A Shared Low-Rank Adaptation Approach to Personalized RLHF

Renpu Liu **Peng Wang** **Donghao Li** **Cong Shen** **Jing Yang***
University of Virginia
{pzw7bx, pw7nc, maj3qx, cong, yangjing}@virginia.edu

Abstract

Reinforcement Learning from Human Feedback (RLHF) has emerged as a pivotal technique for aligning artificial intelligence systems with human values, achieving remarkable success in fine-tuning large language models. However, existing RLHF frameworks often assume that human preferences are relatively homogeneous and can be captured by a single, unified reward model. This assumption overlooks the inherent diversity and heterogeneity across individuals, limiting the adaptability of RLHF to personalized scenarios and risking misalignments that can diminish user satisfaction and trust in AI systems. In this paper, we address these challenges by introducing Low-Rank Adaptation (LoRA) into the personalized RLHF framework. We apply LoRA in the the aggregated parameter space of all personalized reward functions, thereby enabling efficient learning of personalized reward models from potentially limited local datasets. Our approach exploits potential shared structures among the local ground-truth reward models while allowing for individual adaptation, without relying on restrictive assumptions about shared representations as in prior works. We further establish sample complexity guarantees for our method. Theoretical analysis demonstrates the effectiveness of the proposed approach in capturing both shared and individual-specific structures within heterogeneous human preferences, addressing the dual challenge of personalization requirements and practical data constraints. Experimental results on real-world datasets corroborate the efficiency of our algorithm in the personalized RLHF setting.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).
* Corresponding author.

1 Introduction

The rapid development and widespread use of Large Language Models (LLMs) have transformed fields like natural language processing, content generation, and human-computer interaction. Models such as GPT-4 (Achiam et al., 2023), BERT (Devlin et al., 2019), and their successors have exhibited remarkable capabilities in understanding and generating text, enabling applications ranging from automated customer service to advanced creative tools. This "boom" of LLMs has not only broadened AI's potential but also underscored the critical need to ensure that these models align with human values and preferences.

To help this alignment, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2023) plays a key role as a fine-tuning method of LLMs. This method ensures that the generated responses are contextually appropriate and aligned with ethical and social norms (Ouyang et al., 2022). By incorporating human feedback into the fine-tuning process, RLHF bridges the gap between the raw generative power of LLMs and the requirements of real-world applications, improving the quality and safety of AI-generated content.

Current RLHF frameworks, such as Bai et al. (2022); Wang et al. (2024a), essentially assume that human preferences are relatively homogeneous and can be effectively captured by a single, unified reward model. This simplification overlooks the inherent diversity and heterogeneity in human preferences, which can vary significantly across individuals. Such an oversimplification limits the adaptability of RLHF to personalized scenarios and risks, introducing misalignments that could diminish user satisfaction and trust in AI systems. A straightforward approach to handling heterogeneous human preferences is learning personalized reward functions for each labeler using traditional RLHF methods, such as Ouyang et al. (2022). However, this method faces a significant challenge: preference data from individual users may be insufficient to construct accurate reward models for each human labeler. Recently, several studies have proposed empirical methods to address this challenge. For example, Li et al. (2024) introduced a personalized direct preference optimization method within

the personalized RLHF framework. Similarly, Poddar et al. (2024) presented a class of multi-modal RLHF methods that infer user-specific latent variables and then learn personalized reward models conditioned on them. In addition to empirical approaches, some works have provided methods with theoretical guarantees. Specifically, Zhong et al. (2024) conducted a theoretical analysis assuming that human reward functions are linear with shared representations. Extending this line of work, Park et al. (2024) considered a more general setting where the representation function is a general (nonlinear) function of the feature mapping.

On the other hand, since first introduced by Hu et al. (2021), Low-Rank Adaptation (LoRA) has quickly become a prominent method for fine-tuning LLMs to reduce the number of trainable parameters and prevent overfitting (Houlsby et al., 2019; Huang et al., 2023). Some recent works have proposed to combine RLHF with LoRA to enhance the fine-tuning of LLMs using human feedback. For instance, researchers have explored integrating LoRA into the RLHF framework to efficiently incorporate human preferences while maintaining model performance (Santacroce et al., 2023; Sun et al., 2023; Sidahmed et al., 2024). However, these approaches primarily focus on general adaptation and do not address the challenges of heterogeneous feedback from diverse users.

In this paper, we address the challenges of personalized RLHF by introducing personalized LoRA with a shared component into the personalized RLHF framework. By leveraging LoRA, we effectively learn individual reward models that capture human users' heterogeneous preferences with limited data. To the best of our knowledge, LoRA has not been previously explored in the context of personalized RLHF, making our approach a novel contribution to the field. Our major contributions are summarized as follows:

- We propose an algorithm named Personalized LoRA with Shared Component (`P-ShareLoRA`) for RLHF, which leverages the shared components of LoRA modules to learn the personalized reward functions efficiently. Rigorous theoretical analysis demonstrates that `P-ShareLoRA` can effectively reduce sample complexity, compared with both the full-parameter fine-tuning method and the standard LoRA method without parameter sharing. To the best of our knowledge, this is the first work that theoretically demonstrates the benefits of LoRA with shared components in RLHF.
- Unlike existing analytical frameworks for personalized RLHF which typically enforce strict constraints on the reward model structures, such as linear representations (Zhong et al., 2024) or shared representations with linear heads (Park et al., 2024), we develop novel technical approaches to address the challenges from the unstructured reward functions. Specifically, we propose a new Lagrange remainder-based method that allows us to prove that LoRA modules with shared components can approximate the optimal low-rank structure of the ground truth parameter matrix. Building on this, we further prove an upper bound on the distance between the optimal reward function and the learned reward function with shared parameters. The theoretical results demonstrate that the expected return under the policies derived with the learned reward functions are near-optimal (up to a bias term related to the preference diversity among users).
- Experiments on the Reddit TL;DR dataset (Stiennon et al., 2020) validate the effectiveness of the proposed approach. Specifically, our approach achieves a prediction accuracy of 74.65% on Llama-3 8B and 66.93% on GPT-J 6B, which outperforms the SOTA algorithms that achieve 73.25% on Llama-3 8B and 66.13% on GPT-J 6B, respectively. Those empirical results corroborate our theory, demonstrating the advantage of LoRA with shared components for personalized RLHF.

2 Related Works

Reinforcement Learning from Human Feedback. Reinforcement Learning from Human Feedback (RLHF) has demonstrated considerable success across various practical applications, especially in aligning AI models with human values and preferences. One of the most prominent applications of RLHF is in fine-tuning large language models, as exemplified by OpenAI's ChatGPT (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023). Additionally, RLHF has been explored in computer vision tasks (Lee et al., 2023; Xu et al., 2024). Furthermore, RLHF has been widely adopted in domains that involve high-risk decision-making, such as healthcare (Yu et al., 2021), robotics (Abramson et al., 2022; Hwang et al., 2024; Thumm et al., 2024), and autonomous driving (Wu et al., 2023; Chen et al., 2023), where alignment with human preferences is critical for ensuring safety and addressing ethical considerations.

From a theoretical standpoint, studies of RLHF have garnered increasing research interest. Zhu et al. (2023) examine the Bradley-Terry-Luce model (Bradley & Terry, 1952) within the context of a linear reward framework, while Zhan et al. (2023) extend these results to more general classes of reward functions. Similarly, Li et al. (2023) introduce a pessimistic algorithm that is provably efficient for dynamic discrete choice models. All these works focus on settings with offline preference data. In the online setting, Xu et al. (2020) and Pacchiano et al. (2021) study tabular online RLHF. Wang et al. (2024a) theoretically demonstrate that preference-based RL can be directly addressed using existing reward-based RL algorithms by utilizing a preference-to-reward model. Xiong et al. (2024) present a provable iterative Direct Preference Optimization (DPO)

algorithm for online settings. Ye et al. (2024) provide a theoretical analysis of RLHF under a general preference oracle, proposing sample-efficient algorithms for both offline and online settings.

Some recent studies have extended RLHF to personalized alignment for diverse user groups and individuals. Zhao et al. (2023) introduce Group Preference Optimization (GPO), which addresses group-level heterogeneity through a mixture of shared and personalized architectures. Additionally, Ramesh et al. (2024) propose Group Robust Preference Optimization (GRPO), a reward-free RLHF framework that handles heterogeneous preferences by optimizing for worst-case group outcomes. Beyond group-level alignment, other works focus on individual personalization. For instance, Li et al. (2024) develop a Personalized RLHF method that jointly learns a lightweight user model alongside the policy model to capture each user's unique preferences, leading to responses more closely aligned with individual tastes than non-personalized RLHF. Besides, Poddar et al. (2024) introduce a variational latent preference framework that infers a user-specific latent variable on which both the learned reward model and the policy rely.

In addition to the empirical studies, recent works have also established theoretical guarantees for personalized RLHF. Siththaranjan et al. (2023) show that traditional RLHF models that implicitly aggregate preferences can lead to undesirable outcomes. They introduce Distributional Preference Learning (DPL) to mitigate this issue. Chakraborty et al. (2024) group individual reward models into distinct subsets and propose a MaxMin alignment objective inspired by Egalitarian principles. Zhong et al. (2024) investigate a setting where local optimal reward functions share a linear representation combined with personalized linear heads, theoretically demonstrating that aggregating multiple preferences across different parties can overcome the shortcomings of traditional RLHF that only learn a single reward function. Building on this, Park et al. (2024) generalize the reward function model of Zhong et al. (2024) by introducing a general representation function combined with personalized linear heads.

Low-Rank Adaptation (LoRA). The rapid scaling of pre-trained language models has led to significant challenges in fine-tuning these models for downstream tasks due to the substantial computational and storage requirements. To address this, Low-Rank Adaptation (LoRA) has been proposed as an efficient fine-tuning approach (Hu et al., 2021). The vanilla LoRA keeps the original model weights frozen and injects trainable low-rank matrices into each layer of the Transformer architecture. This strategy dramatically reduces the number of trainable parameters and computational overhead, making it feasible to adapt large models on limited hardware resources (Valipour et al., 2022; Zhang et al., 2023; Kopiczko et al., 2023; Dettmers et al., 2024; Hayou et al., 2024; Liu et al., 2024b).

Recently, several studies have focused on implementing LoRA in multi-task settings. Huang et al. (2023) introduce LoraHub, which enables the composition and sharing of LoRA modules trained on diverse tasks. Luo et al. (2024) consider LoRA as a Mixture of Experts (MoE), treating these small adaptation modules as experts focusing on unique aspects. Shen et al. (2024) introduce MixLoRA, treats LoRA modules as experts and uses a dynamic factor selection method to select modules for combination. Tang et al. (2023) propose partial linearization, where they linearize only the adapter modules—the parts adjusted during fine-tuning—and apply "task arithmetic" to combine these linearized adapters from different tasks. In the federated learning setting, Wang et al. (2024b) introduces a stacking-based aggregation technique for LoRA adapters, enabling efficient fine-tuning across clients.

To effectively learn LoRA modules in a multi-task setting, some recent studies consider sharing partial parameters among different tasks or clients. Sun et al. (2024) introduce FFA-LoRA, which keeps one of the LoRA modules fixed while updating only the other during local training. Similarly, Kuo et al. (2024) propose a method in which certain parameters within the locally downloaded LoRA modules remain unchanged, while the rest are updated. HydraLoRA (Tian et al., 2024) extends this idea by incorporating LoRA modules with a shared low-rank matrix in a Mixture-of-Experts (MoE) framework. Additionally, FedSA-LoRA (Guo et al., 2024) observes that in a federated learning setup, one transformation matrix primarily captures generalizable knowledge, while the other learns client-specific adaptations. Building on this insight, they employ a hybrid approach that combines shared global components with personalized local updates. *To the best of our knowledge, the theoretical implications of using shared LoRA parameters in RLHF remain unexplored.*

3 Problem Formulation

Notation. Bold uppercase letters (e.g., $\mathbf{X}$) denote matrices. The function $\mathrm{diag}(x_1, \ldots, x_d)$ represents a $d \times d$ diagonal matrix with diagonal entries $x_1, \ldots, x_d$. The inner product of vectors x and y is denoted by $\langle x, y \rangle$, and the Euclidean norm of a vector x is represented by $\|x\|_2$. For a matrix $\mathbf{X}$, the operator (spectral) norm is denoted by $\|\mathbf{X}\|_2$, and its Frobenius norm by $\|\mathbf{X}\|_F$. The k-th largest singular value of $\mathbf{X}$ is denoted by $\sigma_k(\mathbf{X})$. For a matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, we use $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{d_1 d_2}$ to denote the vector obtained by column-wise vectorizing $\mathbf{X}$, i.e., $\mathrm{vec}(\mathbf{X})^\top = [x_1^\top, \ldots, x_{d_2}^\top]$, where x_i is the i-th column of $\mathbf{X}$. The identity matrix of size $d \times d$ is denoted by $\mathbf{I}_d$.

Markov Decision Processes. We consider the tabular finite-horizon Markov Decision Process (MDP) to

model the Reinforcement Learning from Human Feedback (RLHF) setting with N human labelers (or users), each with their own reward function. A MDP $\mathcal{M}$ is represented by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, (P_h)_{h\in[H]}, \mathbf{r} = (r_i)_{i\in[N]})$, where $\mathcal{S}$ is the set of states, defined as all possible prompts or questions; $\mathcal{A}$ is the set of actions, representing potential answers or responses to these questions; H denotes the length of the horizon; $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the state transition probability at step $h \in [H]$, with $\Delta(\mathcal{S})$ being the set of probability distributions over $\mathcal{S}$; and $r_i : \mathcal{T} \mapsto [-R, R]$ is the reward function for each individual $i \in [N]$, where $\mathcal{T} := (\mathcal{S} \times \mathcal{A})^H$ denotes the set of all possible trajectories $\tau = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H)$. The MDP concludes at an absorbing termination state with zero reward after H steps. A policy is defined as a sequence $\pi = (\pi_h)_{h=1}^H$, where each $\pi_h : (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta(\mathcal{A})$ maps the history and current state to a distribution over actions at step h. The expected cumulative reward of a policy π for individual i is given by $J(\pi; r_i) := \mathbb{E}_{\tau\sim\pi}[r_i(\tau)]$.

Relationship between Preference and Reward Functions. Given two trajectories τ_0 and τ_1, we introduce a random variable $o \in \{0, 1\}$ to represent the preference outcome: We set $o = 1$ if $\tau_0 \succ \tau_1$ (i.e., τ_0 is preferred over τ_1), and $o = 0$ if $\tau_0 \prec \tau_1$ (i.e., τ_1 is preferred over τ_0). We model the probability that individual $i \in [N]$ prefers τ_0 over τ_1 as $P_{r_i}(o = 1 \mid \tau_0, \tau_1) = \Phi\big(r_i(\tau_0) - r_i(\tau_1)\big)$, where $\Phi : \mathbb{R} \to [0, 1]$ is a monotonically increasing function satisfying $\Phi(x) + \Phi(-x) = 1$ and $\log \Phi(x)$ is a Lipschitz continuous and strongly convex function. A common choice for Φ is the sigmoid function $\sigma(x) = 1/(1 + e^{-x})$, which maps real-valued inputs to the range $[0, 1]$. This function corresponds to the Bradley-Terry-Luce (BTL) model, which is commonly used to model the relationship between preferences and rewards. We define the *preference probability vector* induced by the reward functions $\boldsymbol{r}$ as $P_{\boldsymbol{r}}(o \mid \tau_0, \tau_1) = (P_{r_1}(o \mid \tau_0, \tau_1), \ldots, P_{r_N}(o \mid \tau_0, \tau_1))^\top$, where $P_{\boldsymbol{r}}$ represents the collective preference probabilities across all individuals, and P_{r_i} denotes the preference probability induced by the reward function r_i for individual i.

Personalized Reward Functions. We consider the naturally diverse individual human preferences and aim to learn personalized reward models for each individual. As a first step, we assume each reward function r_i is parameterized by $\boldsymbol{\Theta}_i \in \mathbb{R}^{d_1\times d_2}$, and we denote it as $r_{\boldsymbol{\Theta}_i} : \mathcal{T} \to \mathbb{R}$. We denote the aggregated reward vector as $\mathbf{r}_{\boldsymbol{\Theta}} := \big(r_{\boldsymbol{\Theta}_1}, \ldots, r_{\boldsymbol{\Theta}_N}\big)^\intercal$, where $\boldsymbol{\Theta} \in \mathbb{R}^{d_1\times Nd_2}$ is the aggregated parameter matrix defined by $\boldsymbol{\Theta} = \big[\boldsymbol{\Theta}_1, \ \ldots, \ \boldsymbol{\Theta}_N\big]$.

Let θ denote the column-wise vectorization of $\boldsymbol{\Theta}$, i.e., $\theta = \mathrm{vec}(\boldsymbol{\Theta})$. Then, we make the following assumption.

