\max_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) \quad \Longleftrightarrow \quad \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}'}) = 0.$$

Proof. The necessity is straightforward. We now focus only on the proof of sufficiency. Observe that $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$. According to Proposition 11, we know

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}'}) = \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}')(\boldsymbol{\theta}^* - \boldsymbol{\theta}') = 0.$$

Then we have

$$\begin{aligned} 0 &= (\boldsymbol{\theta}^* - \boldsymbol{\theta}')^\top \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}')(\boldsymbol{\theta}^* - \boldsymbol{\theta}') \\ &= \frac{1}{n}\sum_{i=1}^n (\boldsymbol{\theta}^* - \boldsymbol{\theta}')^\top \mathbb{E}_{\pi_{\boldsymbol{\theta}'}} \left[\left(\boldsymbol{\phi}(x_i,a) - \mathbb{E}_{\pi_{\boldsymbol{\theta}'}}[\boldsymbol{\phi}(x_i,a')]\right) \left(\boldsymbol{\phi}(x_i,a) - \mathbb{E}_{\pi_{\boldsymbol{\theta}'}}[\boldsymbol{\phi}(x_i,a')]\right)^\top \right] (\boldsymbol{\theta}^* - \boldsymbol{\theta}') \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\pi_{\boldsymbol{\theta}'}} \left[\left((\boldsymbol{\theta}^* - \boldsymbol{\theta}')^\top \left(\boldsymbol{\phi}(x_i,a) - \mathbb{E}_{\pi_{\boldsymbol{\theta}'}}[\boldsymbol{\phi}(x_i,a')]\right) \right)^2 \right]. \end{aligned}$$

Therefore, for any $x_i \in \boldsymbol{S}$ and a, we have

$$(\boldsymbol{\theta}^* - \boldsymbol{\theta}')^\top \boldsymbol{\phi}(x_i,a) = (\boldsymbol{\theta}^* - \boldsymbol{\theta}')^\top \mathbb{E}_{\pi_{\boldsymbol{\theta}'}}[\boldsymbol{\phi}(x_i,a')],$$

which means that

$$\begin{aligned}\frac{\pi_{\boldsymbol{\theta}'}(a|x_i)}{\pi_{\text{ref}}(a|x_i)} &= \frac{\exp(\lambda^{-1}\langle\boldsymbol{\theta}', \boldsymbol{\phi}(x_i,a)\rangle)}{\mathbb{E}_{\pi_{\text{ref}}}\left[\exp(\lambda^{-1}\langle\boldsymbol{\theta}', \boldsymbol{\phi}(x_i,a)\rangle)\right]} \\ &= \frac{\exp(\lambda^{-1}\langle\boldsymbol{\theta}' - \boldsymbol{\theta}^*, \boldsymbol{\phi}(x_i,a)\rangle) \cdot \exp(\lambda^{-1}\langle\boldsymbol{\theta}^*, \boldsymbol{\phi}(x_i,a)\rangle)}{\exp\left(\lambda^{-1}\langle\boldsymbol{\theta}' - \boldsymbol{\theta}^*, \boldsymbol{\phi}(x_i,a)\rangle\right) \cdot \mathbb{E}_{\pi_{\text{ref}}}\left[\exp(\lambda^{-1}\langle\boldsymbol{\theta}^*, \boldsymbol{\phi}(x_i,a)\rangle)\right]} \\ &= \frac{\pi_{\boldsymbol{\theta}^*}(a|x_i)}{\pi_{\text{ref}}(a|x_i)}.\end{aligned}$$

Therefore, it holds that

$$\pi_{\boldsymbol{\theta}'}(a|x_i) = \pi_{\boldsymbol{\theta}^*}(a|x_i), \forall (x_i, a) \in \boldsymbol{S} \times \mathcal{A}. \tag{43}$$

Combining (43) with (4), we have

$$\begin{aligned}J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}'}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}'}(\cdot|x_i)}\left[r(x_i,a) - \log\frac{\pi_{\boldsymbol{\theta}'}(a|x_i)}{\pi_{\text{ref}}(a|x_i)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a} \pi_{\boldsymbol{\theta}'}(a|x_i)\left[r(x_i,a) - \log\frac{\pi_{\boldsymbol{\theta}'}(a|x_i)}{\pi_{\text{ref}}(a|x_i)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a} \pi_{\boldsymbol{\theta}^*}(a|x_i)\left[r(x_i,a) - \log\frac{\pi_{\boldsymbol{\theta}^*}(a|x_i)}{\pi_{\text{ref}}(a|x_i)}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{a\sim\pi_{\boldsymbol{\theta}^*}(\cdot|x_i)}\left[r(x_i,a) - \log\frac{\pi_{\boldsymbol{\theta}^*}(a|x_i)}{\pi_{\text{ref}}(a|x_i)}\right] \\ &= J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*}).\end{aligned}$$

By combining with Proposition 3, we prove the conclusion. □

Remark 4. *This proposition connects our definition of $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ in* (5) *to global optima. It implies that finding a stationary point is equivalent to finding the empirical global optimum, i.e., $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) \geq J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$ for any $\boldsymbol{\theta} \in \mathbb{R}^d$.*

Next, we introduce two lemmas, which state the smoothness of $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$ w.r.t. $\boldsymbol{\theta}$.

Lemma 12. *Under Assumptions 1 and 2, the empirical risk $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$ is L_f-smoothness with coefficient $8RC^3 + C^2$ within the region Θ_R. That says,*

$$\|\nabla^2_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\| \leq 8RC^3 + C^2. \tag{44}$$

Proof. To simplify the notations, we define $\Sigma_x(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}(a|x)}[\boldsymbol{\phi}(x,a)\boldsymbol{\phi}(x,a)^\top]$, and $\boldsymbol{\mu}_x(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}(a|x)}[\boldsymbol{\phi}(x,a)]$. Then we know

$$\begin{aligned}\left\|\nabla^2 J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\right\| &= \frac{1}{n}\left\|\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}(\Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})\boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top)(\boldsymbol{\theta}^* - \boldsymbol{\theta}) - \boldsymbol{V}_{x_i}(\boldsymbol{\theta})\right\| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left(\left\|\nabla_{\boldsymbol{\theta}}(\Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})\boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top)(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right\| + \|V_{x_i}(\boldsymbol{\theta})\|\right).\end{aligned} \tag{45}$$

For the first term, we denote $g_i(\boldsymbol{u}, \boldsymbol{\theta}) = \boldsymbol{u}^\top(\Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})\boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top)\boldsymbol{u}$, then for any x_i we have

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{\theta}}(\Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})\boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top)(\boldsymbol{\theta}^* - \boldsymbol{\theta})\right\| \\ =& \max_{\boldsymbol{u}\in\mathbb{R}^d, \|\boldsymbol{u}\|\leq 1} \boldsymbol{u}^\top \nabla_{\boldsymbol{\theta}}(\Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})\boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top)(\boldsymbol{\theta}^* - \boldsymbol{\theta})\boldsymbol{u} \\ =& \max_{\boldsymbol{u}\in\mathbb{R}^d, \|\boldsymbol{u}\|\leq 1} \left\langle \nabla_{\boldsymbol{\theta}} g_i(\boldsymbol{u}, \boldsymbol{\theta}), \boldsymbol{\theta}^* - \boldsymbol{\theta}\right\rangle.\end{aligned} \tag{46}$$

By applying lemma(42), we have

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}} g_i(\boldsymbol{u}, \boldsymbol{\theta}) \\
=&\nabla_{\boldsymbol{\theta}} \left(\mathbb{E}_{\pi_{\boldsymbol{\theta}}} [\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a)]^2 - \mathbb{E}^2_{\pi_{\boldsymbol{\theta}}} [\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a)] \right) \\
=&\mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|x_i) \cdot \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right)^2 \right] - 2\mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|x_i) \cdot \boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \\
=&\mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\left(\boldsymbol{\phi}(x_i, a) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \right) \cdot \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right)^2 \right] - 2\mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\left(\boldsymbol{\phi}(x_i, a) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \right) \cdot \boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \\
=&\mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{\phi}(x_i, a) \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right)^2 \right] - 2\mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{\phi}(x_i, a) \boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \\
&- \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \left(\mathsf{Var}_{\pi_{\boldsymbol{\theta}}} \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right) - \mathbb{E}^2_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \right).
\end{aligned} \tag{47}$$

Substituting (47) into (46), we obtain

$$\begin{aligned}
&\left\| \nabla_{\boldsymbol{\theta}} (\Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top)(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \right\| \\
\leq& \max_{\boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\| \leq 1} \|\nabla_{\boldsymbol{\theta}} g_i(\boldsymbol{u}, \boldsymbol{\theta})\| \cdot \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| \\
\leq& 2R \left\| \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{\phi}(x_i, a) \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right)^2 \right] \right\| + 4R \left\| \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{\phi}(x_i, a) \boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \right\| \\
&+ 2R \left\| \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \left(\mathsf{Var}_{\pi_{\boldsymbol{\theta}}} \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right) - R\mathbb{E}^2_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \right) \right\| \\
\leq& 2RC^3 + 4RC^3 + 2RC \max \left\{ \mathsf{Var}_{\pi_{\boldsymbol{\theta}}} \left(\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right), \mathbb{E}^2_{\pi_{\boldsymbol{\theta}}} \left[\boldsymbol{u}^\top \boldsymbol{\phi}(x_i, a) \right] \right\} \\
\leq& 8RC^3.
\end{aligned} \tag{48}$$

For the second term in (45), we have

$$\|\boldsymbol{V}_{x_i}(\boldsymbol{\theta})\| = \left\| \Sigma_{x_i}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top \right\| \leq \max \left\{ \|\Sigma_{x_i}(\boldsymbol{\theta})\|, \left\| \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta}) \boldsymbol{\mu}_{x_i}(\boldsymbol{\theta})^\top \right\| \right\} \leq C^2. \tag{49}$$

Then, by combining (45), (48) and (49), we can prove that the Hessian matrix has all eigenvalues with absolute values smaller than $8RC^3 + C^2$. □

Lemma 13. *Let $\boldsymbol{V}_x(\boldsymbol{\theta}) = \mathsf{Var}_{\pi_{\boldsymbol{\theta}}(a|x)}[\boldsymbol{\phi}(a, x)]$, it satisfies the Lipschitz continuous condition*

$$\|\boldsymbol{V}_x(\boldsymbol{\theta}_1) - \boldsymbol{V}_x(\boldsymbol{\theta}_2)\| \leq 3C^{\frac{5}{2}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\frac{1}{2}}, \tag{50}$$

for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$ and x.