Assumption 1. *For any trajectory τ, the reward function $r_{\boldsymbol{\Theta}}(\tau)$ satisfies Lipschitz continuity $\|\nabla_\theta r_{\boldsymbol{\Theta}}(\tau)\| \le L_1$ and Lipschitz smoothness $\|\nabla^2_\theta r_{\boldsymbol{\Theta}}(\tau)\| \le L_2$ for $L_1, L_2 > 0$.*

Note that the gradient operator ∇ and the Laplacian ∇^2 are applied to the vectorized parameter matrix θ. Assumption 1 is a standard assumption similar to those in related RLHF studies, such as Zhu et al. (2023).

Define the set of valid parameters for the reward function as

$$\mathcal{S} := \Big\{\boldsymbol{\Theta} \,\Big|\, \boldsymbol{\Theta}_i \in \mathbb{R}^{d_1\times d_2},\ \|\boldsymbol{\Theta}_i\|_F \le B,\ \forall i \in [N]\Big\}, \tag{3.1}$$

and the corresponding class of reward functions as

$$\mathcal{G}_{\mathbf{r}}(\mathcal{S}) = \Big\{\big(r_{\boldsymbol{\Theta}_i}(\cdot)\big)_{i\in[N]} \,\Big|\, \boldsymbol{\Theta} \in \mathcal{S}\Big\}. \tag{3.2}$$

The boundedness condition $\|\boldsymbol{\Theta}_i\|_F \le B$ (B is a positive constant) in Equation (3.1), together with Assumption 1, ensures that the reward function is bounded, which is a standard assumption adopted in related works (Zhan et al., 2023; Zhong et al., 2024).

Throughout this paper, we let $\mathbf{r}^\star = \big(r_1^\star, \ldots, r_N^\star\big)$ denote the underlying true human reward functions with corresponding ground truth parameters $\boldsymbol{\Theta}^* = [\boldsymbol{\Theta}_1^*, \cdots, \boldsymbol{\Theta}_N^\star]$. We assume that $\mathbf{r}^\star \in \mathcal{G}_{\mathbf{r}}(\mathcal{S})$ to ensure that the true reward functions are within the considered function class.

Learning Personalized Reward Functions via LoRA. Motivated by LoRA that is widely adopted for the fine-tuning of LLMs (Sidahmed et al., 2024), we assume the system starts from initialized reward model parameters $\boldsymbol{\Theta}^{\text{init}} = [\boldsymbol{\Theta}_1^{\text{init}}, \cdots, \boldsymbol{\Theta}_N^{\text{init}}]$. Denote the low-rank adaptation matrice for the reward models as $\Delta\boldsymbol{\Theta} = [\Delta\boldsymbol{\Theta}_1, \cdots, \Delta\boldsymbol{\Theta}_N]$. Then, after the adaptation, the set of valid parameters for the personalized reward model becomes

$$\begin{aligned}\mathcal{S}^{\text{LoRA}} = \Big\{\boldsymbol{\Theta} \,\Big|\, &\boldsymbol{\Theta} = \boldsymbol{\Theta}^{\text{init}} + \Delta\boldsymbol{\Theta}, \mathrm{rank}(\Delta\boldsymbol{\Theta}_i) \le k,\\ &\|\Delta\boldsymbol{\Theta}_i\|_F \le B, \forall i \in [N]\Big\}.\end{aligned} \tag{3.3}$$

Note that the LoRA module is typically represented in a low-rank factorization form, i.e., as the product of two lower-dimensional matrices: $\Delta\boldsymbol{\Theta}_i = \mathbf{B}_i\mathbf{W}_i$, where $\mathbf{B}_i \in \mathbb{R}^{d_1\times k}$ and $\mathbf{W}_i \in \mathbb{R}^{k\times d_2}$. In the function class $\mathcal{G}_{\mathbf{r}|\boldsymbol{\Theta}^{\text{init}}}$, the individual LoRA modules $\Delta\boldsymbol{\Theta}_i$ are independent. To leverage potential common structures among the individual LoRA modules, as observed in recent works (Zhu et al., 2024; Guo et al., 2024; Tian et al., 2024), we assume that the $\mathbf{B}_i$ matrices are shared across all users, i.e., $\mathbf{B}_i = \mathbf{B}$ for all i. Under this constraint, the aggregated matrix $\Delta\boldsymbol{\Theta}$ can be expressed as $\Delta\boldsymbol{\Theta} = \mathbf{B}[\mathbf{W}_1, \cdots, \mathbf{W}_N]$, which implies that $\Delta\boldsymbol{\Theta}$ becomes a low-rank matrix with $\mathrm{rank}(\Delta\boldsymbol{\Theta}) \le k$, since $\mathrm{rank}(\mathbf{B}) \le k$. Consequently, when $\mathbf{B}$ is shared across all LoRA modules, the parameter set is equivalent

to:

$$\mathcal{S}^{\text{ShareLoRA}} = \Big\{ \boldsymbol{\Theta} \;\Big|\; \boldsymbol{\Theta} = \boldsymbol{\Theta}^{\text{init}} + \Delta\boldsymbol{\Theta}, \operatorname{rank}(\Delta\boldsymbol{\Theta}) \le k, \\ \|\Delta\boldsymbol{\Theta}_i\|_F \le B, \forall i \in [N] \Big\}. \tag{3.4}$$

To leverage the potential common structure among individual LoRA modules, we utilize the parameter set $\mathcal{S}^{\text{ShareLoRA}}$, which allows us to learn LoRA modules with shared parameters across users effectively. This low-rank constraint leverages shared structures among users' preferences, allowing the model to capture common patterns while adapting to individual differences. The aggregated low-rank adaptation $\Delta\boldsymbol{\Theta}$ results in local low-rank adaptations $\{\Delta\boldsymbol{\Theta}_i\}$, which incorporate a shared matrix $\mathbf{B}$ and distinct individual adaptation matrices $\mathbf{W}_i$, i.e., $\Delta\boldsymbol{\Theta}_i = \mathbf{B}\mathbf{W}_i$. Intuitively, the shared matrix $\mathbf{B}$ preserves common directions for parameter updating, while $\mathbf{W}_i$ captures individual adaptation along those dimensions.

Given a collection of preference datasets for individual users, denoted as $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})\}_{j=1}^{N_p}$, our objective is to estimate the ground-truth reward function $\mathbf{r}^\star$ by combining the learned shared-parameter LoRA matrices within $\mathcal{S}^{\text{ShareLoRA}}$. We define the aggregated dataset as $\widehat{\mathcal{D}} = \bigcup_{i=1}^{N} \widehat{\mathcal{D}}_i$, with $|\widehat{\mathcal{D}}_i| = N_p$ for all $i \in [N]$. Our analysis can be extended to scenarios where the dataset sizes vary across individuals, i.e., $|\widehat{\mathcal{D}}_i| = N_{p,i}$ for each i. The optimization problem is then formulated as follows:

$$\max_{\boldsymbol{\Theta} \in \mathcal{S}^{\text{ShareLoRA}}} F\left(\boldsymbol{\Theta}; \widehat{\mathcal{D}}\right) = \sum_{i=1}^{N} \sum_{j=1}^{N_p} \log P_{\boldsymbol{\Theta}_i}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right), \tag{3.5}$$

where we use $P_{\boldsymbol{\Theta}}$ denote $P_{r_{\boldsymbol{\Theta}}}$ to simplify the notation.

Algorithm 1 `P-ShareLoRA` for RLHF

1: **Input:** Dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$; initial parameters $\boldsymbol{\Theta}^{\text{init}}$; reference policy $\mu_{i,\text{ref}}$.

2: Obtain model update $\Delta\widehat{\boldsymbol{\Theta}}$ by solving Equation (3.5) :

$$\Delta\widehat{\boldsymbol{\Theta}} \leftarrow \underset{\Delta\boldsymbol{\Theta}:\boldsymbol{\Theta}\in\mathcal{S}^{\text{ShareLoRA}}}{\arg\max} F\left(\boldsymbol{\Theta}; \widehat{\mathcal{D}}\right)$$

3: Construct confidence sets $\{\mathcal{R}_i\}_{i=1}^N$ by

$$\mathcal{R}_i \leftarrow \left\{ r_{\boldsymbol{\Theta}_i} \;\middle|\; \boldsymbol{\Theta}_i = \boldsymbol{\Theta}_i^{\text{init}} + \Delta\boldsymbol{\Theta}_i, \|\Delta\boldsymbol{\Theta}_i - \Delta\widehat{\boldsymbol{\Theta}}_i\|_F^2 \le \zeta \right\} \tag{3.6}$$

4: Compute policy with respect to $\mathcal{R}_i$ for all $i \in [N]$ by

$$\widehat{\pi}_i \leftarrow \arg\max_{\pi\in\Pi} \min_{r_i\in\mathcal{R}_i} \left(J(\pi; r_i) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i(\tau)]\right) \tag{3.7}$$

5: **Output:** $(\Delta\widehat{\boldsymbol{\Theta}}, (\widehat{\pi}_i)_{i\in[N]})$.

4 Algorithm Design and Analysis

4.1 Algorithm: `P-ShareLoRA` for RLHF

In this section, we present our proposed algorithm, Personalized LoRA with Shared Component (`P-ShareLoRA`) for RLHF, to effectively learn personalized reward functions and compute corresponding policies for each individual user.

The algorithm begins by initializing the reward function for each user i by $\boldsymbol{\Theta}_i^{\text{init}}$. The core of the algorithm involves estimating the personalized reward models by optimizing low-rank adaptations $\Delta\widehat{\boldsymbol{\Theta}}$. Specifically, we obtain $\Delta\widehat{\boldsymbol{\Theta}}$ by solving the optimization problem defined in Equation (3.5).

After obtaining $\widehat{\boldsymbol{\Theta}}$, we construct confidence sets $\{\mathcal{R}_i\}$ for each user's reward function parameters. Each set $\mathcal{R}_i$ is designed to ensure that the distance between the parameter matrix of the reward function and the empirical estimation obtained by solving Equation (3.5) remains within a tolerance level ζ, thereby providing a robust confidence region for the reward functions. Finally, we compute each user's personalized policy $\widehat{\pi}_i$ by solving a robust optimization problem. For each individual $i \in [N]$, we determine the policy that maximizes the difference between its expected cumulative reward $J(\pi; r_i)$ and the expected reward of the reference policy $\mu_{i,\text{ref}}$, evaluated under the worst-case reward function within the confidence set $\mathcal{R}_i$. The algorithm outputs the estimated reward model parameters $\widehat{\boldsymbol{\Theta}}$ and the set of personalized policies $(\widehat{\pi}_i)_{i\in[N]}$. We note that without pessimism (i.e., the confidence set of reward functions reduces to a singleton $\mathcal{R}_i = \{r_{\widehat{\boldsymbol{\Theta}}_i}\}$), the optimization objective simplifies to the vanilla RLHF objective. `P-ShareLoRA` is detailed in Algorithm 1.

4.2 Definitions and Assumptions

Before formally presenting our main theoretical results of Algorithm 1, we introduce the following definitions and assumptions. We start by defining two diversity metrics over human preference on different labelers.

Definition 4.1 (Diversity Metrics)**.** *Given the aggregated ground-truth parameter matrix $\boldsymbol{\Theta}^\star = [\boldsymbol{\Theta}_1^\star, \ldots, \boldsymbol{\Theta}_N^\star]$ and initialization parameter matrices $\{\boldsymbol{\Theta}_i^{\text{init}}\}$, we define the difference matrix $\Delta\boldsymbol{\Theta}^\star = [\Delta\boldsymbol{\Theta}_1^\star, \ldots, \Delta\boldsymbol{\Theta}_N^\star]$, where $\Delta\boldsymbol{\Theta}_i^\star = \boldsymbol{\Theta}_i^\star - \boldsymbol{\Theta}_i^{\text{init}}$ for each user i. Let $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{\min\{d_1, Nd_2\}}$ be the singular values of $\Delta\boldsymbol{\Theta}^\star$. We then define the condition number ν and the summation of tail singular values Σ_{tail} as $\nu = \frac{\sigma_k^2}{N}, \Sigma_{\text{tail}} = \sum_{i=k+1}^{\min\{d_1, Nd_2\}} \sigma_i^2$.*

Remark 1. *The condition number ν, as defined in (Tripuraneni et al., 2021), quantifies the alignment of parameter differences between the ground truth model parameters and the initialization across users. Specifically, it considers the magnitude of the k-th largest singular value of the difference matrix $\Delta\boldsymbol{\Theta}^\star$, normalized by the number of users*

N. Note that due to the constraint in $\mathcal{S}$, for fixed $\mathbf{\Theta}^{\text{init}}$, the bounded total energy of $\mathbf{\Theta}^\star$, i.e., $\|\mathbf{\Theta}^\star\|_F^2 \le NB^2$, implies the total energy of $\Delta\mathbf{\Theta}^\star$ is also bounded. Therefore, a larger ν indicates that the top-k leading singular values are significantly larger than the subsequent ones. This dominance suggests that $\Delta\mathbf{\Theta}_i^\star$ across users are primarily aligned along a few principal directions, indicating low diversity. Conversely, a smaller ν indicates high diversity across different directions.

The tail sum Σ_{tail} measures the total variance not captured by the top k singular values of $\Delta\mathbf{\Theta}^\star$. It is calculated by summing the squares of the singular values from σ_{k+1} onward, quantifying the residual "energy" beyond a rank-k approximation. A smaller Σ_{tail} suggests that the top k singular values capture most of the variance, implying that a low-rank adaptation effectively represents the essential variability among users for accurate modeling of reward functions.

These diversity metrics capture the preference diversity among users. Intuitively, users with similar preferences will be less diverse and could benefit more from a shared LoRA model.

Next, to capture the complexity of the reward function class, we introduce the concept of the bracketing number for reward vectors.

Definition 4.2 (Bracketing Number for Reward Vectors (Park et al., 2024))**.** *For a reward vector $\boldsymbol{r} \in \mathcal{G}_{\boldsymbol{r}}$, an ϵ-bracket is a pair of functions (g_1, g_2) such that for all $(\tau_0, \tau_1) \in \mathcal{T} \times \mathcal{T}$, $\|g_1(\tau_0,\tau_1) - g_2(\tau_0,\tau_1)\|_1 \le \epsilon$, and $g_1(\tau_0,\tau_1) \le P_{\boldsymbol{r}}(\cdot|\tau_0,\tau_1) \le g_2(\tau_0,\tau_1)$. The ϵ-bracketing number of $\mathcal{G}_{\boldsymbol{r}}$, denoted by $\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(\epsilon)$, is the minimal number of ϵ-brackets required to cover all $\boldsymbol{r}$ in $\mathcal{G}_{\boldsymbol{r}}$.*

Definition 4.2 is adapted from the definition of bracketing numbers in Park et al. (2024); Zhan et al. (2023), which captures the complexity of the function class in terms of its parameter dimensions.

We assume a uniform concentration property for the expected Euclidean distance between $r_{\mathbf{\Theta}_1}(\tau_0) - r_{\mathbf{\Theta}_1}(\tau_1)$ and $r_{\mathbf{\Theta}_2}(\tau_0) - r_{\mathbf{\Theta}_2}(\tau_1)$ over the offline data. We note that this expected Euclidean distance can be seen as the distance between two reward functions $r_{\mathbf{\Theta}_1}$ and $r_{\mathbf{\Theta}_2}$ (Zhan et al., 2023), therefore the concentration property ensures that with a sufficiently large sample size N, empirical data reliably approximates these distance for all pairs of reward functions in $\mathcal{G}_r$.

Assumption 2 (Uniform Concentration)**.** *Given distributions μ_0 and μ_1, and two reward functions parameterized by $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$, respectively, we define the expected and empirical squared difference of reward discrepancies as*

$$D_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1) = \mathop{\mathbb{E}}_{\tau_0\sim\mu_0,\,\tau_1\sim\mu_1}\big[\big(r_{\mathbf{\Theta}_1}(\tau_0) - r_{\mathbf{\Theta}_1}(\tau_1) - \big(r_{\mathbf{\Theta}_2}(\tau_0) - r_{\mathbf{\Theta}_2}(\tau_1)\big)\big)^2\big],$$

$$\widehat{D}_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1) = \frac{1}{N}\sum_{\{\tau_0^j,\tau_1^j\}\in\mathcal{D}}\big[\big(r_{\mathbf{\Theta}_1}(\tau_0^j) - r_{\mathbf{\Theta}_1}(\tau_1^j) - \big(r_{\mathbf{\Theta}_2}(\tau_0^j) - r_{\mathbf{\Theta}_2}(\tau_1^j)\big)\big)^2\big],$$

where $\mathcal{D}$ is a dataset satisfies $|\mathcal{D}| = N$ and all trajectory pairs $\{\tau_0^j, \tau_1^j\} \in \mathcal{D}$ are sampled from μ_0 and μ_1 respectively. Then, for any $\delta \in (0,1]$, there exists a number $N_{unif}(\mathcal{G}_r,\mu_0,\mu_1,\delta)$ such that for any $N \ge N_{unif}(\mathcal{G}_r,\mu_0,\mu_1,\delta)$, the empirical estimate $\widehat{D}_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1)$ of $D_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1)$ satisfies the following inequality with probability at least $1-\delta$ for all $r_{\mathbf{\Theta}_1}, r_{\mathbf{\Theta}_2} \in \mathcal{G}_r$: $0.9\, D_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1) \le \widehat{D}_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1) \le 1.1\, D_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1)$.

Assumption 2 indicates that the empirical estimate $\widehat{D}_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1)$ closely approximates the true value $D_{\mathbf{\Theta}_1,\mathbf{\Theta}_2}(\mu_0,\mu_1)$ with high probability. This assumption is crucial in our context because it ensures that, given a sufficiently large sample size N, the empirical data provides a reliable approximation of the expected squared differences in reward discrepancies across all pairs of reward functions in $\mathcal{G}_r$. A similar assumption is adopted by Zhan et al. (2023) and proved to be held when the reward function is constructed by a linear representation and linear local head (Zhong et al., 2024). We note that this assumption is analogous to the uniform concentration results commonly used in statistical learning, where empirical estimates converge uniformly to their expected values over a class of functions (see, e.g., Vershynin (2018); Du et al. (2020); Tripuraneni et al. (2021)). It is a mild assumption and can be satisfied for various function classes. For example, polynomial functions of bounded degrees satisfy this assumption.

4.3 Main Results

Building upon the aforementioned definitions and assumptions, we now present our main theoretical results. For ease of exposition, we denote $\mathcal{G}_{\boldsymbol{r}}(\mathcal{S}^{\text{ShareLoRA}})$ by $\mathcal{G}'_{\boldsymbol{r}}$.

First, we demonstrate that the column space of $\Delta\widehat{\mathbf{\Theta}}$, obtained via Algorithm 1, closely approximates the optimal rank-k representation of $\Delta\mathbf{\Theta}^\star$. For the low-rank matrix $\Delta\widehat{\mathbf{\Theta}}$, let its SVD be $\Delta\widehat{\mathbf{\Theta}} = \widehat{\mathbf{B}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^\top$. Consequently, the column space of $\Delta\widehat{\mathbf{\Theta}}$ is spanned by the orthonormal matrix $\widehat{\mathbf{B}}$, i.e., $\text{span}\{\Delta\widehat{\mathbf{\Theta}}\} = \text{span}\{\widehat{\mathbf{B}}\}$.

For $\Delta\mathbf{\Theta}^\star$, we define its optimal rank-k approximation as

$$\mathbf{\Theta}^\diamond = \mathop{\arg\min}_{\Delta\mathbf{\Theta}:\,\text{rank}(\Delta\mathbf{\Theta})=k} \|\Delta\mathbf{\Theta}^\star - \Delta\mathbf{\Theta}\|_F. \tag{4.1}$$

Existing results in low-rank matrix factorization (Golub & Van Loan, 2013) indicate that the solution must satisfy

$\mathbf{\Theta}^\diamond = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top$, where $\mathbf{\Lambda}_k$ is a $k \times k$ diagonal matrix containing the top-k singular values of $\Delta\mathbf{\Theta}^\star$, and $\mathbf{U}_k$ and $\mathbf{V}_k$ are the corresponding left and right singular vectors, respectively. Let $\mathbf{B}^\diamond = \mathbf{U}_k$ and $\mathbf{W}^\diamond = \mathbf{\Lambda}_k \mathbf{V}_k^\top$, which yields $\mathbf{\Theta}^\diamond = \mathbf{B}^\diamond \mathbf{W}^\diamond$. Therefore, the column space of the optimal rank-k estimation of $\Delta\mathbf{\Theta}^\star$ is given by $\mathbf{B}^\diamond$, and the corresponding LoRA module for each individual reward function can be expressed as: $\Delta\mathbf{\Theta}_i = \mathbf{B}\mathbf{W}_i^\diamond$ for all $i \in [N]$, where $\mathbf{W}^\diamond = [\mathbf{W}_1^\diamond \cdots \mathbf{W}_N^\diamond]$.

To quantify the closeness between the subspaces spanned by $\widehat{\mathbf{B}}$ and $\mathbf{B}^\diamond$, we employ the principal angle distance, as detailed in Appendix A. Utilizing this metric, we establish the following theorem.

Theorem 4.1. (Closeness between $\widehat{\mathbf{B}}$ and $\mathbf{B}^\diamond$). *Suppose Assumption 1 holds. For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that*

$$\operatorname{dist}(\widehat{\mathbf{B}}, \mathbf{B}^\diamond) \leq c_1 \sqrt{\frac{1}{NN_p\nu} \log\left(\mathcal{N}_{\mathcal{G}_\mathbf{r}'}\left(\frac{1}{NN_p}\right)\frac{1}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}},$$

where $c_1 > 0$ is a constant.

The detailed proof is deferred to Appendix C.

Remark 2. *In Theorem 4.1, we demonstrate that the principal angle distance between $\widehat{\mathbf{B}}$ and $\mathbf{B}^\diamond$ decreases as the condition number increases. This implies that when the k-th singular value approaches the maximum singular value of $\Delta\mathbf{\Theta}^\star$, which is upper bounded by a constant due to the assumption in Equation* (3.2) *that $\|\Delta\mathbf{\Theta}_i^\star\|_F$ is bounded, the principal angle distance diminishes. This suggests that greater similarity among human users contributes to a more accurate estimate $\widehat{\mathbf{B}}$.*

Furthermore, the bias term in Theorem 4.1, given by $\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}$, decreases as the condition number increases and as the sum of the tail singular values decreases. Specifically, the bias term vanishes when all tail components are zero, meaning it disappears if there exists a ground-truth low-rank representation $\mathbf{B}^\star$ such that $\Delta\mathbf{\Theta}_i^\star = \mathbf{B}^\star\mathbf{W}^\star$ for all $i \in [N]$.

In Theorem 4.1, the principal angle distance is also influenced by the bracketing number $\mathcal{N}_{\mathcal{G}_\mathbf{r}'}$. We establish an upper bound on this quantity in the following proposition:

Proposition 1. *Suppose Assumption 1 holds. Then, the bracketing number for function class $\mathcal{N}_{\mathcal{G}_\mathbf{r}'}$ satisfies*

$$\log\left(\mathcal{N}_{\mathcal{G}_\mathbf{r}'}((NN_p)^{-1})/\delta\right) \leq \mathcal{O}\big(k(d_1 + Nd_2)\log(NN_p/\delta)\big). \tag{4.2}$$

The proof is deferred to Appendix C. We observe that the reward function class $\mathcal{G}_\mathbf{r}(\mathcal{S})$, as defined in Equation (3.2), has a bracketing number satisfying

$$\log\left(\mathcal{N}_{\mathcal{G}_\mathbf{r}}((NN_p)^{-1})/\delta\right) \leq \mathcal{O}\big(Nd_1d_2\log(NN_p/\delta)\big)$$

This result indicates that the bound for $\mathcal{G}_\mathbf{r}'$ is significantly improved compared with full-parameter fine-tuning when $d_1 \gg k$.

Besides, when each LoRA module is learned individually (i.e., $\mathbf{\Theta} \in \mathcal{S}^{\text{LoRA}}$), the reward function class $\mathcal{G}_\mathbf{r}(\mathcal{S}^{\text{LoRA}})$ satisfies

$$\log\left(\mathcal{N}_{\mathcal{G}_\mathbf{r}}((NN_p)^{-1})/\delta\right) \leq \mathcal{O}\big(Nk(d_1+d_2)\log(NN_p/\delta)\big)$$

Compared to Equation (4.2), our shared-component LoRA method reduces the bracketing number by decreasing the term from Nd_1k to d_1k.

Next, we establish a bound on the gap in expected value functions between the target policy $\pi_{i,\text{tar}}$ and the estimated policy $\widehat{\pi}_i$ for each individual $i \in [N]$. In this context, $\pi_{i,\text{tar}}$ serves as a benchmark for evaluating the performance of $\widehat{\pi}_i$; for instance, it may represent the optimal policy $\pi_i^\star$ associated with the true reward function $r_i^\star$.

Theorem 4.2. (Individual Expected Value Function Gap). *Suppose Assumption 1 and Assumption 2 hold. For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the output $\widehat{\pi}_i$ for any client i satisfies*

$$J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \leq c_2 \sqrt{\left(\frac{\log\left(\mathcal{N}_{\mathcal{G}_\mathbf{r}'}(\frac{1}{NN_p})\frac{1}{\delta}\right)}{NN_p\nu} + \frac{kd_2 + \log(\frac{N}{\delta})}{N_p} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + b_i\right)}$$

where b_i is defined as $b_i := \|\Delta\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2$ and $c_2 > 0$ is a constant.

Proof Sketch. We face two core challenges in our analysis. First, the reward functions are inferred from preference data rather than observed directly, introducing estimation noise that must be carefully controlled. Second, due to the low-rank structure imposed on the LoRA modules, the globally optimal shared LoRA may not perfectly capture the ground-truth reward parameters for each local dataset. This misalignment complicates the analysis of how well a single shared solution performs across different local tasks.

To address the first challenge, we leverage the continuity to translate small deviations in preference space into bounded deviations in parameter space. For the second challenge, we develop a Lagrange remainder-based analysis that quantifies the approximation error introduced by the low-rank constraint. Although perfect recovery is not guaranteed, we show that the resulting estimation error remains bounded.

The proof consists of three major steps: (1) Upper bound the distance between the column space between $\Delta\widehat{\mathbf{\Theta}}$ and $\Delta\mathbf{\Theta}^\star$ (Theorem B.1); (2) Analyze the distance between the learned reward function from algorithm $1\widehat{r}_i$ and ground truth reward function $r_i^\star$ (Theorem B.3); (3) Showing the value function of the learn policy is close to the reference policy (Theorem B.2).

In Step 1, we utilize the existing result of MLE estimates

over the preference dataset, upper bound the distance between the estimated share component LoRA matrix with the ground truth parameter matrix, and then use the Davis-Kahan theorem to bound the corresponding distance between the column space of these two matrices.

In Step 2, for learned reward function with parameter matrix $\widehat{\mathbf{\Theta}}_i = \widehat{\mathbf{B}}\widehat{\mathbf{W}}_i$ and optimal low-rank approximated reward function parameterized by $\mathbf{\Theta}_i^\diamond = \mathbf{B}^\diamond \mathbf{W}_i^\diamond$, we decompose the distance between the two functions into two part: distance between $\widehat{\mathbf{B}}$ and $\mathbf{B}^\diamond$, which already bounded in Step 1, and the distance between $\widehat{\mathbf{W}}_i$ and $\mathbf{W}_i^\diamond$. For this distance, we carefully analyze the geometry of the reward function around the local optimal and utilize the Lagrange remainder to construct a delicate quadratic form of the gradient for $\mathbf{W}$, therefore upper bound the distance between $\widehat{\mathbf{W}}_i$ and $\mathbf{W}_i^\diamond$.

In Step 3, we use the result from Step 2 along with Assumption 2 to show that the expected Euclidean distance between $r_{\mathbf{\Theta}_1}(\tau_0) - r_{\mathbf{\Theta}_1}(\tau_1)$ and $r_{\mathbf{\Theta}_2}(\tau_0) - r_{\mathbf{\Theta}_2}(\tau_1)$ is small. Applying the pessimism mechanism from Algorithm 1, we then demonstrate that the difference between the value function of the learned policy and that of the reference policy is upper bounded by the Euclidean distance between reward functions.

A natural extension of the individual expected value function gap is the averaged bound, which provides insights into the general performance across all clients.

Corollary 4.1. (Averaged Expected Value Function Gap). *Suppose Assumption 1 and Assumption 2 hold. For any $\delta \in (0, 1]$, with probability at least $1-\delta$, the output policies $\{\widehat{\pi}_i\}_{i=1}^N$ satisfy*

$$\frac{1}{N}\sum_{i\in[N]} \left(J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)\right)$$
$$\leq c_3 \sqrt{\left(\frac{\log\left(\mathcal{N}_{\mathcal{G}_r'}(\frac{1}{NN_p})\frac{1}{\delta}\right)}{NN_p\nu} + \frac{kd_2 + \log(\frac{N}{\delta})}{N_p} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}\right)},$$

where $c_3 > 0$ is a constant.

Remark 3 (Sample Complexity)**.** *For full-parameter fine-tuning, the sample complexity required to ensure that the averaged expected value function gap is less than ϵ with probability at least $1-\delta$ is $N_p = \mathcal{O}\left(\frac{d_1 d_2}{\epsilon}\log\left(\frac{N}{\delta}\right)\right)$ (Zhu et al., 2023). In contrast, when using Algorithm 1, the sample complexity required to achieve an averaged estimated value function accuracy of $1 - \epsilon - \left(\frac{\Sigma_{\text{tail}}}{N}\right)^{1/4}$ is*

$$N_p = \mathcal{O}\left(\frac{d_1 k + N d_2 k}{N\epsilon}\log\left(\frac{N}{\delta}\right)\right).$$

Therefore, when $d_1 \gg k$, the sample complexity is significantly reduced, with the trade-off being introducing a bias term in the estimation accuracy of the value function.

Moreover, Park et al. (2024) indicate that their representation learning-based method can learn an ϵ-optimal policy with a sample complexity of

$$N_p = \mathcal{O}\left(\frac{d_1 k + N k}{N\epsilon}\log\left(\frac{N}{\delta}\right)\right).$$

Notably, in their setting, d_2 is assumed to be 1, and the ground truth reward functions are posited to share a common representation with linear heads. In contrast, our results demonstrate a similar sample complexity with an additional bias term $\left(\frac{\Sigma_{\text{tail}}}{N}\right)^{1/4}$ in the accuracy. Importantly, this bias term vanishes if a ground-truth low-rank representation $\mathbf{B}^\star$ exists such that $\mathbf{\Theta}_i^\star = \mathbf{B}^\star \mathbf{W}^\star$ for all $i \in [N]$. Hence, we can achieve similar sample complexity but for the more general reward function class and without assuming the existence of ground truth common representation.

5 Experimental Results

Models and Datasets. We implement the baseline algorithms `Share Rep`, `LoRA-local`, and `LoRA-global`, which will be introduced later, alongside our proposed algorithms on two models: GPT-J 6B (Wang & Komatsuzaki, 2021) and Llama-3 8B (Touvron et al., 2023). This setup enables a comparison with the work of Park et al. (2024). Implementation details for all algorithms are provided in Appendix D.2, and the code is publicly available*.

We empirically evaluate our algorithms on the text summarization task using the Reddit TL;DR summarization and human feedback dataset (Stiennon et al., 2020). This dataset contains a broad range of user preferences, which provides a particularly suitable setting for studying personalized feedback and allows us to validate the proposed P-ShareLoRA method for learning individualized reward functions. Following Park et al. (2024), we rank the labelers by the number of annotated comparisons in the training split and select the top five workers. To balance the dataset, we cap each worker's samples to match the worker with the fewest comparisons, resulting in 5,373 samples per worker and 26,865 training samples in total. The same process is applied to the validation set, yielding 1,238 samples per worker and 6,190 validation samples overall.

Baselines. To evaluate our approach, we introduce two naive baselines for comparison: `LoRA-Global`, in which we train one shared LoRA module across all users; and `LoRA-Local`, where for each labeler's preference dataset, we independently train a separate LoRA module, allowing each user's model to fully adapt to their specific preferences without leveraging shared information across users.

To practically solve Equation (3.5), we propose three alternative algorithms to obtain personalized LoRA mod-

*https://github.com/DonghaoLee/Shared-LoRA-Reward

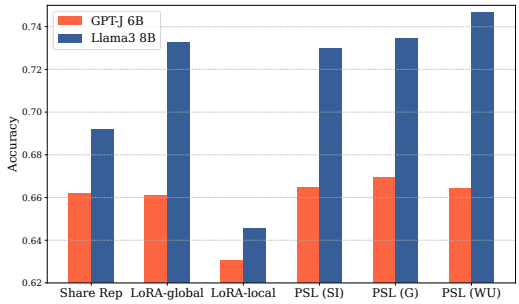

Figure 1: Prediction Accuracy of Different Algorithms.

ules with shared components: `P-ShareLoRA(SI)`, `P-ShareLoRA(G)` and `P-ShareLoRA(WU)`.

The first algorithm, `P-ShareLoRA(SI)`, where SI denotes Standard Initialization, initializes the shared B matrix to zero for all users, while each personalized matrix A_i is initialized with samples from a normal distribution. Both the shared B and the personalized A_i matrices are updated through optimizing the objective function outlined in Equation (3.5) using the adamW (Loshchilov, 2017) optimizer.

The second algorithm, `P-ShareLoRA(G)`, where G denotes Global, initializes the model by pre-training the LoRA module on the entire user dataset, using the configuration from `LoRA-Global`. Training then proceeds in the same manner as in `P-ShareLoRA(SI)`. The third algorithm, `P-ShareLoRA(WU)`, where WU denotes warm-up, employs a few preliminary warm-up steps using a global adaptation module (similar to `P-ShareLoRA(G)`) before proceeding with user-specific training. Following this phase, training continues as in `P-ShareLoRA(SI)`. Detailed pseudocode and parameter settings for each of these algorithms are provided in Appendix D.1.