Proof. Due to the definition, we have

$$\begin{aligned}
&\|\boldsymbol{V}_x(\boldsymbol{\theta}_1) - \boldsymbol{V}_x(\boldsymbol{\theta}_2)\| \\
=& \sup_{\boldsymbol{u}: \|\boldsymbol{u}\| = 1} \boldsymbol{u}^\top (\boldsymbol{V}_x(\boldsymbol{\theta}_1) - \boldsymbol{V}_x(\boldsymbol{\theta}_2)) \boldsymbol{u} \\
\leq& \left| \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)} \left[(\boldsymbol{u}^\top \boldsymbol{\phi}(a, x))^2 \right] - \mathbb{E}_{\pi_{\boldsymbol{\theta}_2}(a|x)} \left[(\boldsymbol{u}^\top \boldsymbol{\phi}(a, x))^2 \right] \right| \\
&+ \left| \left(\boldsymbol{u}^\top \mathbb{E}_{\pi_{\boldsymbol{\theta}_1}(a|x)} [\boldsymbol{\phi}(a, x)] \right)^2 - \left(\boldsymbol{u}^\top \mathbb{E}_{\pi_{\boldsymbol{\theta}_2}(a|x)} [\boldsymbol{\phi}(a, x)] \right)^2 \right| \\
\leq& C^2 \mathsf{TV}(\pi_{\boldsymbol{\theta}_1}(a \mid x), \pi_{\boldsymbol{\theta}_2}(a \mid x)) + 2C^2 \mathsf{TV}(\pi_{\boldsymbol{\theta}_1}(a \mid x), \pi_{\boldsymbol{\theta}_2}(a \mid x)) \\
\leq& 3C^{\frac{5}{2}} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^{\frac{1}{2}},
\end{aligned} \tag{51}$$

which proves our conclusion. Here the last inequality is based on the proof of Lemma 8. □

The following lemma shows that the loss function $f_{\boldsymbol{\theta}}(x)$ is bounded.

Lemma 14. *Under Assumption 1, for any $\boldsymbol{\theta} \in \Theta_R$ and x, we have $|f_{\boldsymbol{\theta}}(x)| \leq 3RC$, which implies $|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})| \leq 3RC$ as well.*

Proof. According to the definition of $f_{\boldsymbol{\theta}}(x)$ in (1), we have:

$$\begin{aligned}
|f_{\boldsymbol{\theta}}(x)| &\leq \max_x \left| \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[r(x,a) - \log \frac{\pi_{\boldsymbol{\theta}}(a|x)}{\pi_{\text{ref}}(a|x)} \right] \right| \\
&= \max_x \left| \sum_{a \in \mathcal{A}} \pi_{\boldsymbol{\theta}}(a|x) \left(\langle \boldsymbol{\theta}^*, \boldsymbol{\phi}(x,a) \rangle - \log \frac{\exp(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x,a) \rangle)}{\sum_{a'} \pi_{\text{ref}}(a'|x) \exp(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x,a) \rangle)} \right) \right| \\
&\leq \max_{x,a} |\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}, \boldsymbol{\phi}(x,a) \rangle| + \max_{x,a} |\log(\exp(\langle \boldsymbol{\theta}, \boldsymbol{\phi}(x,a) \rangle))| \\
&\leq 3RC.
\end{aligned} \tag{52}$$

Furthermore, according to the definition of $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$ in (4), we have:

$$|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})| = \left| \frac{1}{n} \sum_{i=1}^n f_{\boldsymbol{\theta}}(x_i) \right| \leq \max_x |f_{\boldsymbol{\theta}}(x)| \leq 3RC. \tag{53}$$

□

Lemma 15. *For any $\boldsymbol{\theta} \in \Theta_R$, it holds that*

$$\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}})\| \leq \frac{4RC^2}{n}.$$

Proof. Since datasets $\boldsymbol{S}$ and $\boldsymbol{S}'$ differ by one example, W.o.l.g. $x_i \in \boldsymbol{S}$ and $x_i' \in \boldsymbol{S}'$ are different, then

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}})\| =& \left\| \frac{1}{n} (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_i) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_i')) \right\| \\
=& \frac{1}{n} \left\| \left(\boldsymbol{V}_{x_i}(\boldsymbol{\theta}) - \boldsymbol{V}_{x_i'}(\boldsymbol{\theta}) \right) (\boldsymbol{\theta}^* - \boldsymbol{\theta}) \right\| \\
\leq& \frac{1}{n} \left\| \boldsymbol{V}_{x_i}(\boldsymbol{\theta}) - \boldsymbol{V}_{x_i'}(\boldsymbol{\theta}) \right\| \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\| \\
\leq& \frac{2R}{n} (\|\boldsymbol{V}_{x_i}(\boldsymbol{\theta})\| + \|\boldsymbol{V}_{x_i'}(\boldsymbol{\theta})\|) \\
\leq& \frac{4RC^2}{n}.
\end{aligned}$$

where the last inequality is from (49). □

Lemma 16. *For any $\boldsymbol{\theta} \in \Theta_R$, we have*

$$\mathbb{E}_{i_t} \left[|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t}) - \mathbb{E}_{i_t} [\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t})]|^2 \right] \leq 16R^2C^4.$$

Proof. Firstly, due to the selection of the index i_t, $\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t})$ is an unbiased estimation of $\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$, i.e., $\mathbb{E}[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t})] = \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$. Thus we have

$$\begin{aligned}
\mathbb{E} \left[\|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t}) - \mathbb{E}[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t})]\|^2 \right] =& \mathbb{E} \left[\|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\|^2 \right] \\
=& \mathbb{E} \left[\frac{1}{n^2} \left\| \sum_{i \neq i_t} (\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_{i_t}) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_i)) \right\|^2 \right] \\
\leq& \sup_{x_i, x_j \in \boldsymbol{S}} \|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_i) - \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_j)\|^2 \\
=& \sup_{x_i, x_i \in \boldsymbol{S}} \left\| (\boldsymbol{V}_{x_i}(\boldsymbol{\theta}) - \boldsymbol{V}_{x_j}(\boldsymbol{\theta}))(\boldsymbol{\theta}^* - \boldsymbol{\theta}) \right\|^2 \\
\leq& (2 \sup_x \|\boldsymbol{V}_x(\boldsymbol{\theta})\| \cdot \|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|)^2 \\
\leq& 16R^2C^4.
\end{aligned} \tag{54}$$

□

Remark 5. *In fact, according to the proof of Lemma 16, we have a stronger result:*

Given $\boldsymbol{S}$, for any $\boldsymbol{\theta} \in \Theta_R$, we have

$$\sup_{x \in \boldsymbol{S}} \|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})\|^2 \leq 16R^2C^4 = G^2. \tag{55}$$

We now introduce two lemmas that describe the properties of the covariance matrices $\boldsymbol{V}_x(\boldsymbol{\theta})$ and $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$.

Lemma 17. *Under Assumption 4, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and x, it holds that*

$$\mathrm{C}(\boldsymbol{V}_x(\boldsymbol{\theta}_1)) = \mathrm{C}(\boldsymbol{V}_x(\boldsymbol{\theta}_2)), \tag{56}$$

when $d \geq |\mathcal{A}|$. Specifically, for any $\boldsymbol{\theta}$, it holds that

$$\mathrm{C}(\boldsymbol{V}_x(\boldsymbol{\theta})) = \mathrm{C}(\boldsymbol{V}_x(\boldsymbol{\theta}^*_{\boldsymbol{S}})).$$

Proof. According to (11), we have

$$\begin{aligned}
\boldsymbol{V}_x(\boldsymbol{\theta}) &= \mathbb{E}_{\pi_{\boldsymbol{\theta}}}\left[\left(\boldsymbol{\phi}(x,a) - \mathbb{E}[\boldsymbol{\phi}(x,a')]\right)\left(\boldsymbol{\phi}(x,a) - \mathbb{E}[\boldsymbol{\phi}(x,a')]\right)^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{\phi}(x,a)\boldsymbol{\phi}(x,a)^\top\right] - \mathbb{E}[\boldsymbol{\phi}(x,a)]\mathbb{E}[\boldsymbol{\phi}(x,a)]^\top \\
&= \sum_{i=1}^{|\mathcal{A}|} \boldsymbol{\phi}(x,a_i)\pi_{\boldsymbol{\theta}}(a_i|x)\boldsymbol{\phi}(x,a_i)^\top - \left(\sum_{i=1}^{|\mathcal{A}|} \boldsymbol{\phi}(x,a_i)\pi_{\boldsymbol{\theta}}(a_i|x)\right)\left(\sum_{i=1}^{|\mathcal{A}|} \boldsymbol{\phi}(x,a_i)\pi_{\boldsymbol{\theta}}(a_i|x)\right)^\top \\
&= \Phi(x)\Lambda_{\boldsymbol{\theta}}\Phi(x)^\top - \left(\Phi(x)\boldsymbol{p}_{\boldsymbol{\theta}}\right)\left(\Phi(x)\boldsymbol{p}_{\boldsymbol{\theta}}\right)^\top \\
&= \Phi(x)(\Lambda_{\boldsymbol{\theta}} - \boldsymbol{p}_{\boldsymbol{\theta}}\boldsymbol{p}_{\boldsymbol{\theta}}^\top)\Phi(x)^\top,
\end{aligned}$$

where $\Lambda_{\boldsymbol{\theta}} = \mathsf{diag}\{\pi_{\boldsymbol{\theta}}(a_1|x), \cdots, \pi_{\boldsymbol{\theta}}(a_{|\mathcal{A}|}|x)\}$, $\boldsymbol{p}_{\boldsymbol{\theta}} = (\pi_{\boldsymbol{\theta}}(a_1|x), \cdots, \pi_{\boldsymbol{\theta}}(a_{|\mathcal{A}|}|x))^\top$ and $\Phi(x) \triangleq \left(\boldsymbol{\phi}(x,a_1), \boldsymbol{\phi}(x,a_2), \cdots, \boldsymbol{\phi}(x,a_{|\mathcal{A}|})\right) \in \mathbb{R}^{d\times|\mathcal{A}|}$. Let $C_{\boldsymbol{p}_{\boldsymbol{\theta}}} = \Lambda_{\boldsymbol{\theta}} - \boldsymbol{p}_{\boldsymbol{\theta}}\boldsymbol{p}_{\boldsymbol{\theta}}^\top$.