We additionally include the shared representation method by Park et al. (2024) as another baseline, abbreviated as "Share Rep" in Figure 1. In this algorithm, the first 70% of the reward model's layers are frozen as the shared representation, while the remaining 30% are treated as personalized heads.

Results. For each method, we train it for a total of 3 epochs. Specifically, the global pretraining phase in `P-ShareLoRA(G)` is set to two epochs, while in `P-ShareLoRA(WU)` it is set to 0.3 epochs. Following these warm-up phases, we train `P-ShareLoRA(G)` and `P-ShareLoRA(WU)` for one and 2.7 epochs, respectively, ensuring that the total number of training steps remains uniform across all algorithms.

In Figure 1, we present the results of reward model fine-tuning using different algorithms. The reported accuracy represents the average accuracy across the test datasets of the five labelers when preferences are estimated using each algorithm. The abbreviation PSL represents `P-ShareLoRA`.

We observe that for both GPT-J 6B and Llama-3 8B models, our proposed algorithms `P-ShareLoRA(G)` and `P-ShareLoRA(WU)` demonstrate performance improvements over other baseline algorithms. Specifically, `P-ShareLoRA(G)` achieves the most significant enhancement on GPT-J 6B, while `P-ShareLoRA(WU)` performs best on Llama-3 8B. These empirical results validate the effectiveness of our method, which leverages the shared components of LoRA modules to adapt personalized reward functions. Additional experimental results are presented in Appendix D.3.

6 Conclusion

In this work, we introduced a novel algorithm that integrates LoRA into the personalized RLHF framework to effectively align LLMs with diverse user preferences. By applying LoRA to an aggregated parameter matrix, our method captures individual user preferences while leveraging shared structures, thereby improving the sample complexity and enjoying the computational efficiency of LoRA. Theoretical analysis demonstrates that `P-ShareLoRA` results in a low-rank approximation for the ground truth aggregated parameter matrix and achieves near-optimal policy performance, with performance discrepancies controlled by the diversity of user preferences. Empirical evaluations on the Reddit TL;DR dataset exhibit performance improvements compared to baseline algorithms.

Acknowledgement

The work of R. Liu, D. Li and J. Yang was supported in part by the U.S. National Science Foundation under the grants ECCS-2133170 and ECCS-2318759. The work of P. Wang and C. Shen was supported in part by the U.S. National Science Foundation under the grants CNS-2002902, ECCS-2029978, ECCS-2143559, ECCS-2033671, CPS-2313110, and ECCS-2332060.

References

Abramson, J., Ahuja, A., Carnevale, F., Georgiev, P., Goldin, A., Hung, A., Landon, J., Lhotka, J., Lillicrap, T., Muldal, A., et al. Improving multimodal interactive agents with reinforcement learning from human feedback. *arXiv preprint arXiv:2211.11602*, 2022.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with

reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., and Wang, M. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Chen, H., Yuan, K., Huang, Y., Guo, L., Wang, Y., and Chen, J. Feedback is all you need: from chatgpt to autonomous driving. *Science China Information Sciences*, 66(6):1–3, 2023.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *stat*, 1050:17, 2023.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. JHU press, 2013.

Guo, P., Zeng, S., Wang, Y., Fan, H., Wang, F., and Qu, L. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*, 2024.

Hayou, S., Ghosh, N., and Yu, B. Lora+: Efficient low rank adaptation of large models. *arXiv preprint arXiv:2402.12354*, 2024.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, C., Liu, Q., Lin, B. Y., Pang, T., Du, C., and Lin, M. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

Hwang, M., Weihs, L., Park, C., Lee, K., Kembhavi, A., and Ehsani, K. Promptable behaviors: Personalizing multi-objective rewards from human preferences. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674, 2013.

Kopiczko, D. J., Blankevoort, T., and Asano, Y. M. Vera: Vector-based random matrix adaptation. *arXiv preprint arXiv:2310.11454*, 2023.

Kuo, K., Raje, A., Rajesh, K., and Smith, V. Federated lora with sparse communication. *arXiv preprint arXiv:2406.05233*, 2024.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024.

Li, Z., Yang, Z., and Wang, M. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022.

Liu, R., Shen, C., and Yang, J. Federated representation learning in the under-parameterized regime. *arXiv preprint arXiv:2406.04596*, 2024a.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. Dora: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024b.

Loshchilov, I. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luo, T., Lei, J., Lei, F., Liu, W., He, S., Zhao, J., and Liu, K. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Pacchiano, A., Saha, A., and Lee, J. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Park, C., Liu, M., Kong, D., Zhang, K., and Ozdaglar, A. E. Rlhf from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.

Poddar, S., Wan, Y., Ivison, H., Gupta, A., and Jaques, N. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Ramesh, S. S., Hu, Y., Chaimalas, I., Mehta, V., Sessa, P. G., Bou Ammar, H., and Bogunovic, I. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

Santacroce, M., Lu, Y., Yu, H., Li, Y., and Shen, Y. Efficient rlhf: Reducing the memory usage of ppo. *arXiv preprint arXiv:2309.00754*, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shen, Y., Xu, Z., Wang, Q., Cheng, Y., Yin, W., and Huang, L. Multimodal instruction tuning with conditional mixture of lora. *arXiv preprint arXiv:2402.15896*, 2024.

Sidahmed, H., Phatale, S., Hutcheson, A., Lin, Z., Chen, Z., Yu, Z., Jin, J., Komarytsia, R., Ahlheim, C., Zhu, Y., et al. Perl: Parameter efficient reinforcement learning from human feedback. *arXiv preprint arXiv:2403.10704*, 2024.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Sun, S., Gupta, D., and Iyyer, M. Exploring the impact of low-rank adaptation on the performance, efficiency, and regularization of rlhf. *arXiv preprint arXiv:2309.09055*, 2023.

Sun, Y., Li, Z., Li, Y., and Ding, B. Improving lora in privacy-preserving federated learning. *arXiv preprint arXiv:2403.12313*, 2024.

Tang, A., Shen, L., Luo, Y., Zhan, Y., Hu, H., Du, B., Chen, Y., and Tao, D. Parameter efficient multi-task model fusion with partial linearization. *arXiv preprint arXiv:2310.04742*, 2023.

Thumm, J., Agia, C., Pavone, M., and Althoff, M. Text2interaction: Establishing safe and preferable human-robot interaction. *arXiv preprint arXiv:2408.06105*, 2024.

Tian, C., Shi, Z., Guo, Z., Li, L., and Xu, C. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv preprint arXiv:2404.19245*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tripuraneni, N., Jin, C., and Jordan, M. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., and Ghodsi, A. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558*, 2022.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, B. and Komatsuzaki, A. Gpt-j-6b: A 6 billion parameter autoregressive language model. `https://github.com/kingoflolz/mesh-transformer-jax`, 2021.

Wang, Y., Liu, Q., and Jin, C. Is rlhf more difficult than standard rl? a theoretical perspective. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024a.

Wang, Z., Shen, Z., He, Y., Sun, G., Wang, H., Lyu, L., and Li, A. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *arXiv preprint arXiv:2409.05976*, 2024b.

Wu, J., Huang, Z., Hu, Z., and Lv, C. Toward human-in-the-loop ai: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering*, 21:75–91, 2023.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Xu, Y., Wang, R., Yang, L., Singh, A., and Dubrawski, A. Preference-based reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems*, 33:18784–18794, 2020.

Ye, C., Xiong, W., Zhang, Y., Jiang, N., and Zhang, T. Online iterative reinforcement learning from human feedback with general preference model. *arXiv preprint arXiv:2402.07314*, 2024.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline preference-based reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2023.

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

Zhao, S., Dang, J., and Grover, A. Group preference optimization: Few-shot alignment of large language models. *arXiv preprint arXiv:2310.11523*, 2023.

Zhong, H., Deng, Z., Su, W. J., Wu, Z. S., and Zhang, L. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

Zhu, B., Jordan, M., and Jiao, J. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

Zhu, J., Greenewald, K., Nadjahi, K., Borde, H. S. d. O., Gabrielsson, R. B., Choshen, L., Ghassemi, M., Yurochkin, M., and Solomon, J. Asymmetry in low-rank adapters of foundation models. *arXiv preprint arXiv:2402.16842*, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Deferred Definitions and Preliminary Lemmas

In our proof, we assume that all reward models are initialized from the same initial parameter matrix, i.e., $\mathbf{\Theta}_i^0 = \mathbf{\Theta}^{\text{init}}$ for any $i \in [N]$. We note that our results can be straightforwardly generalized to the case with heterogeneous initialization. Additionally, we use $\mathbf{X}^{(N)}$ to represent the column-wise replication of matrix $\mathbf{X}$ N times, i.e., $\mathbf{X}^{(N)} = [\mathbf{X}, \ldots, \mathbf{X}]$.

A.1 Deferred Definitions

Also, we introduce the following deferred definitions:

Definition A.1 (Principal Angle Distance (Jain et al., 2013)). *Given $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{d\cdot k}$ with orthonormal columns, the principal angle distance between their column spaces is defined as*

$$\text{dist}(\mathbf{B}_1, \mathbf{B}_2) = \frac{1}{\sqrt{2}} \|\mathbf{B}_1\mathbf{B}_1^\top - \mathbf{B}_2\mathbf{B}_2^\top\|_F = \|\mathbf{B}_1^\top \bar{\mathbf{B}}_2\|_F,$$

where $\bar{\mathbf{B}}_2$ is an orthonormal basis for the orthogonal complement of $\text{span}(\mathbf{B}_2)$, i.e., $\text{span}(\bar{\mathbf{B}}_2) = \text{span}(\mathbf{B}_2)^\perp$.

The principal angle distance is a standard metric for measuring the distance between subspaces (Jain et al., 2013; Collins et al., 2021).

Definition A.2 (Bracketing Number for Single Reward (Zhan et al., 2023)). *Consider the class $\mathcal{G}_r$ of functions mapping pairs of trajectories $(\tau_0, \tau_1) \in \mathcal{T} \cdot \mathcal{T}$ to preference probability vector. Specifically, each function $r \in \mathcal{G}_r$ maps (τ_0, τ_1) to $P_r(\cdot \mid \tau_0, \tau_1) \in \mathbb{R}^2$. An ϵ-bracket for $\mathcal{G}_r$ is a pair of functions (g_1, g_2) mapping $\mathcal{T}\cdot\mathcal{T}$ to $\mathbb{R}^2$ such that for all $(\tau_0, \tau_1) \in \mathcal{T}\cdot\mathcal{T}$: (1). $g_1(\tau_0, \tau_1) \le g_2(\tau_0, \tau_1)$; (2). $\|g_1(\tau_0, \tau_1) - g_2(\tau_0, \tau_1)\|_1 \le \epsilon$ The ϵ-bracketing number of $\mathcal{G}_r$, denoted by $\mathcal{N}_{\mathcal{G}_r}(\epsilon)$, is the minimal number of ϵ-brackets required to cover $\mathcal{G}_r$ in the following sense: for any function $r \in \mathcal{G}_{\boldsymbol{r}}$, there exists an ϵ-bracket $(g_{b,1}, g_{b,2})$ such that for all $(\tau_0, \tau_1) \in \mathcal{T} \cdot \mathcal{T}$,*

$$g_{b,1}(\tau_0, \tau_1) \le P_r(\cdot \mid \tau_0, \tau_1) \le g_{b,2}(\tau_0, \tau_1).$$

Definition A.3 (Concentrability Coefficient (Zhan et al., 2023)). *Given a reward vector class $\mathcal{G}_{\boldsymbol{r}}$, a human user i, a target policy π_{tar} (which could potentially be the optimal policy $\pi_i^\star$ corresponding to the true reward $r_i^\star$), and a reference policy μ_{ref}, the* concentrability coefficient *is defined as:*

$$C_{\boldsymbol{r}}\left(\mathcal{G}_{\boldsymbol{r}}, \pi_{\text{tar}}, \mu_{\text{ref}}, i\right) := \max\left\{0, \sup_{\boldsymbol{r}\in\mathcal{G}_{\boldsymbol{r}}} \frac{\mathbb{E}_{\tau_0\sim\pi_{\text{tar}},\, \tau_1\sim\mu_{\text{ref}}}\left[r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i(\tau_0) + r_i(\tau_1)\right]}{\sqrt{\mathbb{E}_{\tau_0,\,\tau_1\sim\mu_{\text{ref}}}\left[\left(r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i(\tau_0) + r_i(\tau_1)\right)^2\right]}}\right\}.$$

A.2 Preliminary Lemmas

Before presenting the proof, we introduce a few important lemmas.

Lemma 1 ((Zhan et al. (2023), Lemma 1, reward vector version)). *For any $\delta \in (0, 1]$, if $\boldsymbol{r} \in \mathcal{G}_{\boldsymbol{r}}$, with dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$, $\tau_{i,0}^{(j)} \sim \mu_0$, $\tau_{i,1}^{(j)} \sim \mu_1$, and $o_i^{(j)} \sim P_{r_i^\star}(\cdot|\tau_0^{(j)}, \tau_1^{(j)})$, there exist $C_1 > 0$ such that*

$$\sum_{i\in[N]}\sum_{j\in[N_p]} \log\left(\frac{P_{r_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})}{P_{r_i^\star}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})}\right) \le C_1 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)$$

holds.

Lemma 2 ((Liu et al. (2022), Proposition 14, scalar version)). *For any $\delta \in (0, 1]$, with probability at least $1-\delta$, if $r \in \mathcal{G}_r'$, with dataset $\widehat{\mathcal{D}} = \{(o^{(j)}, \tau_0^{(j)}, \tau_1^{(j)})_{j\in[M]}\}$ where $\tau_0^{(j)} \sim \mu_0$, $\tau_1^{(j)} \sim \mu_1$, and $o^{(j)} \sim P_{r^\star}(\cdot|\tau_0^{(j)}, \tau_1^{(j)})$,*

$$\begin{aligned}
&\mathbb{E}_{\mu_0,\mu_1}\left[\|P_r(\cdot \mid \tau_0^{(j)}, \tau_1^{(j)}) - P_{r^\star}(\cdot \mid \tau_0^{(j)}, \tau_1^{(j)})\|_1^2\right] \\
&\le \frac{C_2}{M}\left(\sum_{j\in[M]} \log\left(\frac{P_{r^\star}(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)})}{P_r(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)})}\right) + \log(\mathcal{N}_{\mathcal{G}_r'}(1/M)/\delta)\right)
\end{aligned}$$

holds where $C_2 > 0$ is a constant.

Lemma 3 ((Liu et al. (2022), Proposition 14, vector version)). *For any $\delta \in (0,1]$, with probability at least $1-\delta$, if $\boldsymbol{r} \in \mathcal{G}'_{\boldsymbol{r}}$, with dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$, $\tau_{i,0}^{(j)} \sim \mu_0$, $\tau_{i,1}^{(j)} \sim \mu_1$, and $o_i^{(j)} \sim P_{r_i^\star}(\cdot|\tau_0^{(j)}, \tau_1^{(j)})$,*

$$\frac{1}{N}\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\|P_{r_i}(\cdot \mid \tau_0^{(j)}, \tau_1^{(j)}) - P_{r_i^\star}(\cdot \mid \tau_0^{(j)}, \tau_1^{(j)})\|_1^2\right]$$
$$\leq \frac{C_2}{NN_p}\left(\sum_{i\in[N]}\sum_{j\in[N_p]} \log\left(\frac{P_{r_i^\star}(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)})}{P_{r_i}(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)})}\right) + \log(\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))/\delta)\right)$$

holds where $C_2 > 0$ is a constant.

Lemma 4. *For any use $i \in [N]$, we have the following inequality holds:*

$$\frac{1}{N}\sum_{i\in[N]} |\log\Phi(r_i^\star(\tau_0) - r_i^\star(\tau_1)) - \log\Phi(r_i^\diamond(\tau_0) - r_i^\diamond(\tau_1))| \leq 2LL_1\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Proof. From the L-Lipschitz continuity of the function $\log\Phi(x)$, for any trajectories τ_0 and τ_1, we have

$$|\log\Phi(r_i^\star(\tau_0) - r_i^\star(\tau_1)) - \log\Phi(r_i^\diamond(\tau_0) - r_i^\diamond(\tau_1))| \leq L\,|r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i^\diamond(\tau_0) + r_i^\diamond(\tau_1)|\,.$$

Recalling that $r_i^\star(\tau) = r(\tau;\boldsymbol{\Theta}_i^\star)$ and $r_i^\diamond(\tau) = r(\tau;\boldsymbol{\Theta}_i^\diamond)$, from the L_1-Lipschitz continuity of the function $r(\tau;\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$, we have

$$|r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i^\diamond(\tau_0) + r_i^\diamond(\tau_1)| \leq 2L'\|\boldsymbol{\Theta}_i^\star - \boldsymbol{\Theta}_i^\diamond\|_F.$$

Therefore,

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in[N]} |r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i^\diamond(\tau_0) + r_i^\diamond(\tau_1)| \\
&\leq \sqrt{\frac{1}{N}\sum_{i\in[N]} \left(r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i^\diamond(\tau_0) + r_i^\diamond(\tau_1)\right)^2} \\
&\leq 2L_1\sqrt{\frac{1}{N}\|\boldsymbol{\Theta}^\star - \boldsymbol{\Theta}^\diamond\|_F^2}.
\end{aligned}$$

Note that $\boldsymbol{\Theta}^\diamond$ can be derived from the truncated singular value decomposition (SVD) of $\boldsymbol{\Theta}^\star$, retaining only the top k singular values and their corresponding singular vectors (Golub & Van Loan, 2013; Liu et al., 2024a). Consequently, we have

$$\frac{1}{N}\sum_{i\in[N]} |r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i^\diamond(\tau_0) + r_i^\diamond(\tau_1)| \leq 2L_1\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Therefore, we finally have

$$\frac{1}{N}\sum_{i\in[N]} |\log\Phi(r_i^\star(\tau_0) - r_i^\star(\tau_1)) - \log\Phi(r_i^\diamond(\tau_0) - r_i^\diamond(\tau_1))| \leq 2LL_1\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

□

Lemma 5. *For local reward models parameterized by $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_N$, suppose there exists a constant $\delta > 0$ such that*

$$\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\left\|P_{\boldsymbol{\Theta}_i}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) - P_{\boldsymbol{\Theta}_i^\star}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)\right\|^2\right] \leq \delta,$$

then, for some constant $C > 0$, it holds that $\operatorname{dist}(\mathbf{B}, \mathbf{B}^\diamond) \leq \frac{\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^\star\|_F^2}{(\delta')^2} \leq C\frac{\delta}{N\nu}$ where $\nu = \sigma_k\left(\frac{(\boldsymbol{\Theta}^\star)^T\boldsymbol{\Theta}^\star}{N}\right)$.

Proof. From the Mean Value Theorem, there exists a constant $C > 0$ such that

$$\|\mathbf{\Theta} - \mathbf{\Theta}^\star\|_F^2 \leq C \sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\mathbf{\Theta}_i}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) - P_{\mathbf{\Theta}_i^\star}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)\|^2\right] \leq C\delta.$$

Define $\delta' := \min_{1\leq i\leq k, k+1\leq j\leq \min\{d_1, Nd_2\}} |\sigma_i(\mathbf{\Theta}^\star) - \sigma_j(\mathbf{\Theta})|$. Then, we observe that

$$\delta' = \min_{1\leq i\leq k, k+1\leq j\leq \min\{d_1, Nd_2\}} |\sigma_i(\mathbf{\Theta}^\star) - \sigma_j(\mathbf{\Theta})| = \sigma_k(\mathbf{\Theta}^\star).$$

Next, by applying the Davis-Kahan Theorem, we obtain

$$\text{dist}^2(\mathbf{B}, \mathbf{B}^\diamond) \leq \frac{\|\mathbf{\Theta} - \mathbf{\Theta}^\star\|_F^2}{(\delta')^2} \leq C\frac{\delta}{\sigma_k^2(\mathbf{\Theta}^\star)} = C\frac{\delta}{N\nu}.$$

This is the desired result. □

B Training from Scratch

In this section, we present our theoretical analysis, focusing on the case where the initialization is set to zero, i.e., the initial parameter matrix is $\mathbf{\Theta}^0 = \mathbf{0}^{d_1\cdot d_2}$. Consequently, $\Delta\mathbf{\Theta}^\star = \mathbf{\Theta}^\star$.

With zero initialization, we define the class of reward functions $\mathcal{G}'_{\mathbf{r}}$, in which a low-rank adaptation matrix with shared representations is learned as the parameter matrix for each individual reward function. Specifically, $\mathcal{G}'_{\mathbf{r}}$ is defined as:

$$\mathcal{G}'_{\boldsymbol{r}} = \left\{(r_{\mathbf{\Theta}_i}(\cdot))_{i\in[N]} \;\middle|\; \mathbf{\Theta} \in \mathbb{R}^{d_1\cdot Nd_2},\ \text{rank}(\mathbf{\Theta}) = k,\ \|\mathbf{\Theta}_i\|_F \leq B,\ \forall i \in [N]\right\},$$

where $\mathbf{\Theta}$ denotes the aggregated parameter matrix across all individuals, subject to a rank constraint k. Additionally, each individual parameter matrix $\mathbf{\Theta}_i$ satisfies the Frobenius norm constraint $\|\mathbf{\Theta}_i\|_F \leq B$. We state our theoretical results under zero initialization as follows:

Theorem B.1. (Closeness between $\widehat{\mathbf{B}}$ and $\mathbf{B}^\diamond$). *For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that*

$$\text{dist}(\widehat{\mathbf{B}}, \mathbf{B}^\diamond) \leq c_1\sqrt{\frac{1}{NN_p\nu}\log(\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))/\delta) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}}.$$

where $c_1 > 0$ is a constant, ν is the condition number as defined earlier, Σ_{tail} represents the aggregate tail singular values and $\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(\cdot)$ denotes the beacketing number of the function class $\mathcal{G}'_{\boldsymbol{r}}$.

Theorem B.2. (Individual Expected Value Function Gap). *For any user $i \in [N]$ and any $\delta \in (0, 1]$, set ζ in Equation* (3.6) *as*

$$\zeta = \sqrt{\frac{c_3}{2L_1^2}\left(\min_{i\in[N]}\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{\log\left(\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))/\delta\right)}{NN_p\nu} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}, \tag{B.1}$$

where $c_3 > 0$ is a constant. Then, with probability at least $1 - \delta$, the output policy $\widehat{\pi}_i$ for client i satisfies

$$\begin{aligned} &J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \\ &\leq \sqrt{c_3\left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{\log\left(\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))/\delta\right)}{NN_p\nu} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}. \end{aligned}$$

Corollary B.1. (Averaged Expected Value Function Gap). *For any $\delta \in (0, 1]$, set ζ as in Equation* (B.1)*. If $N \geq \mu^2\Sigma_{\text{tail}}$, then, with probability at least $1 - \delta$, the output policies $\{\widehat{\pi}_i\}_{i=1}^N$ satisfy the following inequality:*

$$\frac{1}{N}\sum_{i=1}^N (J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)) \leq c_4\sqrt{\frac{\log(\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))/\delta)}{NN_p}} + \sqrt{\frac{\Sigma_{\text{tail}}}{N}},$$

where $c_4 > 0$ is a constant.

B.1 Proof of Theorem B.1

Theorem B.1. (Closeness between $\widehat{\mathbf{B}}$ and $\mathbf{B}^{\diamond}$). *For any $\delta \in (0,1]$, with probability at least $1-\delta$, it holds that*

$$\operatorname{dist}(\widehat{\mathbf{B}}, \mathbf{B}^{\diamond}) \leq c_1 \sqrt{\frac{1}{NN_p\nu} \log(\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))/\delta) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}}.$$

where $c_1 > 0$ is a constant, ν is the condition number as defined earlier, Σ_{tail} represents the aggregate tail singular values and $\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(\cdot)$ denotes the beacketing number of the function class $\mathcal{G}'_{\boldsymbol{r}}$.

Proof. Consider the events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined by the satisfaction of the conditions in Lemma 1 and Lemma 3, respectively, with the confidence parameter adjusted to $\delta \leftarrow \delta/2$. This adjustment guarantees that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1-\delta$. Consequently, we conduct our analysis conditioned on the event $\mathcal{E}_1 \cap \mathcal{E}_2$.

From Lemma 4 we have

$$\sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\boldsymbol{\Theta}_i^\star}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) \leq \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\boldsymbol{\Theta}_i^\diamond}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) + cN_p\sqrt{N\,\Sigma_{\text{tail}}}.$$

Using the definition of $\widehat{\boldsymbol{\Theta}}$ gives:

$$\sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\boldsymbol{\Theta}_i^\star}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) \leq \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\widehat{\boldsymbol{\Theta}}_i}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) + cN_p\sqrt{N\,\Sigma_{\text{tail}}}.$$

Therefore, it follows that:

$$\begin{aligned}\sum_{i\in[N]}\sum_{j\in[N_p]} \log\left(\frac{P_{r_i^\star}\left(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)}\right)}{P_{r_i}\left(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)}\right)}\right) &\leq \sum_{i\in[N]}\sum_{j\in[N_p]} \log\left(\frac{P_{r_i^\diamond}\left(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)}\right)}{P_{r_i}\left(o^{(j)} \mid \tau_0^{(j)}, \tau_1^{(j)}\right)}\right) + cN_p\sqrt{N\,\Sigma_{\text{tail}}} \\ &\leq \log\left(\frac{\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}\left(\frac{1}{NN_p}\right)}{\delta}\right) + cN_p\sqrt{N\,\Sigma_{\text{tail}}}.\end{aligned}$$

By Lemma 3 we have:

$$\begin{aligned}&\frac{1}{N}\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\left\|P_{\boldsymbol{\Theta}_i}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) - P_{\boldsymbol{\Theta}_i^\star}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)\right\|_1^2\right] \\ &\leq \frac{C_2}{NN_p}\left(\sum_{i\in[N]}\sum_{j\in[N_p]} \log\left(\frac{P_{\boldsymbol{\Theta}_i^\star}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)}{P_{\boldsymbol{\Theta}_i}\left(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)}\right) + \log\left(\frac{\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}\left(\frac{1}{NN_p}\right)}{\delta}\right)\right) \\ &\leq \frac{C_2}{NN_p}\left(C_1 \log\left(\frac{\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}\left(\frac{1}{NN_p}\right)}{\delta}\right) + cN_p\sqrt{N\,\Sigma_{\text{tail}}} + \log\left(\frac{\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}\left(\frac{1}{NN_p}\right)}{\delta}\right)\right) \\ &= \frac{C_3}{NN_p}\log\left(\frac{\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}\left(\frac{1}{NN_p}\right)}{\delta}\right) + C_4\sqrt{\frac{\Sigma_{\text{tail}}}{N}},\end{aligned}$$

for any $\mathbf{r}_{\boldsymbol{\Theta}} \in \mathcal{R}(\widehat{\mathcal{D}})$, where $C_3 = C_2(C_1+1)$. By the mean value theorem, for any $\mathbf{r}_{\boldsymbol{\Theta}} \in \mathcal{R}(\widehat{\mathcal{D}})$, we obtain:

$$\begin{aligned}&\frac{1}{N}\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_{\boldsymbol{\Theta}_i}(\tau_{i,0}) - r_{\boldsymbol{\Theta}_i}(\tau_{i,1})) - (r_i^\star(\tau_{i,0}) - r_i^\star(\tau_{i,1}))\right|^2\right] \\ &\leq \frac{\kappa^2}{N}\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\left\|P_{\boldsymbol{\Theta}_i}(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}, i) - P_{\boldsymbol{\Theta}_i^\star}(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}, i)\right\|_1^2\right] \\ &\leq \frac{C_3\kappa^2}{NN_p}\log\left(\frac{\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}(1/(NN_p))}{\delta}\right) + C_4\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.\end{aligned} \tag{B.2}$$

Therefore, combining with Lemma 5 gives

$$\text{dist}^2(\mathbf{B}, \mathbf{B}^\diamond) \leq \frac{CC_3}{NN_p\nu} \log(\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))/\delta) + \frac{CC_4}{\nu}\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Also, we obtain

$$\|\mathbf{B} - \mathbf{B}^\diamond\|_F^2 \leq \text{dist}^2(\mathbf{B}, \mathbf{B}^\diamond) \leq 2\frac{CC_3}{NN_p\nu} \log(\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))/\delta) + 2\frac{CC_4}{\nu}\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

This proves the theorem. □

B.2 Proof of Theorem B.2

Before formally proving Theorem B.2, we present the following theorem as an intermediate result.

Theorem B.3. *For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that*

$$\begin{aligned}
&\frac{1}{N_p}\sum_{j\in[N_p]} \left|(r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,1}^{(j)})) - (r_{\mathbf{\Theta}_i^\star}(\tau_{i,0}^{(j)}) - r_{\mathbf{\Theta}_i^\star}(\tau_{i,1}^{(j)}))\right|^2 \\
&\leq C_8\left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right),
\end{aligned}$$

where $C_8 > 0$ is a constant.

Proof. Recall that for a function $r_{\mathbf{\Theta}}$ parameterized by the matrix $\mathbf{\Theta} \in \mathbb{R}^{d_1 \cdot Nd_2}$, we use r_θ to denote the same function parameterized by the vector θ, where $\theta = \text{vec}(\mathbf{\Theta})$. To begin, by leveraging the continuity of $r_{\mathbf{\Theta}}(\cdot)$, we can establish the following inequality:

$$\begin{aligned}
&\left|(r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,1}^{(j)})) - (r_{\mathbf{\Theta}_i^\star}(\tau_{i,0}^{(j)}) - r_{\mathbf{\Theta}_i^\star}(\tau_{i,1}^{(j)}))\right|_F^2 \\
&\leq 2\left|(r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,1}^{(j)})) - (r_{\mathbf{\Theta}_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\mathbf{\Theta}_i^\diamond}(\tau_{i,1}^{(j)}))\right|^2 + 4L\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|^2 \\
&= 2\left|(r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})) - (r_{\theta_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\theta_i^\diamond}(\tau_{i,1}^{(j)}))\right|^2 + 4L\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2.
\end{aligned} \tag{B.3}$$

Next, we focus on obtaining an upper bound for the first part of the right-hand side of the above inequality. Using the Lagrange form of the remainder in the Taylor expansion of $r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)})$, we get

$$r_{\theta_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) = \nabla_\theta r_{\bar{\theta}_i}(\tau_{i,0}^{(j)})^\top(\theta_i^\diamond - \widehat{\theta}_i).$$

Therefore, there exist $\bar{\theta}_0$ and $\bar{\theta}_1$ such that

$$(r_{\theta_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\theta_i^\diamond}(\tau_{i,1}^{(j)})) - (r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})) = \left(\nabla_\theta r_{\bar{\theta}_0}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\bar{\theta}_1}(\tau_{i,1}^{(j)})\right)^\top(\theta_i^\diamond - \widehat{\theta}_i).$$

Then, we obtain the following:

$$\begin{aligned}
&\left|(r_{\theta_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\theta_i^\diamond}(\tau_{i,1}^{(j)})) - (r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)}))\right|^2 \\
&\leq 2\left|\left(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\right)^\top(\theta_i^\diamond - \widehat{\theta}_i)\right|^2 + 4\left|\left(\nabla_\theta r_{\bar{\theta}_0}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)})\right)^\top(\theta_i^\diamond - \widehat{\theta}_i)\right|^2 \\
&\quad + 4\left|\left(\nabla_\theta r_{\bar{\theta}_1}(\tau_{i,1}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\right)^\top(\theta_i^\diamond - \widehat{\theta}_i)\right|^2 \\
&\leq 2\underbrace{\left|\left(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\right)^\top(\theta_i^\diamond - \widehat{\theta}_i)\right|^2}_{\mathcal{A}_{i,j}} + 16L_1\underbrace{\|\theta_i^\diamond - \widehat{\theta}_i\|^2}_{\mathcal{B}_i}.
\end{aligned} \tag{B.4}$$

Our remaining proof contains three major steps: (1) **Step 1:** bounding the summation of $\mathcal{A}_{i,j}$ over j; (2) **Step 2:** bounding the term $\mathcal{B}_i$; and (3) **Step 3:** combining the bounds for $\mathcal{A}_{i,j}$ and $\mathcal{B}_i$ to obtain the final result. We now proceed with the proof of the first step.

Step 1: Bounding the summation of $\mathcal{A}_{i,j}$ over j.