By Lemma 24, it holds that:

$$\mathrm{C}(\boldsymbol{V}_x(\boldsymbol{\theta}_1)) = \mathrm{C}\left(\Phi(x)C_{\boldsymbol{p}_{\boldsymbol{\theta}_1}}\Phi(x)^\top\right) = \{\Phi(x)w | w^\top v = 0, v = (1,\cdots,1)^\top\} = \mathrm{C}\left(\Phi(x)C_{\boldsymbol{p}_{\boldsymbol{\theta}_2}}\Phi(x)^\top\right) = \mathrm{C}(\boldsymbol{V}_x(\boldsymbol{\theta}_2)),$$

which proves our conclusion. □

Lemma 18. *Under Assumption 4, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and dataset $\boldsymbol{S}$, it holds*

$$\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_1)) = \mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_2)), \tag{57}$$

when $d \geq |\mathcal{A}|$. Specifically, for any $\boldsymbol{\theta}$, it holds that

$$\mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})) = \mathrm{C}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})).$$

Proof. We prove this lemma by induction on n, i.e., the number of elements in the set $\boldsymbol{S}$. When $n = 1$, (57) becomes Lemma 17. Thus Lemma 17 implies the conclusion.

For $n = 2$, we prove that

$$\mathrm{C}\left(\boldsymbol{V}_{x_1}(\boldsymbol{\theta}_1) + \boldsymbol{V}_{x_2}(\boldsymbol{\theta}_1)\right) = \mathrm{C}\left(\boldsymbol{V}_{x_1}(\boldsymbol{\theta}_2) + \boldsymbol{V}_{x_2}(\boldsymbol{\theta}_2)\right),$$

given the conditions that

$$\mathrm{C}\left(\boldsymbol{V}_{x_1}(\boldsymbol{\theta}_1)\right) = \mathrm{C}\left(\boldsymbol{V}_{x_1}(\boldsymbol{\theta}_2)\right) \text{ and } \mathrm{C}\left(\boldsymbol{V}_{x_2}(\boldsymbol{\theta}_1)\right) = \mathrm{C}\left(\boldsymbol{V}_{x_2}(\boldsymbol{\theta}_2)\right).$$

Then, the conclusion is implied by Lemma 22.

Now, suppose the lemma holds for $n = k \geq 2$, since $\mathrm{C}(\sum_{i=1}^k \boldsymbol{V}_{x_i}(\boldsymbol{\theta}_1)) = \mathrm{C}(\sum_{i=1}^k \boldsymbol{V}_{x_i}(\boldsymbol{\theta}_2))$ and $\mathrm{C}(\boldsymbol{V}_{x_{k+1}}(\boldsymbol{\theta}_1)) = \mathrm{C}(\boldsymbol{V}_{x_{k+1}}(\boldsymbol{\theta}_2))$, we prove the conclusion for $n = k+1$ by Lemma 17. Then we prove our conclusion. □

Lemma 19. *Under Assumption 2, for a given initial parameter $\boldsymbol{\theta}_1$ satisfying $\|\boldsymbol{\theta}_1\| \leq D$, any parameter $\boldsymbol{\theta}$ obtained by* (14) *and* (18) *satisfies*

$$\|\boldsymbol{\theta}\| \leq 3D. \tag{58}$$

Proof. We first analyze the case where the parameters are updated using (14). According to Lemma 11, the learning rate $\eta = \frac{1}{L_f}$, for any $t \geq 1$, we have

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^* = \boldsymbol{\theta}_t + \eta\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}) - \boldsymbol{\theta}^* = \boldsymbol{\theta}_t - \boldsymbol{\theta}^* + \eta\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)(\boldsymbol{\theta}^* - \boldsymbol{\theta}_t) = \left(I - \eta\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*). \tag{59}$$

According to the definition of $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)$ and (49), for any $\boldsymbol{u}$ satisfies $\|\boldsymbol{u}\| = 1$, it holds that

$$\boldsymbol{u}^\top (I - \eta \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t))\boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{u} - \eta \cdot \boldsymbol{u}^\top \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)\boldsymbol{u} \geq \boldsymbol{u}^\top \boldsymbol{u}(1 - \eta \cdot C^2) = 1 - \frac{1}{8RC+1} \geq 0. \tag{60}$$

Additionally, by the positive semi-definiteness of covariance matrix $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})$, it holds that

$$\boldsymbol{u}^\top (I - \eta \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t))\boldsymbol{u} = \boldsymbol{u}^\top \boldsymbol{u} - \eta \cdot \boldsymbol{u}^\top \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)\boldsymbol{u} \leq \boldsymbol{u}^\top \boldsymbol{u} = 1, \tag{61}$$

which means that the spectral radius of the matrix $I - \eta \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)$ is no more than 1. Combining with (59), we have

$$\begin{aligned} \|\boldsymbol{\theta}_{t+1}\| &\leq \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\| = \| \left(I - \eta \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_t)\right) (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)\| + \|\boldsymbol{\theta}^*\| \\ &\leq \|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\| \leq 3D. \end{aligned} \tag{62}$$

Analogously, the parameters obtained from (18) are also bounded by $\|\boldsymbol{\theta}\| \leq 3D$. □

This lemma demonstrates that with proper initialization, the parameters obtained by gradient-based algorithms are naturally constrained to lie within a compact set.

C Some Technical Lemmas

Lemma 1. (Concentration inequality) Under Assumptions 1, 2 and 5, for any $n \geq 1$ and $\rho \in (0, 1)$, the following bound holds with probability at least $1 - \rho$ over the random draw of the dataset $\boldsymbol{S}$,

$$|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})| \leq 3RC\sqrt{\frac{2\log(2/\rho)}{n}}. \tag{7}$$

Proof. According to Propositions 4 and 3, it holds that

$$J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) = J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*}).$$

Let the dataset be $\boldsymbol{S} = \{x_1, \cdots, x_n\}$. We define the random variable X_i for each sample as: $X_i = f_{\boldsymbol{\theta}^*}(x_i)$. Since the samples x_i are drawn i.i.d. from the context space $\mathcal{X}$, X_i are independent random variables, therefore

$$J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) = J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*}) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then we can calculate the expectation of X_i as follows:

$$\mathbb{E}[X_i] = \mathbb{E}_{\boldsymbol{S}}[J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})] = \mathbb{E}_{\boldsymbol{S}}[J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*})] = J(\pi_{\boldsymbol{\theta}^*}).$$

Furthermore, according to Lemma 14, X_i is bounded by:

$$|X_i| = |f_{x_i}(\pi_{\boldsymbol{\theta}^*})| \leq 3RC.$$

We apply the Chernoff-Hoeffding inequality for bounded random variables X_i. For the empirical mean of n independent random variables $X_i \in [-3RC, 3RC]$, the probability of deviating from the expectation by more than ϵ is bounded by:

$$\mathbb{P}\left(\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})\right| \geq \epsilon\right) = \mathbb{P}\left(\left|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geq n\epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(6RC)^2}\right). \tag{63}$$

We set the failure probability in (63) to ρ, with a simple identity transformation, we complete the proof of this Lemma. □

Lemma 2. *By taking $\eta = \frac{1}{L_f}$ (L_f is defined in Lemma 12), the parameter $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ satisfies*

$$\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\|^2 \leq \frac{12L_f RC}{T}. \tag{15}$$

Proof. Owing to the smoothness of $J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}})$, we know that

$$J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{t+1}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}) \geq \langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\rangle + \frac{L_f}{2}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2 = \frac{1}{2L_f}\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\|^2. \tag{64}$$

Taking a telescoping sum of the above inequality, we obtain

$$\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\|^2 \leq \frac{2L_f}{T}(J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{T+1}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_1})) \leq \frac{12L_f RC}{T}, \tag{65}$$

where the last inequality is obtained by Assumption 1 and Lemma 14. □

Lemma 4. *By taking* $\eta_t = \frac{1}{2L_f\sqrt{t}}$($L_f$ is defined in Lemma 12), for any $\rho' \in (0,1)$, with probability at least $1-\rho'$, the parameter $\boldsymbol{\theta}_{\boldsymbol{S},T}^{\mathrm{SGA}}$ satisfies

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S},T}^{\mathrm{SGA}}})\right\|^2 \leq \frac{6L_f}{\sqrt{T}}\left(\left(6RC + \frac{2R^2C^4}{L_f}\right) + \frac{2R^2C^4}{L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{1}{\rho'}\right)\right). \tag{66}$$

where $\lambda > 0$ satisfies $\frac{e^\lambda - \lambda - 1}{\lambda} \leq \frac{L_f}{8R^2C^4}$.

Proof. Let the stochastic gradient noise be

$$\nu_t = \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(x_{i_t}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}).$$

By the L_f-smoothness of the empirical objective $J_{\boldsymbol{S}}$ according to Lemma 12, for the SGA update rule in (18), we have the following equality:

$$\begin{aligned} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{t+1}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}) &\geq \langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\rangle - \frac{L_f}{2}\left\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\right\|^2 \\ &= \langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \eta_t(\nu_t + \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}))\rangle - \frac{L_f}{2}\eta_t^2\left\|\nu_t + \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 \\ &= \left(\eta_t - \frac{L_f}{2}\eta_t^2\right)\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 - \frac{L_f}{2}\eta_t^2\|\nu_t\|^2 + (\eta_t - L_f\eta_t^2)\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \nu_t\rangle. \end{aligned} \tag{67}$$

By rearranging the terms, we have

$$\left(\eta_t - \frac{L_f}{2}\eta_t^2\right)\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 \leq J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{t+1}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}) + \frac{L_f}{2}\eta_t^2\|\nu_t\|^2 - (\eta_t - L_f\eta_t^2)\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \nu_t\rangle. \tag{68}$$

Summing the (68) from $t=1$ to T, we obtain that:

$$\begin{aligned} \sum_{t=1}^{T}\left(\eta_t - \frac{L_f}{2}\eta_t^2\right)\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 &\leq \Delta J + \frac{L_f}{2}\sum_{t=1}^{T}\eta_t^2\|\nu_t\|^2 - \sum_{t=1}^{T}(\eta_t - L_f\eta_t^2)\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \nu_t\rangle \\ &\leq \Delta J + \frac{L_f G^2}{2}\sum_{t=1}^{T}\frac{1}{4L_f^2 t} - \sum_{t=1}^{T}(\eta_t - L_f\eta_t^2)\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \nu_t\rangle \\ &\leq \Delta J + \frac{G^2}{8L_f}(1+\log(T)) \underbrace{- \sum_{t=1}^{T}(\eta_t - L_f\eta_t^2)\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t}), \nu_t\rangle}_{\text{Term 1}}. \end{aligned} \tag{69}$$

where $\Delta J = J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{T+1}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_1})$, the second equality is from Lemma 16 that for any t, we have $\|\nu_t\|^2 \in [0, G^2]$.

We now need to prove the left term in (69) is small with high probability.