Regarding the term $\mathcal{A}_{i,j}$, let us denote $w_i^\diamond = \mathrm{vec}(\mathbf{W}_i^\diamond)$. Then, we have

$$\begin{aligned}
&\Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top (\theta_i^\diamond - \widehat{\theta}_i) \\
&= \Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top \big(\theta_i^\diamond + (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i^\diamond - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i^\diamond - \widehat{\theta}_i\big).
\end{aligned}$$

Utilizing the fact that $\theta_i^\diamond = (\mathbf{I}_{d_2} \otimes \mathbf{B}^\diamond) w_i^\diamond$, it follows that

$$\begin{aligned}
\mathcal{A}_{i,j} &\le 2\left|\Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top \Big((\mathbf{I}_{d_2} \otimes \mathbf{B}^\diamond) w_i^\diamond - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i^\diamond\Big)\right|^2 \\
&\quad + 2\left|\Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top \Big((\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i^\diamond - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) \widehat{w}_i\Big)\right|^2 \\
&\underset{(i)}{\le} 4L\|(\mathbf{I}_{d_2} \otimes \mathbf{B}^\diamond) - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})\|^2 \|w_i^\diamond\|^2 + 2\left|\Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})(w_i^\diamond - \widehat{w}_i)\right|^2 \\
&\underset{(ii)}{=} 4L\|\mathbf{B}^\diamond - \widehat{\mathbf{B}}\|^2 \|\mathbf{W}_i^\diamond\|_F^2 + 2\left|\Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})(w_i^\diamond - \widehat{w}_i)\right|^2,
\end{aligned}$$

where inequality (i) follows from the L-Lipschitz continuity of the function $r_\theta(\cdot)$ with respect to θ, and equality (ii) is derived from the facts that $\|(\mathbf{I}_{d_2} \otimes \mathbf{B}^\diamond) - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})\|^2 = \|\mathbf{B}^\diamond - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})\|^2$ and $\|w_i^\diamond\|^2 = \|\mathbf{W}_i^\diamond\|_F^2$. Next, we define

$$\widehat{\mathbf{\Sigma}}_i = \frac{1}{N_p} \sum_{j \in N_p} (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})^\top \Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)\Big(\nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - \nabla_\theta r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})\Big)^\top (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}).$$

Following the definition of $\widehat{\mathbf{\Sigma}}_i$, we further derive the following inequality:

$$\frac{1}{N_p} \sum_{j \in N_p} \mathcal{A}_{i,j} \le 4L\|\mathbf{B}^\diamond - \widehat{\mathbf{B}}\|^2 \|\mathbf{W}_i^\diamond\|_F^2 + 2\|w_i^\diamond - \widehat{w}_i\|_{\widehat{\mathbf{\Sigma}}_i}^2. \tag{B.5}$$

Now, we consider the following optimization problem:

$$\max_{w_i} \quad f(w_i) := \frac{1}{N_p} \sum_{j \in [N_p]} \log P_{(\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}).$$

The solution to this optimization problem is given by $\widehat{w}_i = \arg\max_w f(w_i)$. To proceed with the analysis, let us denote $x_i^{(j)} = r_{(\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i}(\tau_{i,0}^{(j)}) - r_{(\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i}(\tau_{i,1}^{(j)})$. Using this notation, the gradient of the objective function can be expressed as follows:

$$\begin{aligned}
\nabla f(w_i) = &\frac{1}{N_p} \sum_{j \in [N_p]} \left(\frac{\Phi'(x_i^{(j)})}{\Phi(x_i^{(j)})} \mathbf{1}(o_i^{(j)} = 0) - \frac{\Phi'(-x_i^{(j)})}{\Phi(-x_i^{(j)})} \mathbf{1}(o_i^{(j)} = 1)\right) \\
&\cdot (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}})^\top \big(\nabla r_{(\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i}(\tau_{i,1}^{(j)})\big),
\end{aligned}$$

and

$$\begin{aligned}\nabla^2 f(w_i) =& \frac{1}{N_p}\sum_{j\in[N_p]}\left(\frac{\Phi'(x_i^{(j)})}{\Phi(x_i^{(j)})}\mathbf{1}(o_i^{(j)}=0) - \frac{\Phi'(-x_i^{(j)})}{\Phi(-x_i^{(j)})}\mathbf{1}(o_i^{(j)}=1)\right)\\ &\cdot(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\big(\nabla^2 r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla^2 r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\big)(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})\\ &+\frac{1}{N_p}\sum_{j\in[N_p]}\left(\frac{\Phi''(x_i^{(j)})\Phi(x_i^{(j)}) - \Phi'(x_i^{(j)})^2}{\Phi(x_i^{(j)})^2}\mathbf{1}(o_i^{(j)}=0)\right.\\ &\left.+\frac{\Phi''(-x_i^{(j)})\Phi(-x_i^{(j)}) - \Phi'(-x_i^{(j)})^2}{\Phi(-x_i^{(j)})^2}\mathbf{1}(o_i^{(j)}=1)\right)\\ &\cdot(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\big(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\big)\\ &\cdot\big(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\big)^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}}).\end{aligned} \tag{B.6}$$

From the Lagrange form of the remainder in the Taylor expansion, there exist $\bar{w}_i$ such that

$$f(\widehat{w}_i) = f(w_i^\diamond) + \nabla f(w_i^\diamond)^\top(\widehat{w}_i - w_i^\diamond) + (\widehat{w}_i - w_i^\diamond)^\top\nabla^2 f(\bar{w}_i)(\widehat{w}_i - w_i^\diamond). \tag{B.7}$$

To handle the $(\widehat{w}_i - w_i^\diamond)^\top\nabla^2 f(\bar{w}_i)(\widehat{w}_i - w_i^\diamond)$ term, we define

$$\begin{aligned}\mathbf{\Sigma}_i^\diamond =& \frac{1}{N_p}\sum_{j\in N_p}(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right)\cdot\\ &\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right)^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}}),\end{aligned}$$

and let c_1 and c_1' be the maximum and minimum positive constants, respectively, such that for any i, $\|w_i\|\le B$, and any vector u, the following inequality holds:

$$\begin{aligned}c_1\, u^\top\mathbf{\Sigma}_i^\diamond u \le& \frac{1}{N_p}\sum_{j\in N_p} u^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\right)\\ &\cdot\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\right)^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})u \le c_1'\, u^\top\mathbf{\Sigma}_i^\diamond u.\end{aligned} \tag{B.8}$$

Combining this with inequality (B.6), we obtain:

$$\begin{aligned}(\widehat{w}_i - w_i^\diamond)^\top\nabla^2 f(\bar{w}_i)(\widehat{w}_i - w_i^\diamond) \le& \frac{1}{N_p}(\widehat{w}_i - w_i^\diamond)^\top\sum_{j\in[N_p]}\left(\frac{\Phi'(x_i^{(j)})}{\Phi(x_i^{(j)})}\mathbf{1}(o_i^{(j)}=0) - \frac{\Phi'(-x_i^{(j)})}{\Phi(-x_i^{(j)})}\mathbf{1}(o_i^{(j)}=1)\right)\\ &\cdot(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\big(\nabla^2 r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla^2 r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\big)(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})(\widehat{w}_i - w_i^\diamond)\\ &- c_1c_2(\widehat{w}_i - w_i^\diamond)^\top\mathbf{\Sigma}_i^\diamond(\widehat{w}_i - w_i^\diamond),\end{aligned}$$

where $c_2 = \min_x\left(\frac{\Phi'(x)^2 - \Phi''(x)\Phi(x)}{\Phi(x)^2}\right)$. Then, from the smoothness of r_θ we have:

$$\begin{aligned}&\frac{1}{N_p}\sum_{j\in[N_p]}(\widehat{w}_i - w_i^\diamond)^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\big(\nabla^2 r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,0}^{(j)}) - \nabla^2 r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i}(\tau_{i,1}^{(j)})\big)(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})(\widehat{w}_i - w_i^\diamond)\\ &\le L_2(\widehat{w}_i - w_i^\diamond)^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})(\widehat{w}_i - w_i^\diamond)\\ &= L_2\|\widehat{w}_i - w_i^\diamond\|^2\end{aligned}$$

Let $c_3 = \max_x\left(\Phi'(x)/\Phi(x)\right)$ we have

$$(\widehat{w}_i - w_i^\diamond)^\top\nabla^2 f(\bar{w}_i)(\widehat{w}_i - w_i^\diamond) \le -c_1c_2(\widehat{w}_i - w_i^\diamond)^\top\mathbf{\Sigma}_i^\diamond(\widehat{w}_i - w_i^\diamond) + c_3L_2\|\widehat{w}_i - w_i^\diamond\|^2.$$

Combining with Equation (B.7) gives

$$c_1c_2(\widehat{w}_i - w_i^\diamond)^\top\mathbf{\Sigma}_i^\diamond(\widehat{w}_i - w_i^\diamond) - c_3L_2\|\widehat{w}_i - w_i^\diamond\|^2 \le \nabla f(w_i^\diamond)^\top(w_i^\diamond - \widehat{w}_i). \tag{B.9}$$

From the smoothness of $f(w)$, we have

$$f(\widehat{w}_i) \leq f(w_i^\diamond) + \nabla f(w_i^\diamond)^\top (\widehat{w}_i - w_i^\diamond) + \frac{L_2'}{2}\|\widehat{w}_i - w_i^\diamond\|^2,$$

using the fact $f(\widehat{w}_i) \geq f(w_i^\diamond)$ we have

$$\|\widehat{w}_i - w_i^\diamond\|^2 \leq \frac{2}{L_2'}\nabla f(w_i^\diamond)^\top (w_i^\diamond - \widehat{w}_i). \tag{B.10}$$

Then, combining the above inequality with Equation (B.9), we conclude that for any $\lambda > 0$ the following inequality holds

$$\begin{aligned} c_1 c_2 \|\widehat{w}_i - w_i^\diamond\|_{\mathbf{\Sigma}^\diamond}^2 &\leq \left(1 + 2c_3\frac{L_2}{L_2'}\right)\nabla f(w_i^\diamond)^\top (w_i^\diamond - \widehat{w}_i) \\ &\leq \left(1 + 2c_3\frac{L_2}{L_2'}\right)|\nabla f(w_i^\diamond)^\top (w_i^\diamond - \widehat{w}_i)| \\ &\leq \left(1 + 2c_3\frac{L_2}{L_2'}\right)\|\nabla f(w_i^\diamond)\|_{(\mathbf{\Sigma}^\diamond + \lambda\mathbf{I})^{-1}}\|w_i^\diamond - \widehat{w}_i\|_{\mathbf{\Sigma}^\diamond + \lambda\mathbf{I}}. \end{aligned} \tag{B.11}$$

Observe that for any $\lambda > 0$, the introduced $\lambda\mathbf{I}$ term will ensure $\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I}$ is a full rank since $\mathbf{\Sigma}_i^\diamond$ is a PSD matrix. For all i, we define a random vector $V \in \mathbb{R}^{N_p}$ as follows:

$$V_{i,j} = \begin{cases} \frac{\Phi'(r_{\theta_i^\star}(\tau_{i,0}^{(j)}) - r_{\theta_i^\star}(\tau_{i,1}^{(j)}))}{\Phi(r_{\theta_i^\star}(\tau_{i,0}^{(j)}) - r_{\theta_i^\star}(\tau_{i,1}^{(j)}))} & \text{w.p.} \quad \Phi(r_{\theta_i^\star}(\tau_{i,0}^{(j)}) - r_{\theta_i^\star}(\tau_{i,1}^{(j)})) \\ -\frac{\Phi'(r_{\theta_i^\star}(\tau_{i,1}^{(j)}) - r_{\theta_i^\star}(\tau_{i,0}^{(j)}))}{\Phi(r_{\theta_i^\star}(\tau_{i,1}^{(j)}) - r_{\theta_i^\star}(\tau_{i,0}^{(j)}))} & \text{w.p.} \quad \Phi(r_{\theta_i^\star}(\tau_{i,1}^{(j)}) - r_{\theta_i^\star}(\tau_{i,0}^{(j)})) \end{cases}$$

Also, define $V_i' \in \mathbb{R}^{N_p}$ as follows:

$$V_{i,j}' = \begin{cases} \frac{\Phi'(r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)}))}{\Phi(r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)}))} & \text{w.p.} \quad \Phi(r_{\theta_i^\star}(\tau_{i,0}^{(j)}) - r_{\theta_i^\star}(\tau_{i,1}^{(j)})) \\ -\frac{\Phi'(r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)}) - r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}))}{\Phi(r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)}) - r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}))} & \text{w.p.} \quad \Phi(r_{\theta_i^\star}(\tau_{i,1}^{(j)}) - r_{\theta_i^\star}(\tau_{i,0}^{(j)})) \end{cases}$$

Therefore, $\nabla f(w_i^\diamond)$ can be rewritten as

$$\begin{aligned} \nabla f(w_i^\diamond) =& \frac{1}{N_p}\sum_{j\in[N_p]} V_{i,j}'(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{\mathbf{I}\otimes\mathbf{B}(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right) \\ =& \frac{1}{N_p}\sum_{j\in[N_p]} (V_{i,j}' - V_{i,j})(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right) \\ &+ \frac{1}{N_p}\sum_{j\in[N_p]} V_{i,j}(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right). \end{aligned}$$

Then we obtain

$$\begin{aligned} &\|\nabla f(w_i^\diamond)\|_{(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1}} \\ &\leq \left\|\frac{1}{N_p}\sum_{j\in[N_p]} (V_{i,j}' - V_{i,j})(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right)\right\|_{(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1}} \\ &\quad + \left\|\frac{1}{N_p}\sum_{j\in[N_p]} V_{i,j}(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top\left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)})\right)\right\|_{(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1}}. \end{aligned} \tag{B.12}$$

Next, we bound the first term on the right-hand side of Equation (B.12). By the Mean Value Theorem, we have $\left|\frac{\Phi'(x)}{\Phi(x)} - \frac{\Phi'(y)}{\Phi(y)}\right| \leq \xi|x-y|$, for $x, y \in [-2R_{\max}, 2R_{\max}]$. Therefore, we can write:

$$\begin{aligned}
|V'_{i,j} - V_{i,j}| &\leq \xi \left| r_{\theta_i^\star}(\tau_{i,0}^{(j)}) - r_{\theta_i^\star}(\tau_{i,1}^{(j)}) - r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) + r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)}) \right| \\
&\overset{(i)}{\leq} 2L\xi \left\| \theta_i^\star - \theta_i^\diamond + \theta_i^\diamond - (\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond \right\| \\
&\leq 2L\xi \left\| \theta_i^\star - \theta_i^\diamond \right\| + 2L\xi \left\| (\mathbf{I}_{d_2}\otimes\mathbf{B}^\diamond) - (\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}}) \right\| \cdot \|w_i^\diamond\| \\
&= 2L\xi \|\mathbf{\Theta}_i^\star - \mathbf{B}^\diamond \mathbf{W}_i^\diamond\|_F + 2L\xi \left\| \mathbf{B}^\diamond - \widehat{\mathbf{B}} \right\| \|\mathbf{W}_i^\diamond\|_F,
\end{aligned}$$

where inequality (i) follows from the L-Lipschitz continuity of $r_{\mathbf{\Theta}}(\cdot)$.

Then, we have

$$\begin{aligned}
&\left\| \frac{1}{N_p} \sum_{j\in[N_p]} (V'_{i,j} - V_{i,j}) \widehat{\mathbf{B}}^\top \left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})\mathbf{W}_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})\mathbf{W}_i^\diamond}(\tau_{i,1}^{(j)}) \right) \right\|_{(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1}} \\
&\leq 2CL\xi \|\mathbf{\Theta}_i^\star - \mathbf{B}^\diamond \mathbf{W}_i^\diamond\|_F + 2CL\xi \left\| \mathbf{B}^\diamond - \widehat{\mathbf{B}} \right\| \|\mathbf{W}_i^\diamond\|_F
\end{aligned} \tag{B.13}$$

for constant C.

Next, we bound the second term on the right-hand side of Equation (B.12). Let $V_i \in \mathbb{R}^{N_p}$ be the vector such that $[V_i]_j = V_{i,j}$ for all $j \in [N_p]$ and we define

$$\mathbf{M}_i := \frac{1}{N_p^2} \mathbf{G}_i^\top (\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1} (\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top \mathbf{G}_i$$

where

$$\mathbf{G}_i = \left[\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(1)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(1)}) \quad \cdots \quad \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(N_p)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(N_p)}) \right].$$

As shown in Zhu et al. (2023), the matrix $\mathbf{M}_i$ satisfies the following properties:

$$\mathrm{Tr}(\mathbf{M}_i) \leq \frac{d_2 k}{N_p}, \qquad \mathrm{Tr}\left(\mathbf{M}_i^2\right) \leq \frac{d_2 k}{N_p^2}, \qquad \|\mathbf{M}_i\|_F \leq \frac{1}{N_p}.$$

Furthermore, consider that the variables $V_{i,j}$ are centered sub-Gaussian random variables, as $\mathbb{E}[V_{i,j}] = 0$ and $V_{i,j}$ are bounded. Consequently, by applying Bernstein's inequality, we obtain

$$\begin{aligned}
&\left\| \frac{1}{N_p} \sum_{j\in[N_p]} V_{i,j} (\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})^\top \left(\nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,0}^{(j)}) - \nabla r_{(\mathbf{I}_{d_2}\otimes\widehat{\mathbf{B}})w_i^\diamond}(\tau_{i,1}^{(j)}) \right) \right\|_{(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1}} \\
&= \sqrt{V_i^\top \mathbf{M}_i V_i} \leq C_4 \sqrt{\frac{k d_2 + \log(N/\delta)}{N_p}},
\end{aligned} \tag{B.14}$$

with probability at least $1 - \delta/(2N)$, where $C_4 > 0$ is constant.

Subsequently, by substituting Equation (B.13) and Equation (B.14) into Equation (B.12), we obtain

$$\begin{aligned}
&\|\nabla f(\mathbf{W}_i^\diamond)\|_{(\mathbf{\Sigma}_i^\diamond + \lambda\mathbf{I})^{-1}} \\
&\leq 2CL\xi \|\mathbf{\Theta}_i^\star - \mathbf{B}^\diamond \mathbf{W}_i^\diamond\|_F + 2CL\xi \left\| \mathbf{B}^\diamond - \widehat{\mathbf{B}} \right\| \cdot \|\mathbf{W}_i^\diamond\|_F + C_4 \sqrt{\frac{k d_2 + \log(N/\delta)}{N_p}} \\
&\leq 2CL\xi \|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F + 2CL\xi B \left\| \mathbf{B}^\diamond - \widehat{\mathbf{B}} \right\| + C_4 \sqrt{\frac{k d_2 + \log(N/\delta)}{N_p}},
\end{aligned}$$

where the final inequality leverages the facts that $\mathbf{B}^\diamond \mathbf{W}_i^\diamond = \mathbf{\Theta}_i^\diamond$ and $\|\mathbf{W}_i^\diamond\|_F \leq B$.

Furthermore, utilizing Theorem B.1, we obtain

$$\|\nabla f(\mathbf{W}_i^\diamond)\|^2_{(\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I})^{-1}} \tag{B.15}$$

$$\leq C'\left(\|\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2+\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right)+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log(N/\delta)}{N_p}\right), \tag{B.16}$$

where $C'>0$ is a constant.

Notably, from Equation (B.11), by defining $c=\frac{1+2c_3}{c_1c_2}$ we have

$$\|\widehat{w}_i-w_i^\diamond\|_{\mathbf{\Sigma}^\diamond}\leq\sqrt{c^2\|\nabla f(w_i^\diamond)\|^2_{(\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I})^{-1}}+2c\lambda B\|\nabla f(w_i^\diamond)\|_{(\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I})^{-1}}}. \tag{B.17}$$

Therefore, by setting

$$\lambda=\frac{cC'}{2B}\sqrt{\|\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2+\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right)+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log(N/\delta)}{N_p}},$$

and by combining Equation (B.16) with Equation (B.19), we obtain

$$\|\widehat{w}_i-w_i^\diamond\|^2_{\mathbf{\Sigma}^\diamond} \tag{B.18}$$

$$\leq\sqrt{2}\,cC'\left(\|\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2+\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right)+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log(N/\delta)}{N_p}\right). \tag{B.19}$$

Note that by combining Equation (B.5) with Equation (B.8) we have

$$\frac{1}{N_p}\sum_{j\in N_p}\mathcal{A}_{i,j}\leq 4LB^2\left\|\mathbf{B}^\diamond-\widehat{\mathbf{B}}\right\|^2+2c_1'\left\|w_i^\diamond-\widehat{w}_i\right\|^2_{\mathbf{\Sigma}_i^\diamond}. \tag{B.20}$$

From Theorem B.1, we have

$$\|\mathbf{B}-\mathbf{B}^\diamond\|_F^2\leq 2\frac{CC_3}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right)+2\frac{CC_4}{\nu}\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}. \tag{B.21}$$

Thus, by combining Equation (B.19), Equation (B.20), and Equation (B.21), we conclude that there exists a constant $C_5>0$ such that

$$\frac{1}{N_p}\sum_{j\in N_p}\mathcal{A}_{i,j}\leq C_5\left(\|\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|^2+\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right)+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log(N/\delta)}{N_p}\right). \tag{B.22}$$

Step 2: Bounding term $\mathcal{B}_i$.

Note that from Equation (B.10), we have

$$\|\widehat{w}_i-w_i^\diamond\|^2\leq\frac{2}{L_2}\nabla f(w_i^\diamond)^\top(w_i^\diamond-\widehat{w}_i)\leq\|\nabla f(w_i^\diamond)\|_{(\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I})^{-1}}\|w_i^\diamond-\widehat{w}_i\|_{\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I}}.$$

Therefore, we obtain

$$\|\widehat{w}_i-w_i^\diamond\|^2\leq\sqrt{\lambda^2\|\nabla f(w_i^\diamond)\|^2_{(\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I})^{-1}}+2\|\nabla f(w_i^\diamond)\|_{(\mathbf{\Sigma}_i^\diamond+\lambda\mathbf{I})^{-1}}\|w_i^\diamond-\widehat{w}_i\|_{\mathbf{\Sigma}_i^\diamond}}.$$

Let

$$\lambda\leq\min\left\{1,\ \frac{cC'}{2B}\sqrt{\|\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2+\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right)+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log(N/\delta)}{N_p}}\right\},$$

using Equation (B.16) and Equation (B.17), we obtain

$$\|\widehat{w}_i - w_i^\diamond\|^2 \le C_6 \left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu} \log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p} \right), \tag{B.23}$$

for a constant $C_6 \ge 0$. Leveraging the smoothness of $r_\theta(\cdot)$ with respect to θ, we obtain the bound

$$\begin{aligned}\mathcal{B}_{i,j} &\le \left(\|\widehat{\theta}_i - (\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i^\diamond\| + \|(\mathbf{I}_{d_2} \otimes \widehat{\mathbf{B}}) w_i^\diamond - \theta_i^\diamond\| \right)^2 \\ &\le \left(\|\widehat{w}_i - w_i^\diamond\| + B\|\widehat{\mathbf{B}} - \mathbf{B}^\diamond\| \right)^2.\end{aligned}$$

Therefore, applying Theorem 4.1 and using Equation (B.23) we obtain

$$\mathcal{B}_{i,j} \le C_7 \left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu} \log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p} \right) \tag{B.24}$$

Step 3: Putting $\mathcal{A}_{i,j}$ and $\mathcal{B}_{i,j}$ together.

Combining Equation (B.4), Equation (B.22) and Equation (B.24) we have

$$\begin{aligned}&\frac{1}{N_p} \sum_{j\in[N_p]} \left| (r_{\theta_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\theta_i^\diamond}(\tau_{i,1}^{(j)})) - (r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})) \right|^2 \\ &\quad \le C_7 \left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu} \log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p} \right),\end{aligned}$$

where $C_7 > 0$ is a constant. Therefore, combining with Equation (B.3) we obtain

$$\begin{aligned}&\frac{1}{N_p} \sum_{j\in[N_p]} \left| (r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,1}^{(j)})) - (r_{\mathbf{\Theta}_i^\star}(\tau_{i,0}^{(j)}) - r_{\mathbf{\Theta}_i^\star}(\tau_{i,1}^{(j)})) \right|^2 \\ &\quad \le \frac{1}{N_p} \sum_{j\in[N_p]} \left| (r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})) - (r_{\theta_i^\diamond}(\tau_{i,0}^{(j)}) - r_{\theta_i^\diamond}(\tau_{i,1}^{(j)})) \right|^2 + 4L_1\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|^2 \\ &\quad \le C_8 \left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu} \log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p} \right),\end{aligned}$$

where $C_8 > 0$ is a constant. The proof is thus complete. □

Also, we introduce the following short lemma to upper bound the expected squared difference between the true reward differences and their estimates use Appendix B.2.

Lemma 6. *Assume Assumption 2 holds. For any $\delta \in (0, 1]$, if $N \ge N_{unif}(\mathcal{G}_r, \mu_0, \mu_1, \delta)$, with probability at least $1 - \delta$, we have*

$$\begin{aligned}&\mathop{\mathbb{E}}_{\tau_0\sim\mu_0, \tau_1\sim\mu_1} \left[\left| (r_{\theta_i^\star}(\tau_0) - r_{\theta_i^\star}(\tau_1)) - (r_{\widehat{\theta}_i}(\tau_0) - r_{\widehat{\theta}_i}(\tau_1)) \right|^2 \right] \\ &\quad \le C_9 \left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu} \log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p} \right),\end{aligned}$$

where $C_8 > 0$ is a constant.

Proof. From Assumption 2, if $N \ge N_{\text{unif}}(\mathcal{G}_r, \mu_0, \mu_1, \delta)$, with probability at least $1 - \delta$, we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\tau_0\sim\mu_0, \tau_1\sim\mu_1} \left[\left| (r_{\theta_i^\star}(\tau_0) - r_{\theta_i^\star}(\tau_1)) - (r_{\widehat{\theta}_i}(\tau_0) - r_{\widehat{\theta}_i}(\tau_1)) \right|^2 \right] \\ &\quad \le \frac{1.1}{N_p} \sum_{j\in[N_p]} \left| (r_{\theta_i^\star}(\tau_{i,0}^{(j)}) - r_{\theta_i^\star}(\tau_{i,1}^{(j)})) - (r_{\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\theta}_i}(\tau_{i,1}^{(j)})) \right|^2 \\ &\quad \le C_8 \left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu} \log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p} \right),\end{aligned}$$

where $C_8 > 0$ is a constant. This is the desired result. □

With the assistance of Lemma 6, we are now prepared to prove Theorem B.2.

Theorem B.2. (Individual Expected Value Function Gap). *For any user $i \in [N]$ and any $\delta \in (0, 1]$, set ζ in Equation* (3.6) *as*

$$\zeta = \sqrt{\frac{c_3}{2L_1^2}\left(\min_{i\in[N]} \|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{\log\left(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}'}(1/(NN_p))/\delta\right)}{NN_p\nu} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}, \tag{B.1}$$

where $c_3 > 0$ is a constant. Then, with probability at least $1 - \delta$, the output policy $\widehat{\pi}_i$ for client i satisfies

$$\begin{aligned}
&J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \\
&\leq \sqrt{c_3\left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{\log\left(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}'}(1/(NN_p))/\delta\right)}{NN_p\nu} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}.
\end{aligned}$$

Proof. To simplify notation, let $C_{\boldsymbol{r}} = C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,\text{tar}}, \mu_{i,\text{ref}}, i)$. Following the approach in Park et al. (2024), define $r_\pi^{i,\text{inf}} := \arg\min_{\boldsymbol{r}\in\mathcal{R}_i}\left(J(\pi, r_i) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i(\tau)]\right)$. By the continuity of r and the definition of $\mathcal{R}_i$, for any policy π, we have

$$|\widehat{r}_i(\tau_{i,1}) - \widehat{r}_i(\tau_{i,0})) - (r_\pi^{i,\text{inf}}(\tau_{i,1}) - r_\pi^{i,\text{inf}}(\tau_{i,0})| \leq L_1\zeta.$$

Thus, it follows that

$$\begin{aligned}
&J(\pi_{i,\text{tar}}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \\
&= (J(\pi_{i,\text{tar}}; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i^\star(\tau)]) - (J(\widehat{\pi}_i; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i^\star(\tau)]) \\
&\leq (J(\pi_{i,\text{tar}}; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i^\star(\tau)]) - (J(\pi_{i,\text{tar}}; r_{\pi_{i,\text{tar}}}^{i,\text{inf}}) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_{\pi_{i,\text{tar}}}^{i,\text{inf}}(\tau)]) \\
&\quad + (J(\widehat{\pi}_i; r_{\widehat{\pi}_i}^{i,\text{inf}}) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}(r_{\widehat{\pi}_i}^{i,\text{inf}}(\tau))) - (J(\widehat{\pi}_i; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i^\star(\tau)]) \\
&\leq (J(\pi_{i,\text{tar}}; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_i^\star(\tau)]) - (J(\pi_{i,\text{tar}}; r_{\pi_{i,\text{tar}}}^{i,\text{inf}}) - \mathbb{E}_{\tau\sim\mu_{i,\text{ref}}}[r_{\pi_{i,\text{tar}}}^{i,\text{inf}}(\tau)]) \\
&\leq \mathbb{E}_{\tau_{i,0}\sim\pi_{i,\text{tar}},\tau_{i,1}\sim\mu_{i,\text{ref}}}[(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (r_{\pi_{i,\text{tar}}}^{i,\text{inf}}(\tau_{i,1}) - r_{\pi_{i,\text{tar}}}^{i,\text{inf}}(\tau_{i,0}))] + L_1\zeta \\
&\leq C_{\boldsymbol{r}}\sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (\widehat{r}_i(\tau_{i,1}) - \widehat{r}_i(\tau_{i,0}))\right|^2\right]} + L_1\zeta \\
&\leq \sqrt{CC_{\boldsymbol{r}}^2\left(\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}
\end{aligned}$$

where $C > 0$ is a constant. The proof is thus complete. □

B.3 Proof of Corollary B.1

Corollary B.1. (Averaged Expected Value Function Gap). *For any $\delta \in (0, 1]$, set ζ as in Equation* (B.1)*. If $N \geq \mu^2\Sigma_{\text{tail}}$, then, with probability at least $1 - \delta$, the output policies $\{\widehat{\pi}_i\}_{i=1}^N$ satisfy the following inequality:*

$$\frac{1}{N}\sum_{i=1}^N (J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)) \leq c_4\sqrt{\frac{\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}'}(1/(NN_p))/\delta)}{NN_p}} + \sqrt{\frac{\Sigma_{\text{tail}}}{N}},$$

where $c_4 > 0$ is a constant.

Proof. From Theorem B.2, by summing over $i \in [N]$, we obtain the following inequality:

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in[N]} J(\pi_{i,\text{tar}}; r_i^\star) - J(\widehat{\pi}_i'; r_i^\star) \\
&\leq \sqrt{c_3\left(\frac{1}{N}\sum_{i\in[N]}\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}.
\end{aligned}$$

Furthermore, we can derive the following bound:

$$\frac{1}{N}\sum_{i\in[N]} J(\pi_{i,\text{tar}}; r_i^\star) - J(\widehat{\pi}_i'; r_i^\star)$$
$$\leq \sqrt{c_4\left(\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2+\log(N/\delta)}{N_p}\right)}.$$

Here, the inequality holds because $\frac{\Sigma_{\text{tail}}}{N} \leq \frac{1}{\mu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}$ for $N \geq \mu^2\Sigma_{\text{tail}}$. This proves the corollary. □

C Deferred Proofs in Section 4.3

First, we introduce two auxiliary lemmas.

Lemma 7. *For reward function r, suppose Assumption 1 holds. Then, we have*

$$\frac{1}{N}\sum_{i\in[N]}\left|\log\Phi(r_i^\star(\tau_0)-r_i^\star(\tau_1)) - \log\Phi(r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1))\right| \leq 2LL'\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Proof. From the L-Lipschitz continuity of the function $\log\Phi(x)$, for any trajectories τ_0 and τ_1, we have

$$\left|\log\Phi(r_i^\star(\tau_0)-r_i^\star(\tau_1)) - \log\Phi(r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)-r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1))\right|$$
$$\leq L\left|r_i^\star(\tau_0)-r_i^\star(\tau_1)-r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1)\right|.$$

From the L'-Lipschitz continuity of the function $r(\tau;\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$, we have

$$\left|r_i^\star(\tau_0)-r_i^\star(\tau_1)-r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1)\right| \leq 2L'\|\boldsymbol{\Theta}_i^\star-\boldsymbol{\Theta}^{\text{init}}-\boldsymbol{\Theta}_i^\diamond\|_F.$$

Therefore, we have

$$\frac{1}{N}\sum_{i\in[N]}\left|r_i^\star(\tau_0)-r_i^\star(\tau_1)-r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1)\right|$$
$$\leq \sqrt{\frac{1}{N}\sum_{i\in[N]}\left(r_i^\star(\tau_0)-r_i^\star(\tau_1)-r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1)\right)^2}$$
$$\leq 2L'\sqrt{\frac{1}{N}\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}^{\text{init},(N)}-\boldsymbol{\Theta}^\diamond\|_F^2}$$
$$= 2L'\sqrt{\frac{1}{N}\|\Delta\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}^\diamond\|_F^2}.$$

Note that $\boldsymbol{\Theta}^\diamond$ can be derived from the truncated SVD of $\Delta\boldsymbol{\Theta}^\star$, retaining the top k singular values (Golub & Van Loan, 2013; Liu et al., 2024a). Consequently, we have

$$\frac{1}{N}\sum_{i\in[N]}\left|r_i^\star(\tau_0)-r_i^\star(\tau_1)-r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1)\right| \leq 2L'\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Then we obtain

$$\frac{1}{N}\sum_{i\in[N]}\left|\log\Phi(r_i^\star(\tau_0)-r_i^\star(\tau_1)) - \log\Phi(r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_0)+r_{\boldsymbol{\Theta}^{\text{init}}+\boldsymbol{\Theta}_i^\diamond}(\tau_1))\right| \leq 2LL'\sqrt{\frac{\Sigma_{\text{tail}}}{N}},$$

which completes the proof. □

Lemma 8. *For local reward models parameterized by* $\{\mathbf{\Theta}_i\}_{i=1}^N$ *with* $\mathbf{\Theta}_i = \mathbf{\Theta}^{\text{init}} + \Delta\mathbf{\Theta}_i$*, if there exists a constant* $\delta > 0$ *such that*

$$\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\mathbf{\Theta}_i}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) - P_{\mathbf{\Theta}_i^\star}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)\|^2\right] \le \delta,$$

then for $\mathbf{B}$, $\mathbf{B}^\diamond \in \mathbb{R}^{d_1\cdot k}$ *with orthonormal columns satisfies* $\text{span}(\mathbf{B}) = \text{span}(\Delta\mathbf{\Theta})$ *and* $\text{span}(\mathbf{B}^\diamond) = \text{span}(\Delta\mathbf{\Theta}^\star)$*, there exists a constant* $C > 0$ *such that*

$$\text{dist}(\mathbf{B}, \mathbf{B}^\diamond) \le \frac{\|\Delta\mathbf{\Theta} - (\mathbf{\Theta}^\star - \mathbf{\Theta}^{\text{init},(N)})\|_F^2}{(\delta')^2} \le C\frac{\delta}{N\nu},$$

where $\nu = \sigma_k\left(\frac{(\Delta\mathbf{\Theta}^\star)^T\Delta\mathbf{\Theta}^\star}{N}\right)$.