By using Freedman's inequality (Theorem 1.6 inFreedman (1975)), and combining $\mathbb{E}[\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_t), \nu_t\rangle] = 0$, for any $\lambda \neq 0$, with probability at least $1-\rho_1$, we have

$$\begin{aligned} -\sum_{t=1}^{T} m_t\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_t), \nu_t\rangle &\leq \frac{e^\lambda - \lambda - 1}{\lambda}\sum_{t=1}^{T}\mathbb{E}\left[\|m_t\langle \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_t), \nu_t\rangle\|^2 | \mathcal{F}_t\right] + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right) \\ &\leq \frac{e^\lambda - \lambda - 1}{\lambda}\sum_{t=1}^{T}\frac{\eta_t^2}{4}\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_t)\|^2 \cdot \mathbb{E}\left[\|\nu_t\|^2 | \mathcal{F}_{t-1}\right] + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right) \\ &\leq \frac{e^\lambda - \lambda - 1}{\lambda}\cdot\frac{G^2}{4}\cdot\sum_{t=1}^{T}\eta_t^2\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}_t)\|^2 + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right), \end{aligned} \tag{70}$$

where $m_t = \eta_t - L_f\eta_t^2 \leq \frac{\eta_t}{2}$ for any $t \geq 1$.

Combining (70) with (69), it holds that

$$\sum_{t=1}^{T}\left(\eta_t - \frac{L_f}{2}\eta_t^2 - \frac{e^\lambda - \lambda - 1}{\lambda}\cdot\frac{G^2}{4}\eta_t^2\right)\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 \leq \Delta J + \frac{G^2}{8L_f}(1+\log(T)) + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right). \tag{71}$$

Based on the property of the function $\frac{e^\lambda-\lambda-1}{\lambda}$, we can choose λ such that $\frac{e^\lambda-\lambda-1}{\lambda} \leq \frac{2L_f}{G^2}$. Therefore, for any $t \geq 1$, we have

$$\eta_t - \frac{L_f}{2}\eta_t^2 - \frac{e^\lambda-\lambda-1}{\lambda}\cdot\frac{G^2}{4}\eta_t^2 \geq \eta_t\left(1-\frac{L_f}{2}\frac{1}{2L_f\sqrt{t}} - \frac{2L_f}{G^2}\cdot\frac{G^2}{4}\frac{1}{2L_f\sqrt{t}}\right) \geq \frac{\eta_t}{2}. \tag{72}$$

Take (72) into (71), then with probability at least $1-\rho_1$, it holds that

$$\begin{aligned}\frac{1}{\sum_{t=1}^T \frac{\eta_t}{2}}\sum_{t=1}^T \frac{\eta_t}{2}\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 &\leq \frac{1}{\sum_{t=1}^T \frac{\eta_t}{2}}\cdot\sum_{t=1}^T(\eta_t - \frac{L_f}{2}\eta_t^2 - \frac{e^\lambda-\lambda-1}{\lambda}\cdot\frac{G^2}{4}\eta_t^2)\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_t})\right\|^2 \\ &\leq \frac{2L_f}{\sqrt{T+1}-1}\left(\Delta J + \frac{G^2}{8L_f}(1+\log(T)) + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right)\right) \\ &\leq \frac{6L_f}{\sqrt{T}}\left(\left(\Delta J + \frac{G^2}{8L_f}\right) + \frac{G^2}{8L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right)\right).\end{aligned} \tag{73}$$

According to the definition of $\boldsymbol{\theta}_{\boldsymbol{S},T}^{\mathrm{SGA}}$ in 7, it holds that

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_{\boldsymbol{S},T}^{\mathrm{SGA}}})\right\|^2 \leq \mathbb{E}_\tau\left[\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}_\tau})\right\|^2\right] \leq \frac{6L_f}{\sqrt{T}}\left(\left(\Delta J + \frac{G^2}{8L_f}\right) + \frac{G^2}{8L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{1}{\rho}\right)\right). \tag{74}$$

According to Lemmas 14 and 16, we have $\Delta J \leq 6RC$ and $G^2 = 16R^2C^4$. By combining with (74), we prove our conclusion. □

D Proofs for Section 5.1

Theorem 1. *Under Assumption 3, when*

$$n > \frac{128C^4}{\sigma^2}\left(\log\left(\frac{1}{\rho}\right) + d\log\left(\frac{432RC^5}{\sigma^2}\right)\right), \tag{10}$$

the ground truth $\boldsymbol{\theta}^*$ is the unique empirical stationary point in Θ_R with probability at least $1-\rho$.

Proof. As clarified before, to make $\boldsymbol{\theta}_{\boldsymbol{S}}^* = \boldsymbol{\theta}^*$, it is sufficient to show $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}_{\boldsymbol{S}}^*)$ is positive-definite. This can be done by combining the concentration of $\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*)$ and the continuity obtained from Lemma 13. Firstly, let us show the concentration result. By defining $\boldsymbol{V}(\boldsymbol{\theta}) = \mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})]$, and i.i.d. Rademacher random variable $\{\epsilon_i\}_{i=1}^n$ independent with $\boldsymbol{S}$, for any $\gamma > 0$, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\gamma\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\|\right)\right] =& \mathbb{E}\left[\exp\left(\gamma\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|=1}\boldsymbol{u}^\top\left(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\right)\boldsymbol{u}\right)\right] \\ \leq& \mathbb{E}\left[\exp\left(2\gamma\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|=1}\boldsymbol{u}^\top\left(\frac{1}{n}\sum_{i=1}^n\epsilon_i\boldsymbol{V}_{x_i}(\boldsymbol{\theta})\right)\boldsymbol{u}\right)\right] \\ =& \mathbb{E}\left[\mathbb{E}\left[\exp\left(2\gamma\sup_{\boldsymbol{u}:\|\boldsymbol{u}\|=1}\frac{1}{n}\sum_{i=1}^n\epsilon_i\boldsymbol{u}^\top\boldsymbol{V}_{x_i}(\boldsymbol{\theta})\boldsymbol{u}\right) \mid \boldsymbol{S}\right]\right] \\ \leq& \exp\left(\frac{8C^4\gamma^2}{n}\right),\end{aligned} \tag{75}$$

where the first inequality is from Proposition 4.11 (b) in Wainwright (2019), the last inequality is obtained from $\|\boldsymbol{V}_x(\boldsymbol{\theta})\| \leq C^2$, and the sub-Gaussian property of Rademacher random variable. With the above inequality, by applying Markov's inequality, we have

$$\mathbb{P}\left(\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\| \geq \delta\right) \leq \inf_\gamma \exp\left(-\gamma\delta + \frac{8C^4\gamma^2}{n}\right) = \exp\left(-\frac{n\delta^2}{32C^4}\right), \tag{76}$$

by taking $\gamma = \frac{n\delta}{16C^4}$. Since $\mathrm{diam}(\Theta_{\mathrm{R}}) \leq \mathrm{R}$, from (Zhang et al., 2017) we know that Θ_R has an r-covering $\mathcal{C}$ such that for any $\boldsymbol{\theta} \in \Theta_R$ there exists a $\boldsymbol{\theta}' \in \mathcal{C}$ satisfies $\|\boldsymbol{\theta}-\boldsymbol{\theta}'\| \leq r$ and $|\mathcal{C}| \leq (3R/r)^d$. Then by applying Bonferroni inequality and (76), we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\mathcal{C}}\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\| \geq \delta\right) \leq |\mathcal{C}|\exp\left(-\frac{n\delta^2}{32C^4}\right) \leq \exp\left(d\log\left(\frac{3R}{r}\right) - \frac{n\delta^2}{32C^4}\right). \tag{77}$$

By applying Lemma 13, we know that when $\{\sup_{\boldsymbol{\theta}\in\mathcal{C}} \|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\| \leq \delta\}$ happens, for any $\boldsymbol{\theta} \in \Theta_R$ and its corresponded $\boldsymbol{\theta}' \in \mathcal{C}$,

$$\|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\| \leq \|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}')\| + \|\boldsymbol{V}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta}')\| + \|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}') - \boldsymbol{V}(\boldsymbol{\theta}')\| \leq 6C^{\frac{5}{2}} r^{\frac{1}{2}} + \delta. \tag{78}$$

Then, by Weyl's theorem, we have

$$|\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})) - \sigma_{\min}(\boldsymbol{V}(\boldsymbol{\theta}))| \leq \|\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}) - \boldsymbol{V}(\boldsymbol{\theta})\| \leq 6C^{\frac{5}{2}} r^{\frac{1}{2}} + \delta, \tag{79}$$

which implies $\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta})) > 0$ by taking $r = (1/6C^{\frac{5}{2}})^2\delta^2$ and $\delta < \sigma/2$. By plugging these into (77), we obtain our conclusion. □

E Proofs for Section 5.2

Theorem 2. *Under Assumptions 1, 2 and 5, we have*

$$\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}) \leq \left(\frac{\Gamma_1}{n} + 2RC\sqrt{\frac{\Gamma_1}{n}}\right), \tag{80}$$

where

$$\Gamma_1 = \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ 4nR\epsilon_n + \frac{8RC^3 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}))} \right\}.$$

Then, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$,

$$\left| J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) \right| \leq \mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}) \log(n) \log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right). \tag{81}$$

Proof. As clarified in Section 4, the generalization error is implied by invoking uniform algorithmic stability. To obtain this, for any $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ and $\boldsymbol{\theta}^*_{\boldsymbol{S}}$ satisfies $\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}})\| = \|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\boldsymbol{\theta}^*_{\boldsymbol{S}'})\| = 0$, we explore the algorithmic stability. We first evaluate the gradient norm of $\boldsymbol{\theta}^*_{\boldsymbol{S}'}$ with respect to the objective $J_{\boldsymbol{S}}$. According to Lemma 15, it holds that

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'})\| &= \|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}) + \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\boldsymbol{\theta}^*_{\boldsymbol{S}'})\| \\ &\leq \|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\boldsymbol{\theta}^*_{\boldsymbol{S}'})\| + \|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\boldsymbol{\theta}^*_{\boldsymbol{S}'})\| \leq \frac{4RC^2}{n}.\end{aligned}$$

By using Lemma 6, it holds that

$$\begin{aligned}&\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}) = \sup_{x\in\mathcal{X},\boldsymbol{S},\boldsymbol{S}'} |f_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}(x) - f_{\boldsymbol{\theta}^*_{\boldsymbol{S}'}}(x)| \\ &\leq \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ 4R\epsilon_n + \frac{8RC^3 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{n\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}))} \right\} + 2CR \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ \sqrt{4R\epsilon_n + \frac{8RC^3 \cdot \sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{n\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}'}))}} \right\} \\ &= \frac{\Gamma_1}{n} + 2CR\sqrt{\frac{\Gamma_1}{n}},\end{aligned} \tag{82}$$

which proves our conclusion by invoking Proposition 2. □

Theorem 3. *Under Assumptions 1, 2, 5, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$,*

$$\begin{aligned}|J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})| &\leq \left(\frac{\Gamma_1}{n} + 2RC\sqrt{\frac{\Gamma_1}{n}}\right)\log(n)\log\left(\frac{2}{\rho}\right) + 6RC\sqrt{\frac{\log(2/\rho)}{n}} + 3RC\sqrt{\frac{2\log(4/\rho)}{n}} \\ &= \tilde{\mathcal{O}}\left(n^{-\frac{1}{2}}\right).\end{aligned} \tag{83}$$

Proof. Firstly, we decompose the suboptimality gap of $\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}$ as follows:

$$\left|J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})\right| \leq \left|J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| + \left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})\right|. \tag{84}$$

Then, according to Theorem 2, for $\rho_1 \in (0,1)$, with probability at least $1-\rho_1$, it holds that

$$\left|J(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| \leq \left(\frac{\Gamma_1}{n} + 2RC\sqrt{\frac{\Gamma_1}{n}}\right)\log(n)\log\left(\frac{1}{\rho_1}\right) + 6RC\sqrt{\frac{\log(1/\rho_1)}{n}}. \tag{85}$$

Now we turn to the second term in (84). By using Lemma 1, for $\rho_2 \in (0,1)$, with probability at least $1-\rho_2$, it holds that

$$\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})\right| \leq 3RC\sqrt{\frac{2\log(2/\rho_2)}{n}}. \tag{86}$$

Let $\rho_1 = \rho_2 = \frac{\rho}{2}$, by combining (85) and (86), we complete the proof. □

F Proofs for Section 6.1

Lemma 3. *Under Assumptions 1, 2 and 5, any $\boldsymbol{\theta}^*_{\boldsymbol{S}} \in \Theta_{\boldsymbol{S}}$ and $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ from Definition 6 satisfy:*

$$\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| = \mathcal{O}\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right). \tag{16}$$

Proof. According to Lemma 2, we have

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\right\| \leq 2\sqrt{\frac{3L_f RC}{T}} \quad \text{and} \quad \left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right\| = 0.$$

By using Lemma 6, it holds that

$$\begin{aligned}
&\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|f_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}(x_i) - f_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}(x_i)\right| \leq \sup_{x\in\mathcal{X}}\left|f_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}(x) - f_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}(x)\right| \\
\leq& 4R\epsilon_n + 4C\sqrt{\frac{3L_f RC}{T}}\frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}))} \\
&+ 2RC\sqrt{4R\epsilon_n + 4C\sqrt{\frac{3L_f RC}{T}}\frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}))}}.
\end{aligned}$$

We highlight that the bound above is of the form $X + \alpha\sqrt{X}$. In the regime where n and T are large, X is small enough such that the linear term X becomes negligible compared to the square root term $\sqrt{X}$ (i.e., $X = o(\sqrt{X})$). Thus, the bound is dominated by the square root component. Applying the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain that

$$\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| \lesssim \sqrt{\epsilon_n} + \sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}}}T^{-\frac{1}{4}} = \mathcal{O}(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}).$$

where $\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}))}$.

We complete the proof of this theorem. □

Theorem 4. *Under Assumptions 1, 2 and 5, for the GA output $\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}$ defined in Definition 6, we have*

$$\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{GA}_{\boldsymbol{S},T}) \lesssim \sqrt{\epsilon_n} + \left(\sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}}} + \sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right)\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right),$$

where

$$\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'} = \sup_{\boldsymbol{S},\boldsymbol{S}'}\frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}))} \quad \text{and} \quad \Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}}\frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}))}.$$

Then, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$:

$$\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\right| \leq \mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{GA}_{\boldsymbol{S},T})\log(n)\log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} = \tilde{\mathcal{O}}\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right). \tag{87}$$

Proof. Similarly to Theorem 2, we prove the algorithmic stability. Given $\boldsymbol{S}, \boldsymbol{S}'$, we can obtain $\boldsymbol{\theta}^*_{\boldsymbol{S},T}$ and $\boldsymbol{\theta}^*_{\boldsymbol{S}',T}$ defined in Definition 6, respectively. According to Lemma 2, we have

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\right\| \leq 2\sqrt{\frac{3L_f RC}{T}} \quad \text{and} \quad \left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}})\right\| \leq 2\sqrt{\frac{3L_f RC}{T}}.$$

According to Lemma 15, it holds that

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}})\right\| \leq \left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}})\right\| + \left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}})\right\| \leq 2\sqrt{\frac{3L_f RC}{T}} + \frac{4RC^2}{n}.$$

By using Lemma 6, it holds that

$$\begin{aligned}
&\mathcal{E}_{\text{stab}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}) = \sup_{x\in\mathcal{X},\boldsymbol{S},\boldsymbol{S}'} |f_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}(x) - f_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}}(x)| \\
&\leq \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ 4R\epsilon_n + 4C\sqrt{\frac{3L_f RC}{T}}\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} + \left(4C\sqrt{\frac{3L_f RC}{T}} + \frac{8RC^3}{n}\right)\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'} \right\} \\
&\quad + 2RC \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ \sqrt{4R\epsilon_n + 4C\sqrt{\frac{3L_f RC}{T}}\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} + \left(4C\sqrt{\frac{3L_f RC}{T}} + \frac{8RC^3}{n}\right)\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'}} \right\} \\
&\leq \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ 4R\epsilon_n + 4C\sqrt{\frac{3L_f RC}{T}}\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} + \left(4C\sqrt{\frac{3L_f RC}{T}} + \frac{8RC^3}{n}\right)\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'} \right\} \\
&\quad + 2RC \sup_{\boldsymbol{S},\boldsymbol{S}'} \left\{ \sqrt{4R\epsilon_n} + \sqrt{4C\sqrt{\frac{3L_f RC}{T}}\left(\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} + \Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'}\right)} + \sqrt{\frac{8RC^3}{n}\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'}} \right\} \\
&\lesssim \sqrt{\epsilon_n} + \left(\sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}}} + \sqrt{\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right)\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right).
\end{aligned} \tag{88}$$

where $\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}))}$ and $\Gamma^{\mathrm{GA}}_{\boldsymbol{S},\boldsymbol{S}'} = \sup_{\boldsymbol{S},\boldsymbol{S}'} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S}',T}))}$.

We can now complete the proof of the theorem by applying Proposition 2. □

Theorem 5. *Under Assumptions 1, 2, and 5, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho$ over the randomness of $\boldsymbol{S}$: $\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J(\pi_{\boldsymbol{\theta}^*})\right| = \tilde{\mathcal{O}}\left(T^{-\frac{1}{4}} + n^{-\frac{1}{2}}\right)$.*

Proof. Recall the suboptimality gap decomposition in Definition 3, we have

$$\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J(\pi_{\boldsymbol{\theta}^*})\right| \leq \left|J(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}})\right| + \left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{GA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| + \left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})\right|.$$

By combining the results of Theorem 4, Lemma 3 and Lemma 1, we complete the proof of this theorem. □

G Proofs for Section 6.2

When we consider a randomized algorithm $\mathcal{H} : \mathcal{X}^n \times \Omega \to \mathbb{R}^d$, the standard notion of uniform stability defined in Definition 5 is insufficient to provide a high-probability generalization guarantee. Instead, we adopt the notion of uniform stability with high probability (Yuan & Li, 2023).

Definition 8. *A randomized algorithm $\mathcal{H}$ has $\mathcal{E}_{\mathrm{stab}}-$uniform stability with probability at least $1-\rho'$ for some ρ' over the random draw of $\zeta \in \Omega$, if for any two datasets $\boldsymbol{S}, \boldsymbol{S}' \in \mathcal{X}^n$ that differ by at most one example, it holds that:*

$$\mathbb{P}\left\{ \sup_{x\in\mathcal{X},\boldsymbol{S},\boldsymbol{S}'} |f_{\mathcal{H}(\boldsymbol{S},\zeta)}(x) - f_{\mathcal{H}(\boldsymbol{S}',\zeta)}(x)| \leq \mathcal{E}_{\mathrm{stab}} \right\} \geq 1-\rho'. \tag{89}$$

Corollary 2. *Suppose that the randomness of $\mathcal{H}$ is independent of the training set $\boldsymbol{S}$. Then the bound in Proposition2 naturally implies that with probability at least $1-\rho-\rho'$ over $\boldsymbol{S}$ and ζ,*

$$\left|J(\pi_{\mathcal{H}(\boldsymbol{S},\zeta)}) - J_{\boldsymbol{S}}(\pi_{\mathcal{H}(\boldsymbol{S},\zeta)})\right| \lesssim \mathcal{E}_{\mathrm{stab}} \log(n) \log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} + \mathcal{O}(\rho'). \tag{90}$$

Theorem 6. *Under Assumptions 1, 2 and 5, for any $\rho' \in (0,1)$ and parameter $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$, we have*

$$\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}) \lesssim \sqrt{\epsilon_n} + \left(\sqrt{\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}}} + \sqrt{\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right)\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right),$$

where

$$\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}))} \quad \text{and} \quad \Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'} = \sup_{\boldsymbol{S},\boldsymbol{S}'} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',\mathrm{T}}))}.$$

Then, for any $\rho \in (0,1)$, the following bound holds with probability at least $1-\rho-\rho'$ over the randomness of $\boldsymbol{S}$ and i_t in SGA:

$$\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}})\right| \le \mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T})\log(n)\log\left(\frac{1}{\rho}\right) + 6RC\sqrt{\frac{\log(1/\rho)}{n}} = \tilde{\mathcal{O}}\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right). \tag{91}$$

Proof. We first analyze the uniform stability of the algorithm. For the given dataset $\boldsymbol{S}, \boldsymbol{S}'$, we can obtain two parameters $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$ and $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}$ defined in Definition 7, respectively. According to Lemma 4, with probability at least $1-\frac{\rho'}{2}$, it holds that:

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}})\right\| \le \sqrt{\frac{6L_f}{\sqrt{T}}\left(\left(6RC + \frac{2R^2C^4}{L_f}\right) + \frac{2R^2C^4}{L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{2}{\rho'}\right)\right)} \triangleq \Delta(T,\rho')$$

and $$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}})\right\| \le \sqrt{\frac{6L_f}{\sqrt{T}}\left(\left(6RC + \frac{2R^2C^4}{L_f}\right) + \frac{2R^2C^4}{L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{2}{\rho'}\right)\right)} = \Delta(T,\rho').$$

Furthermore, according to Lemma 15, it holds that

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}})\right\| &\le \left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}}) - \nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}})\right\| + \left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}'}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}})\right\| \\
&\le \sqrt{\frac{6L_f}{\sqrt{T}}\left(\left(6RC + \frac{2R^2C^4}{L_f}\right) + \frac{2R^2C^4}{L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{2}{\rho'}\right)\right)} + \frac{4RC^2}{n} \\
&= \Delta(T,\rho') + \frac{4RC^2}{n}.
\end{aligned}$$

According to Definition 8, by using Lemma 6, with probability at least $1-\rho'$, it holds that:

$$\begin{aligned}
\mathcal{E}_{\mathrm{stab}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}) \le& \sup_{\boldsymbol{S},\boldsymbol{S}'}\left\{4R\epsilon_n + 2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} + 2C\left(\Delta(T,\rho') + \frac{4RC^2}{n}\right)\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}\right\} \\
&+ 2RC\sup_{\boldsymbol{S},\boldsymbol{S}'}\left\{\sqrt{4R\epsilon_n + 2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} + 2C\left(\Delta(T,\rho') + \frac{4RC^2}{n}\right)\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right\} \\
\le& \sup_{\boldsymbol{S},\boldsymbol{S}'}\left\{4R\epsilon_n + 2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} + 2C\left(\Delta(T,\rho') + \frac{4RC^2}{n}\right)\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}\right\} \\
&+ 2RC\sup_{\boldsymbol{S},\boldsymbol{S}'}\left\{\sqrt{4R\epsilon_n} + \sqrt{2C\Delta(T,\rho')\left(\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} + \Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}\right)} + \sqrt{\frac{8RC^3}{n}\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right\} \\
\lesssim& \sqrt{\epsilon_n} + \left(\sqrt{\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}}} + \sqrt{\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'}}\right)\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right),
\end{aligned} \tag{92}$$

where $\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}))}$ and $\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}'} = \sup_{\boldsymbol{S},\boldsymbol{S}'} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}))}$.

Then we can complete the proof using Corollary 2.

□

Lemma 5. *Under Assumptions 1, 2 and 5, for any $\rho' \in (0,1)$, with probability at least $1-\frac{\rho'}{2}$,*

$$\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| = \tilde{\mathcal{O}}\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right). \tag{20}$$

Proof. According to Lemma 4, with probability at least $1 - \frac{\rho'}{2}$, it holds that:

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right\| = 0. \quad \text{and}$$

$$\left\|\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}})\right\| \leq \sqrt{\frac{6L_f}{\sqrt{T}}\left(\left(6RC + \frac{2R^2C^4}{L_f}\right) + \frac{2R^2C^4}{L_f}\log(T) + \frac{1}{\lambda}\log\left(\frac{2}{\rho'}\right)\right)} \triangleq \Delta(T, \rho')$$

By using Lemma 6, with probability at least $1 - \rho'$, it holds that

$$\begin{aligned}
&\left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|f_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}(x_i) - f_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}(x_i)\right| \leq \sup_{x\in\mathcal{X}}\left|f_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}(x) - f_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}(x)\right| \\
\leq& 4R\epsilon_n + 2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} + 2RC\sqrt{4R\epsilon_n + 2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}}} \\
\leq& 4R\epsilon_n + 2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} + 2RC\sqrt{4R\epsilon_n} + \sqrt{2C\Delta(T,\rho')\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}}} \\
=& \mathcal{O}\left(T^{-\frac{1}{8}}\right).
\end{aligned}$$

where $\Gamma^{\mathrm{SGA}}_{\boldsymbol{S},\boldsymbol{S}} = \sup_{\boldsymbol{S}} \frac{\sigma_{\max}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))}{\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^*_{\boldsymbol{S}}))\sigma^+_{\min}(\boldsymbol{V}_{\boldsymbol{S}}(\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}))}$.

We complete the proof of this theorem. □

Theorem 7. *Under Assumptions 1, 2 and 5, for any $\rho, \rho' \in (0,1)$, the following bound holds with probability at least $1-\rho-\rho'$ over the randomness of $\boldsymbol{S}$:* $\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J(\pi_{\boldsymbol{\theta}^*})\right| = \tilde{\mathcal{O}}\left(T^{-\frac{1}{8}} + n^{-\frac{1}{2}}\right)$.

Proof. Recall the suboptimality gap decomposition in Definition 3, we have

$$\left|J(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J(\pi_{\boldsymbol{\theta}^*})\right| \leq \left|J(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}})\right| + \left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}}) - J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}})\right| + \left|J_{\boldsymbol{S}}(\pi_{\boldsymbol{\theta}^*_{\boldsymbol{S}}}) - J(\pi_{\boldsymbol{\theta}^*})\right|.$$

By combining the results of Theorem 6, Lemma 1 and Lemma 5, we complete the proof of this theorem. □

Remark 6. *Crucially, while both the optimization and generalization bounds are probabilistic, they are not independent events. The optimization bound depends on the randomness of the single parameter $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$, while the generalization bound depends on the randomness of both $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S},T}$ and the perturbed parameter $\boldsymbol{\theta}^{\mathrm{SGA}}_{\boldsymbol{S}',T}$. Consequently, the event where the generalization bound holds strictly implies the satisfaction of the optimization bound.*

H Detailed Experiments Setup

In this appendix, we provide the detailed configurations for the experiments presented in Section 7.

H.1 The Results under Positive Definite Covariance Matrix

This setup corresponds to the results in Section 7.1.

Feature Construction. We construct the feature space using a standard orthonormal basis $\mathcal{X} = \{x_1, \cdots, x_d\}$ in $\mathbb{R}^d$. x is directly sampled from the basis set $\mathcal{S}$ according to a non-uniform distribution: the specific direction x_d is selected with a probability p (where p is very small), while the remaining $d-1$ directions are selected uniformly. Consequently, by defining the action space to encompass opposite directions, the set of feature vectors corresponding to all responses for a given x is simply $\{x, -x\}$, i.e., the feature vectors for a given x are defined as $\boldsymbol{\phi}(x, a_1) = x, \boldsymbol{\phi}(x, a_2) = -x$.

Hyper-parameters. We set the ambient dimension to $d = 20$. We vary the sample size n to observe the convergence behavior. Furthermore, we assume the ground truth parameter $\boldsymbol{\theta}^*$ and all feasible parameters are bounded by a unit ball.

Verification of Assumption 3 According to Lemma 17, for any $x \in \mathcal{S}$, it holds that

$$\mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})] = \sum_{i=1}^{d-1} \frac{1-p}{d-1} \boldsymbol{V}_{x_i}(\boldsymbol{\theta}) + p\boldsymbol{V}_{x_d}(\boldsymbol{\theta})$$
$$= \sum_{i=1}^{d-1} \frac{1-p}{d-1} \frac{4}{(\exp(\langle \boldsymbol{\theta}, x_i\rangle) + \exp(-\langle \boldsymbol{\theta}, x_i\rangle))^2} x_i x_i^\top + \frac{4p}{(\exp(\langle \boldsymbol{\theta}, x_d\rangle) + \exp(-\langle \boldsymbol{\theta}, x_d\rangle))^2} x_d x_d^\top.$$

From the definition of $\{x_1, \cdots, x_d\}$, it is straightforward to see that each x_i is an eigenvector of matrix $\mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})]$, with a corresponding eigenvalue satisfies

$$\sigma_i(\mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})]) = \begin{cases} \frac{1-p}{d-1} \frac{4}{(\exp(\langle \boldsymbol{\theta}, x_i\rangle) + \exp(-\langle \boldsymbol{\theta}, x_i\rangle))^2} & , i \in \{1, 2, \cdots, d-1\}; \\ p \cdot \frac{4}{(\exp(\langle \boldsymbol{\theta}, x_i\rangle) + \exp(-\langle \boldsymbol{\theta}, x_i\rangle))^2} & , i = d. \end{cases}$$

Therefore, for any $\boldsymbol{\theta}$ bounded by a unit ball, it holds that

$$\min_{\boldsymbol{\theta}} \sigma_{\min}(\mathbb{E}_x[\boldsymbol{V}_x(\boldsymbol{\theta})]) = \frac{4p}{e + e^{-1}} > 0.$$

This confirms that Assumption 3 holds theoretically. We vary the sample size n and check whether the empirical covariance matrix is strictly positive definite.

Evaluation. We estimate the probability of strict empirical coverage $(1-\rho)$ as a function of sample size n using Monte Carlo simulation. In our setting, checking whether the empirical covariance matrix $\boldsymbol{V}_{\boldsymbol{S}}$ is positive definite simplifies to a coverage problem. Since the feature vectors are drawn from a standard orthonormal basis $\mathcal{X} = \{x_1, \cdots, x_d\}$, the matrix $\boldsymbol{V}_{\boldsymbol{S}}$ is strictly positive definite if and only if the sampled features span the entire $\mathbb{R}^d$ space. Mathematically, this requires that every basis vector in $\mathcal{X}$ appears at least once in the dataset.

For each rare-feature probability p and sample size n, we conduct $m = 10,000$ independent trials. In each trial, we independently sample n feature vectors from the specified distribution. We record a trial as "successful" if the set of unique vectors collected has a cardinality of d (i.e., full coverage). The reported probability $1-\rho$ is the fraction of successful trials out of the m repetitions.

H.2 Generalization under Insufficient Coverage

This setup corresponds to the results in Section 7.2.

Feature Construction. We construct the feature space constrained to a low-dimensional subspace spanned by an orthonormal basis set $\mathcal{A} = \{a_1, \cdots, a_{d_{\text{eff}}}\}$ in $\mathbb{R}^d$, where $d_{\text{eff}} < d$. For a given x, there are two responses corresponding to it, the feature vectors are defined as $\boldsymbol{\phi}(x, a_{i_u}) = a_{i_u} a_{i_u}^\top x$ and $\boldsymbol{\phi}(x, a_{i_v}) = a_{i_v} a_{i_v}^\top x$.

In detail, we generate a vector x from a standard Gaussian distribution and normalize it to satisfy the boundedness assumption, ensuring it is non-orthogonal to the effective subspace $\mathcal{A}$. At the same time, we randomly select two indices $i_u, i_v \in \{1, \ldots, d_{\text{eff}}\}$.

Hyper-parameters. We set the ambient dimension d and the effective subspace dimension $d_{\text{eff}} < d$, creating a rank-deficient environment. The ground truth parameter $\boldsymbol{\theta}^*$ is generated strictly within this effective subspace.

Case 1: Suboptimality gap scaling with sample size n. We vary the sample size n from 7 to 22 to study the generalization error scaling. To verify the conclusion of Theorems 2, 4 and 6, we calculate the empirical stationary point exactly, run Gradient Ascent (GA) and Stochastic Gradient Ascent (SGA) for sufficiently many steps to ensure convergence, respectively. Specifically, based on our convergence analysis, we set the iteration steps $T^{\text{GA}} = 50n^2$ and $T^{\text{SGA}} = 10n^4$. We evaluate the generalization performance under two distinct settings: $d = 38, d_{\text{eff}} = 32$ and $d = 32, d_{\text{eff}} = 24$.

Case 2: Suboptimality gap decreasing with optimization steps T. We set the ambient dimension $d = 18$ and the effective subspace dimension $d_{\text{eff}} = 14$, creating a rank-deficient environment. We fix the sample size at two representative levels, $n = 7$ (high rank deficiency) and $n = 10$ (moderate rank deficiency). We vary the optimization steps T on a scale from 10^1 to 10^7 to observe the trajectory of the learned policy.

Table 1: Performance of output-head-only and full-parameter post-training on Qwen2.5-3B-Instruct across several benchmarks. Traning only on the output head while fix the learned representation.

Setting	AIME	Olymp.	Minerva	AMC23	Math500	GSM8K	GPQA	MMLU-P	ARC	Avg.
Base	7.60	28.59	27.21	43.59	64.80	87.11	19.00	49.64	86.29	45.98
Head-only	8.85	29.62	27.94	45.31	66.80	87.11	19.60	52.14	86.62	47.11
Full-param.	9.47	29.33	28.30	47.18	67.20	87.18	20.60	53.03	86.28	47.62

Verification of Assumption 5. By design, every feature vector $\boldsymbol{\phi}(x, a)$ lies within the d_{eff}-dimensional effective subspace. According to Lemma 17, When $n < d_{\text{eff}}$, the empirical covariance matrix $\boldsymbol{V_S}$ can only span at most an $n-$ dimensional subspace. Consequently, for any new feature vector $\boldsymbol{\phi}$, it cannot be fully represented by the column space of $\boldsymbol{V_S}$. It decomposes into a component within $\mathrm{C}(\boldsymbol{V_S})$ and a non-zero residual term $\boldsymbol{r_S}$ orthogonal to it. As the sample size n increases, the subspace spanned by $\boldsymbol{V_S}$ expands, causing the magnitude of the residual $\boldsymbol{r_S}$ to diminish statistically. This geometric property perfectly simulates the conditions characterized in Assumption 5, allowing us to validate the dimension-independent bounds derived in our theorems.

Evaluation. We first generate an independent set $\boldsymbol{S}_{\text{train}}$ of size n for training. To rigorously evaluate the theoretical quantities, we generate a large, independent test set $\boldsymbol{S}_{\text{test}}$ of size 15000 to approximate the expected objective.

For **Case 1**, the learned $\boldsymbol{\theta}$ is the parameter obtained after the full training duration. For **Case 2**, $\boldsymbol{\theta}$ corresponds to the parameter at each check-pointed iteration step t. For any learned $\boldsymbol{\theta}$, the Suboptimality Gap is estimated by $|J_{\boldsymbol{S}_{\text{test}}}(\pi_{\boldsymbol{\theta}^*}) - J_{\boldsymbol{S}_{\text{test}}}(\pi_{\boldsymbol{\theta}})|$.

I Experiments Extended to Real post-training

To further examine the practical relevance of our theoretical framework, we extend our experiments to real LLM post-training. We focus on two aspects closely related to our analysis. First, we restrict parameter updates to the final output head to examine whether the linearized policy setting can remain effective in practice. Second, we post-train the same model on datasets with different feature coverage to study how training-data coverage affects downstream generalization.

I.1 Output Head-Only Post-Training

Connections to Our Theoretical Analysis. Our theoretical framework is built on Definitions 1 and 2, where a fixed feature representation $\boldsymbol{\phi}(x, a)$ is assumed and learning is performed through a linear parameter $\boldsymbol{\theta}$ acting on this representation. Such a linear structure is convenient for theoretical analysis, but differs from practical LLM post-training, where updating the full model induces a highly nonlinear mapping from model parameters to the output distribution.

To examine whether this linear abstraction can still capture a meaningful training regime in practice, we consider output-head-only post-training. Specifically, we freeze the pretrained backbone and update only the final output head. Under this restriction, the learned representation remains fixed while only the linear readout is optimized, making the resulting training structure more closely aligned with the policy class considered in our analysis.

Experiment Setup. We conduct post-training on Qwen2.5-3B-Instruct (Yang et al., 2024) using Math12K (Lightman et al., 2024) with GRPO (Shao et al., 2024). In the head-only setting, we freeze the entire backbone and update only the final output head. We compare this setting against both the pretrained base model and standard full-parameter GRPO, where all model parameters are updated. We evaluate the resulting models on AIME24, OlympiadBench, Minerva Math, AMC23, Math500, GSM8K, GPQA, MMLU-Pro, and ARC, and report the average performance across all benchmarks.

Results. As shown in Table 1, head-only post-training improves the average performance from 45.98 to 47.11, compared with 47.62 achieved by full-parameter GRPO. These results show that substantial post-training gains can be obtained while keeping the pretrained backbone fixed and

Table 2: Post-training performance with different training-data coverage.

Model	AIME	Olymp.	Minerva	AMC23	Math500	GSM8K	GPQA	MMLU-P	ARC	Avg.
Qwen2.5-3B	6.56	24.74	23.90	34.53	61.80	82.41	18.60	32.14	74.92	39.96
+High Coverage data	7.29	29.63	27.57	41.17	66.60	84.99	18.40	44.82	80.60	**44.56**
+Low Coverage data	8.85	29.04	25.74	37.19	64.40	83.32	16.20	43.75	79.60	43.12
LLaMA3.1-8B	6.77	16.30	22.79	25.70	49.20	85.97	15.40	48.39	83.95	39.39
+High Coverage data	9.38	20.00	27.94	26.56	49.60	86.96	20.00	47.68	83.28	**41.26**
+Low Coverage data	6.04	14.81	25.74	26.09	50.20	87.04	18.20	50.36	83.61	40.23

optimizing only the linear output head. Moreover, these results suggest that the linear parameterization with a fixed representation assumed in our theoretical analysis remains practically meaningful for LLM post-training.

I.2 Post-Training with Different Data Coverage

Connections to Our Theoretical Analysis. Our key feature assumption (Assumption 5) formalizes the fact that, with finite training data, the subspace spanned by the empirical feature covariance matrix $\boldsymbol{V}_{\boldsymbol{S}}$ may not fully cover the entire feature space. As the collected data become more diverse, however, this subspace can provide broader feature coverage, reducing the residual component of an unseen feature that lies outside the covered subspace. Our theoretical results, e.g., Theorems 4 and 6, further show that the magnitude of this residual term affects the generalization bound, with a smaller residual leading to better generalization guarantees. This suggests that post-training on data with broader feature coverage can improve model generalization.

While the importance of training data is well recognized, our analysis provides a quantitative metric for data selection based on the coverage of the feature space induced by the empirical feature covariance matrix. Based on this metric, we partition Math12K into two equally sized subsets with different levels of feature-space coverage and perform the same post-training procedure on each subset. More experimental details are provided as follows.

Experiment Setup. We construct two subsets of Math12K with the same size but different feature coverage. Concretely, for each prompt x, we sample one response from the pretrained model (i.e., Qwen2.5-3B) and extract the last-layer hidden representation immediately before the output head as its feature $\boldsymbol{\phi}_x$. Given a selected set $\boldsymbol{S}$, we define the regularized empirical covariance matrix as

$$\boldsymbol{V}_{\boldsymbol{S}} = \lambda I + \sum_{x \in \boldsymbol{S}} \boldsymbol{\phi}_x \boldsymbol{\phi}_x^\top,$$

and measure the coverage contribution of an unselected prompt x using the ellipsoidal bonus

$$b(x \mid \boldsymbol{S}) = \boldsymbol{\phi}_x^\top \boldsymbol{V}_{\boldsymbol{S}}^{-1} \boldsymbol{\phi}_x.$$

We then greedily select the prompt with the largest bonus,

$$x^\star = \arg\max_{x \notin \boldsymbol{S}} b(x \mid \boldsymbol{S}),$$

which favors features lying along directions that are less covered by the currently selected data. We repeat this procedure until 6K prompts are selected to form the high-coverage subset, while the remaining prompts constitute the low-coverage subset. We then separately perform post-training on the two subsets under identical training settings. We conduct experiments on Qwen2.5-3B and LLaMA3.1-8B (Grattafiori et al., 2024), respectively.

Results. As shown in Table 2, post-training on the high-coverage subset achieves better average performance than using the low-coverage subset on both models. Specifically, the average performance improves from 43.12 to 44.56 on Qwen2.5-3B and from 40.23 to 41.26 on LLaMA3.1-8B. These results indicate that, under the same training-data budget and training configuration, data with broader feature coverage can lead to better post-training generalization. This empirical observation is consistent with our theoretical analysis, where broader feature coverage reduces the residual component outside the covered subspace and consequently leads to a tighter generalization guarantee.

J Empirical Validation of Assumptions 4 and 5

J.1 Discussion on Assumption 4

To examine the linear-independence condition in Assumption 4, we randomly sample 2,000 prompts from Math12K and independently generate 64 responses for each prompt. For each response, we extract the last-layer hidden representation immediately before the output head as its feature $\boldsymbol{\phi}(x, a)$ on different models. Then we construct a 64×2048 feature matrix for each prompt, whose rows correspond to the features of the sampled responses. We then compute the rank of each feature matrix.

Results. For Qwen2.5-3B, **1,997 out of 2,000 matrices (**99.85%**)** are full rank. We further repeat the same experiment on LLaMA3.1-8B-Instruct, where all **2,000 matrices (**100%**)** are full rank. These results provide empirical support for the linear-independence condition in Assumption 4 on real LLMs.

J.2 Discussion on Assumption 5

To examine the feature-coverage condition in Assumption 5, we extract response features from Qwen2.5-3B and construct the empirical covariance matrix $\boldsymbol{V_S}$ using training sets of different sizes n. For each n and corresponding empirical covariance matrix $\boldsymbol{V_S}$, we evaluate unseen features on a held-out test set and compute their residual components relative to the column space $\mathcal{C}(\boldsymbol{V_S})$.

Results. We report the 95th-percentile residual norm over the test samples. As the training-set size increases, the residual decreases **from** 0.14 **at** $n = 64$ **to** 0.001 **at** $n = 128$ **and** 0.00087 **at** $n = 4096$. This decreasing trend empirically supports the feature-coverage behavior characterized in Assumption 5.

Discussion on the Residual Decay Rate. We emphasize that the exact $\epsilon_n = \mathcal{O}(n^{-1})$ decay rate in Assumption 5 is not essential to our analysis. Our proofs only require the residual term $\boldsymbol{r_S}$ to decrease with the sample size n, while a slower decay would lead to a correspondingly slower bound, with the residual-dependent term scaling as $\mathcal{O}(\sqrt{\epsilon_n})$. The residual is more naturally interpreted as a coverage or covariance-approximation error arising from the gap between the empirical covariance $\boldsymbol{V_S}$ and the population covariance $\mathbb{E}_x[\boldsymbol{V}_x]$. Under general matrix concentration, one can expect at least $\mathcal{O}(n^{-1/2})$-scale control, while faster decay, such as $\mathcal{O}(n^{-1})$, can arise under stronger low-variance conditions. We therefore adopt the $\mathcal{O}(n^{-1})$ rate primarily for a cleaner statement of our theoretical results, rather than because our analysis is restricted to this particular regime.

K Some Lemmas on Algebra.

Lemma 20. *Let $\mathbb{U}$ and $\mathbb{V}$ be any two subspaces of the vector space $\mathbb{R}^d$. Then, the orthogonal complement of their intersection, $(\mathbb{U} \cap \mathbb{V})^\perp$, is equal to the sum of their individual orthogonal complements, $\mathbb{U}^\perp$ and $\mathbb{V}^\perp$. That is:*

$$\mathbb{U}^\perp \cap \mathbb{V}^\perp = (\mathbb{U} + \mathbb{V})^\perp.$$

Proof. We prove the equality by showing both inclusions. First, we show that $(\mathbb{U}+\mathbb{V})^\perp \subseteq \mathbb{U}^\perp \cap \mathbb{V}^\perp$. Clearly we have $\mathbb{U} \subseteq \mathbb{U} + \mathbb{V}$, it is straightforward to show that

$$(\mathbb{U} + \mathbb{V})^\perp \subseteq \mathbb{U}^\perp,$$

similarly, we also have

$$(\mathbb{U} + \mathbb{V})^\perp \subseteq \mathbb{V}^\perp,$$

combining these two conditions, we have

$$(\mathbb{U} + \mathbb{V})^\perp \subseteq \mathbb{U}^\perp \cap \mathbb{V}^\perp.$$

On the other hand, we will show that $\mathbb{U}^\perp \cap \mathbb{V}^\perp \subseteq (\mathbb{U}+\mathbb{V})^\perp$. Let x be an arbitrary vector in $\mathbb{U}^\perp \cap \mathbb{V}^\perp$, which means that for any $y_1 \in \mathbb{U}$ and $y_2 \in \mathbb{V}$, we have

$$\langle x, y_1 \rangle = \langle x, y_2 \rangle = \langle x, y_1 + y_2 \rangle = 0.$$

Since y_1 and y_2 are chosen arbitrarily, we have $x \in \mathbb{U} + \mathbb{V}$. Therefore, we can obtain that

$$\mathbb{U}^\perp \cap \mathbb{V}^\perp \subseteq (\mathbb{U} + \mathbb{V})^\perp.$$

This completes the proof of the lemma.

□

Lemma 21. *Let $A, B \in \mathbb{R}^{n\times n}$ be two positive semi-definite matrices, then it holds that*

$$N(A + B) = N(A) \cap N(B). \tag{93}$$

Proof. We prove the equality by showing both inclusions. First, we show that $\mathrm{N}(A + B) \subseteq \mathrm{N}(A) \cap \mathrm{N}(B)$. Let $\boldsymbol{x}$ be an arbitrary vector in $\mathrm{N}(A + B)$, which means that $\boldsymbol{x}$ satisfies

$$(A + B)\boldsymbol{x} = 0,$$

then we have

$$\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{x}^\top B\boldsymbol{x} = \boldsymbol{x}^\top (A + B)\boldsymbol{x} = 0,$$

moreover, given that A and B are both positive semi-definite, it follows that

$$\boldsymbol{x}^\top A\boldsymbol{x} = \boldsymbol{x}^\top B\boldsymbol{x} = 0,$$

it is straightforward to know that $\boldsymbol{x}$ belonging to both $\mathrm{N}(A)$ and $\mathrm{N}(B)$, i.e., $\boldsymbol{x} \in \mathrm{N}(A) \cap \mathrm{N}(B)$. Therefore, we conclude that

$$\mathrm{N}(A + B) \subseteq \mathrm{N}(A) \cap \mathrm{N}(B).$$

On the other hand, we will show that $\mathrm{N}(A) \cap \mathrm{N}(B) \subseteq \mathrm{N}(A + B)$. Similarly, for an arbitrary vector $\boldsymbol{x} \in \mathrm{N}(A) \cap \mathrm{N}(B)$, we have

$$\boldsymbol{x}^\top (A + B)\boldsymbol{x} = \boldsymbol{x}^\top A\boldsymbol{x} = \boldsymbol{x}^\top B\boldsymbol{x} = Ax = B\boldsymbol{x} = 0.$$

Therefore, it holds that $(A + B)\boldsymbol{x} = 0$, i.e. $\boldsymbol{x} \in \mathrm{N}(A + B)$. Therefore, we conclude that

$$\mathrm{N}(A) \cap \mathrm{N}(B) \subseteq \mathrm{N}(A + B).$$

This completes the proof of the lemma. □

Lemma 22. *Let $A, B \in \mathbb{R}^{n\times n}$ be two positive semi-definite matrices, it holds that*

$$\mathrm{C}(A + B) = \mathrm{C}(A) + \mathrm{C}(B). \tag{94}$$

Proof. For the positive semi-definite matrix $A + B$, we have

$$\mathrm{C}(A + B) = \mathrm{C}((A + B)^\top) = \mathrm{N}(A + B)^\perp,$$

where the first equality holds because the matrix $A + B$ is symmetric, while the second equality follows from the Fundamental Theorem of Linear Algebra. Combining with lemma20 and lemma21, it holds that

$$\mathrm{C}(A+B) = \mathrm{N}(A+B)^\perp = (\mathrm{N}(A)\cap\mathrm{N}(B))^\perp = \mathrm{N}(A)^\perp+\mathrm{N}(B)^\perp = \mathrm{C}(A^\top)+\mathrm{C}(B^\top) = \mathrm{C}(A)+\mathrm{C}(B),$$

which proves our conclusion. □

Lemma 23. *Given a positive vector $\boldsymbol{p} = (p_1, \cdots, p_n)^\top \in \mathbb{R}^n$ that satisfies $\sum_{i=1}^n p_i = 1$, denote $C_{\boldsymbol{p}}$ as:*

$$C_{\boldsymbol{p}} = \mathsf{diag}\{p_1, \cdots, p_n\} - \boldsymbol{p}\boldsymbol{p}^\top.$$

Then we have that the rank of the matrix $C_{\boldsymbol{p}}$ is $n-1$, and the column space of $C_{\boldsymbol{p}}$ can be demonstrated as

$$\mathrm{C}(C_{\boldsymbol{p}}) = \{\boldsymbol{w} | \boldsymbol{w}^\top \boldsymbol{v} = 0, \boldsymbol{v} = (1, \cdots, 1)^\top\}.$$

Proof. Let vector $\boldsymbol{v}$ be any vector falls in the null space of matrix $C_{\boldsymbol{p}}$, then we have

$$0 = C_{\boldsymbol{p}}\boldsymbol{v} = \left(\mathsf{diag}\{p_1, \cdots, p_n\} - \boldsymbol{p}\boldsymbol{p}^\top\right)\boldsymbol{v} = \mathsf{diag}\{p_1, \cdots, p_n\}\boldsymbol{v} - (\boldsymbol{p}^\top\boldsymbol{v})\boldsymbol{p},$$

then for any $i, j \in \{1, 2, \cdots, n\}$, it holds that

$$0 = p_i v_i - (\boldsymbol{p}^\top\boldsymbol{v})p_i = p_j v_j - (\boldsymbol{p}^\top\boldsymbol{v})p_j,$$

therefore, it holds that

$$v_i = (\boldsymbol{p}^\top\boldsymbol{v}) = v_j,$$

for vector $\boldsymbol{p}$ is a positive vector. It follows that the null space of matrix $C_{\boldsymbol{p}}$ is $\mathrm{N}(C_{\boldsymbol{p}}) = \{\boldsymbol{v}|\boldsymbol{v} = k(1, \cdots, 1)^\top, k \in \mathbb{R}\}$. According to the Fundamental Theorem of Linear Algebra, we can further express the row space of $C_{\boldsymbol{p}}$ as

$$\mathrm{C}(C_{\boldsymbol{p}}^\top) = \{\boldsymbol{w}|\boldsymbol{w}^\top\boldsymbol{v} = 0, \boldsymbol{v} = (1, \cdots, 1)^\top\}.$$

Furthermore, since $C_{\boldsymbol{p}}$ is a symmetric matrix, its row space and column space are identical, which proves the lemma. □

Lemma 24. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix with full column rank and $C_{\boldsymbol{p}}$ be a matrix satisfying Lemma 23, where $m > n$. Then, the column spaces of matrices $AC_{\boldsymbol{p}}A^\top$ and $AC_{\boldsymbol{p}}$ are identical, and both can be expressed in the form of $\{A\boldsymbol{w}|\boldsymbol{w}^\top\boldsymbol{v} = 0, \boldsymbol{v} = (1, \cdots, 1)^\top\}$, i.e.*

$$\mathrm{C}(AC_{\boldsymbol{p}}A^\top) = \mathrm{C}(AC_{\boldsymbol{p}}) = \{A\boldsymbol{w}|\boldsymbol{w}^\top\boldsymbol{v} = 0, \boldsymbol{v} = (1, \cdots, 1)^\top\}.$$

Proof. We begin by showing that $\mathrm{C}(AC_{\boldsymbol{p}}A^\top)$ is contained in $\mathrm{C}(AC_{\boldsymbol{p}})$. For any $\boldsymbol{y} \in \mathrm{C}(AC_{\boldsymbol{p}}A^\top)$, there exists a $\boldsymbol{u}$ such that

$$\boldsymbol{y} = AC_{\boldsymbol{p}}A^\top \cdot \boldsymbol{u} = AC_{\boldsymbol{p}} \cdot (A^\top\boldsymbol{u}),$$

thus it is clear that we also have $\boldsymbol{y} \in \mathrm{C}(AC_{\boldsymbol{p}})$.

On the other hand, for any $\boldsymbol{y} \in \mathrm{C}(AC_{\boldsymbol{p}})$, we can find a $\boldsymbol{v}$ such that

$$\boldsymbol{y} = AC_{\boldsymbol{p}} \cdot \boldsymbol{v}.$$

Since matrix A has full column rank, it follows that there exists a vector $\boldsymbol{u}$ such that

$$A^\top\boldsymbol{u} = \boldsymbol{v}.$$

We have thus far proven that $\mathrm{C}(AC_{\boldsymbol{p}}A^\top) = \mathrm{C}(AC_{\boldsymbol{p}})$. Furthermore, using Lemma 23, it is straightforward to show that

$$\mathrm{C}(AC_{\boldsymbol{p}}) = \{A\boldsymbol{w}|\boldsymbol{w}^\top\boldsymbol{v} = 0, \boldsymbol{v} = (1, \cdots, 1)^\top\},$$

which proves our conclusion. □