Proof. From the Mean Value Theorem, there exists a constant $C > 0$ such that

$$\|\mathbf{\Theta} - \mathbf{\Theta}^\star\|_F^2 \le C\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\mathbf{\Theta}_i}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right) - P_{\mathbf{\Theta}_i^\star}\left(\cdot \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right)\|^2\right] \le C\delta.$$

Define $\delta' := \min_{1\le i\le k, k+1\le j\le \min\{d_1, Nd_2\}} |\sigma_i(\Delta\mathbf{\Theta}^\star) - \sigma_j(\Delta\mathbf{\Theta})|$. Then, we observe that

$$\delta' = \min_{1\le i\le k, k+1\le j\le \min\{d_1, Nd_2\}} |\sigma_i(\Delta\mathbf{\Theta}^\star) - \sigma_j(\Delta\mathbf{\Theta})| = \sigma_k(\Delta\mathbf{\Theta}^\star).$$

Next, by applying the Davis-Kahan Theorem, we obtain

$$\text{dist}^2(\mathbf{B}, \mathbf{B}^\diamond) \le \frac{\|\Delta\mathbf{\Theta}^\star - \Delta\mathbf{\Theta}\|_F^2}{(\delta')^2} \le C\frac{\delta}{\sigma_k^2(\Delta\mathbf{\Theta}^\star)} = C\frac{\delta}{N\nu}.$$

This proves the lemma. □

C.1 Proof of Proposition 1

In this section, we introduce the upper bound for the bracketing number of function class $\mathcal{G}_{\mathbf{r}}(\mathcal{S}^{\text{ShareLoRA}})$, which is denoted by $\mathcal{G}'_{\mathbf{r}}$.

Proposition 1. *Suppose Assumption 1 holds. Then, the bracketing number for function class* $\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}$ *satisfies*

$$\log\left(\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}((NN_p)^{-1})/\delta\right) \le \mathcal{O}\big(k(d_1 + Nd_2)\log(NN_p/\delta)\big). \tag{4.2}$$

Proof. We start from the zero initialization case, therefore $\mathcal{G}'_{\boldsymbol{r}}$ is equilvant to:

$$\mathcal{G}'_{\boldsymbol{r}} = \left\{(r_{\mathbf{\Theta}_i}(\cdot))_{i\in[N]} \;\middle|\; \mathbf{\Theta} \in \mathbb{R}^{d_1\cdot Nd_2},\ \text{rank}(\mathbf{\Theta}) = k,\ \|\mathbf{\Theta}_i\|_F \le B,\ \forall i \in [N]\right\}.$$

Similar to the proof in Zhan et al. (2023, Proposition 1), we denote by $\mathcal{F}$ the function class

$$\mathcal{F}_{\boldsymbol{r}} = \left\{(f_i(\cdot))_{i\in[N]} \;\middle|\; f_i(\tau_0,\tau_1) = P_{r_i}(o = 1 \mid \tau_0, \tau_1), (r_i(\cdot))_{i\in[N]} \in \mathcal{G}'_{\boldsymbol{r}}\right\}.$$

Let $\mathcal{I}_{\mathcal{F}}(\epsilon)$ denote the ϵ-bracket number with respect to the ℓ_∞ norm. Therefore, there exist a set $\bar{\mathcal{F}}$ satisfies $|\bar{\mathcal{F}}| = \mathcal{I}_{\mathcal{F}}(\epsilon/4N)$ such that for any $(f_i(\cdot))_{i\in[N]} \in \mathcal{G}_{\boldsymbol{r}}$, there exist $\left(\bar{f}_i(\cdot)\right)_{i\in[N]} \in \bar{\mathcal{F}}$ such that

$$\sup_{\tau_0,\tau_1} \left|f_i(\tau_0,\tau_1) - \bar{f}_i(\tau_0,\tau_1)\right| \le \frac{\epsilon}{4N}, \quad \forall i \in [N].$$

Given $(\bar{f}_i)_{i\in[N]}$, construct a bracket (g_1, g_2):

$$[g_1(o=1|\tau_0,\tau_1)]_i = \bar{f}_i - \frac{\epsilon}{4N}, \quad [g_1(o=0|\tau_0,\tau_1)]_i = 1 - \bar{f}_i - \frac{\epsilon}{4N},$$
$$[g_2(o=1|\tau_0,\tau_1)]_i = \bar{f}_i + \frac{\epsilon}{4N}, \quad [g_2(o=0|\tau_0,\tau_1)]_i = 1 - \bar{f}_i + \frac{\epsilon}{4N}.$$

Then, we observe that (g_1, g_2) satisfies $g_1(\tau_0, \tau_1) \leq g_2(\tau_0, \tau_1)$, $\|g_1(\tau_0, \tau_1) - g_2(\tau_0, \tau_1)\|_1 \leq \epsilon$ and $g_1(\tau_0, \tau_1) \leq P_{\boldsymbol{r}}(\cdot \mid \tau_0, \tau_1) \leq g_2(\tau_0, \tau_1)$. Therefore, our goal is to bound $\mathcal{I}_{\mathcal{F}}(\epsilon/4N)$. From the mean value theorem, for $a, b \in [-2R, 2R]$, there exist constant $C_R = \max_{a\in[-2R,2R]} |\Phi'(a)|$ such that

$$|\Phi(b) - \Phi(a)| \leq C_R|b - a|.$$

Denote $\boldsymbol{f} = [f_1, \cdots, f_N]^\top$, we obtain

$$\begin{aligned}
\left|\boldsymbol{f}_i(\tau_0, \tau_1) - \bar{\boldsymbol{f}}_i(\tau_0, \tau_1)\right| &\leq C_R|\boldsymbol{r}(\tau_0) - \boldsymbol{r}(\tau_1) - \boldsymbol{r}'(\tau_0) + \boldsymbol{r}'(\tau_1)| \\
&\leq 2C_R L_1 \|\mathrm{vec}(\boldsymbol{\Theta}) - \mathrm{vec}(\boldsymbol{\Theta}')\| \\
&= 2C_R L_1 \left\|\mathrm{diag}(\mathbf{B})\mathrm{vec}(\mathbf{W}) - \mathrm{diag}(\mathbf{B}')\mathrm{vec}(\mathbf{W}')\right\| \\
&\leq 2C_R L_1 \left\|\mathrm{vec}(\mathbf{W}) - \mathrm{vec}(\mathbf{W}')\right\| + 2C_R L_1 B \left\|\mathrm{diag}(\mathbf{B}) - \mathrm{diag}(\mathbf{B}')\right\| \\
&\leq \max\{1, B\} 2C_R L_1 \left\| \begin{bmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vec}(\mathbf{W}) \end{bmatrix} - \begin{bmatrix} \mathrm{vec}(\mathbf{B}') \\ \mathrm{vec}(\mathbf{W}') \end{bmatrix} \right\|. \qquad \text{(C.1)}
\end{aligned}$$

Denote $C'_R = \max\{1, B\} \cdot 2C_R L_1$. From Equation (C.1), we conclude that the $\epsilon/4N$-bracket number $\mathcal{I}_{\mathcal{F}}(\epsilon/4N)$ is bounded by the ϵ'-covering number of a $(d_1 k + N d_2 k)$-dimensional ball centered at the origin with radius B with respect to the ℓ_2 norm, where

$$\epsilon' = \frac{\epsilon}{4N \cdot \max\{1, B\} \cdot 2C_R L_1}.$$

According to Wainwright (2019), this covering number is upper bounded by $\mathcal{O}\left((d_1 k + N d_2 k) \log\left(\frac{N}{\epsilon}\right)\right)$. Therefore, for $\epsilon = 1/(NN_p)$, we conclude that the covering number $\mathcal{N}_{\mathcal{G}'_{\boldsymbol{r}}}(1/(NN_p))$ is upper bounded by $\mathcal{O}\left((d_1 k + N d_2 k) \log(NN_p)\right)$. We note that for $\mathcal{G}'_{\mathbf{r}|\boldsymbol{\Theta}^{\mathrm{init}}}$, following the same proof process, we can show that

$$\mathcal{G}'_{\mathbf{r}|\boldsymbol{\Theta}^{\mathrm{init}}} \leq \mathcal{O}\left((d_1 k + N d_2 k). \log(NN_p)\right).$$

The proof is thus complete. □

C.2 Proof of Theorem 4.1

We note that the proof of Theorem 4.1 is a natural extension of the argument used in Theorem B.1 as detailed in Appendix B.

Theorem 4.1. (Closeness between $\widehat{\mathbf{B}}$ and $\mathbf{B}^\diamond$). *Suppose Assumption 1 holds. For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, it holds that*

$$\begin{aligned}
&\mathrm{dist}(\widehat{\mathbf{B}}, \mathbf{B}^\diamond) \\
&\leq c_1 \sqrt{\frac{1}{NN_p\nu} \log\left(\mathcal{N}_{\mathcal{G}'_{\mathbf{r}}}\left(\frac{1}{NN_p}\right)\frac{1}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\mathrm{tail}}}{N}}},
\end{aligned}$$

where $c_1 > 0$ is a constant.

Proof. We define the event $\mathcal{E}_1, \mathcal{E}_2$ as satisfying Lemma 1, Lemma 3 with $\delta \leftarrow \delta/2$, respectively, so we have $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) > 1 - \delta$. We will only consider the under event $\mathcal{E}_1 \cap \mathcal{E}_2$. From Lemma 4, we have

$$\sum_{i\in[N]} \sum_{j\in[N_p]} \log P_{\boldsymbol{\Theta}_i^\star}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) \leq \sum_{i\in[N]} \sum_{j\in[N_p]} \log P_{\boldsymbol{\Theta}_i^\diamond + \boldsymbol{\Theta}^{\mathrm{init}}}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) + cN_p\sqrt{N\Sigma_{\mathrm{tail}}},$$

Then, from the definition of $\widehat{\boldsymbol{\Theta}}$, we have

$$\sum_{i\in[N]} \sum_{j\in[N_p]} \log P_{\boldsymbol{\Theta}_i^\star}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) \leq \sum_{i\in[N]} \sum_{j\in[N_p]} \log P_{\widehat{\boldsymbol{\Theta}}_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) + cN_p\sqrt{N\Sigma_{\mathrm{tail}}}.$$

Therefore, similar to the proof in Corollary B.1, we obtain that

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in[N]}\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_{\mathbf{\Theta}_i}(\tau_{i,0})-r_{\mathbf{\Theta}_i}(\tau_{i,1}))-(r_i^\star(\tau_{i,0})-r_i^\star(\tau_{i,1}))\right|^2\right]\\
&\le\frac{\kappa^2}{N}\sum_{i\in[N]}\mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\mathbf{\Theta}_i}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)},i)-P_{\mathbf{\Theta}_i^\star}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)},i)\|_1^2\right]\\
&\le\frac{C_3\kappa^2}{NN_p}\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)+C_4\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.
\end{aligned}\tag{C.2}$$

Combining this with Lemma 8, we obtain

$$\text{dist}^2(\mathbf{B},\mathbf{B}^\diamond)\le\frac{CC_3}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))}{\delta}\right)+\frac{CC_4}{\nu}\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Additionally, we have

$$\|\mathbf{B}-\mathbf{B}^\diamond\|_F^2\le\text{dist}^2(\mathbf{B},\mathbf{B}^\diamond)\le\frac{CC_3}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))}{\delta}\right)+\frac{CC_4}{\nu}\kappa^2\sqrt{\frac{\Sigma_{\text{tail}}}{N}}.$$

Hence, the proof is complete. □

C.3 Proof of Theorem 4.2

We set the tolerance level ζ in Algorithm 1 to satisfy

$$\zeta\le c_4\sqrt{\min_{i\in[N]}\|\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2+\frac{\log\left(\mathcal{N}_{\mathcal{G}_{\mathbf{r}}'}\left(\frac{1}{NN_p}\right)/\delta\right)}{NN_p\nu}+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log\left(\frac{N}{\delta}\right)}{N_p}},\tag{C.3}$$

where $c_4>0$ is a constant. In Theorem 4.2, we build upon the proof established for Theorem B.2 in Appendix B.

Theorem 4.2. (Individual Expected Value Function Gap). *Suppose Assumption 1 and Assumption 2 hold. For any $\delta\in(0,1]$, with probability at least $1-\delta$, the output $\widehat{\pi}_i$ for any client i satisfies*

$$\begin{aligned}
&J(\pi_{i,tar};r_i^\star)-J(\widehat{\pi}_i;r_i^\star)\\
&\le c_2\sqrt{\left(\frac{\log\left(\mathcal{N}_{\mathcal{G}_{\mathbf{r}}'}(\frac{1}{NN_p})\frac{1}{\delta}\right)}{NN_p\nu}+\frac{kd_2+\log(\frac{N}{\delta})}{N_p}+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+b_i\right)}
\end{aligned}$$

where b_i is defined as $b_i:=\|\Delta\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2$ and $c_2>0$ is a constant.

Proof. Recall that $\Delta\mathbf{\Theta}_i^\star=\mathbf{\Theta}_i^\star-\mathbf{\Theta}^{\text{init}}$ and $\mathbf{\Theta}^\diamond=\arg\min_{\mathbf{\Theta}:\text{rank}(\mathbf{\Theta})=k}\|\mathbf{\Theta}-\Delta\mathbf{\Theta}^\star\|$. By leveraging the continuity of $r_{\mathbf{\Theta}}(\cdot)$, we can establish the following inequality:

$$\begin{aligned}
&\left|(r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,0}^{(j)})-r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,1}^{(j)}))-(r_{\mathbf{\Theta}_i^\star}(\tau_{i,0}^{(j)})-r_{\mathbf{\Theta}_i^\star}(\tau_{i,1}^{(j)}))\right|^2\\
&\le 2\left|(r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,0}^{(j)})-r_{\widehat{\mathbf{\Theta}}_i}(\tau_{i,1}^{(j)}))-(r_{\mathbf{\Theta}^{\text{init}}+\mathbf{\Theta}_i^\diamond}(\tau_{i,0}^{(j)})-r_{\mathbf{\Theta}^{\text{init}}+\mathbf{\Theta}_i^\diamond}(\tau_{i,1}^{(j)}))\right|^2+4L\|\Delta\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|^2.
\end{aligned}$$

Therefore, similar to our proof in Appendix B.2, we will show that with probability at least $1-\delta$, we have

$$\begin{aligned}
&J(\pi_{i,\text{tar}};r_i)-J(\widehat{\pi}_i;r_i^\star)\\
&\le c_2\sqrt{\left(\|\Delta\mathbf{\Theta}_i^\star-\mathbf{\Theta}_i^\diamond\|_F^2+\frac{\log\left(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta\right)}{NN_p\nu}+\frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}+\frac{kd_2+\log(N/\delta)}{N_p}\right)}.
\end{aligned}$$

This concludes the proof. □

C.4 Proof of Corollary 4.1

We set the tolerance level ζ in Algorithm 1 as defined in Equation (C.3). Similarly, the proof of Corollary 4.1 follows that of Corollary B.1 presented in Appendix B.

Corollary 4.1. (Averaged Expected Value Function Gap). *Suppose Assumption 1 and Assumption 2 hold. For any $\delta \in (0, 1]$, with probability at least $1 - \delta$, the output policies $\{\widehat{\pi}_i\}_{i=1}^N$ satisfy*

$$\frac{1}{N}\sum_{i\in[N]}\left(J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)\right) \leq c_3\sqrt{\left(\frac{\log\left(\mathcal{N}_{\mathcal{G}_r'}(\frac{1}{NN_p})\frac{1}{\delta}\right)}{NN_p\nu} + \frac{kd_2 + \log(\frac{N}{\delta})}{N_p} + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}}\right)},$$

where $c_3 > 0$ is a constant.

Proof. From Theorem 4.2, by summing over $i \in [N]$, we obtain the following inequality:

$$\frac{1}{N}\sum_{i\in[N]} J(\pi_{i,\text{tar}}; r_i^\star) - J(\widehat{\pi}_i'; r_i^\star) \leq c_2\sqrt{\left(\frac{1}{N}\sum_{i\in[N]}\|\mathbf{\Theta}_i^\star - \mathbf{\Theta}_i^\diamond\|_F^2 + \frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}.$$

Similar to the proof for Corollary B.1, we have

$$\frac{1}{N}\sum_{i\in[N]} J(\pi_{i,\text{tar}}; r_i^\star) - J(\widehat{\pi}_i'; r_i^\star) \leq c_3\sqrt{\left(\frac{1}{NN_p\nu}\log\left(\frac{\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p))}{\delta}\right) + \frac{1}{\nu}\sqrt{\frac{\Sigma_{\text{tail}}}{N}} + \frac{kd_2 + \log(N/\delta)}{N_p}\right)}.$$

This is the desired result. □

D Experiment Details

D.1 Algorithms

In this section, we present the practical algorithms used for empirical evaluation. Algorithm 2 outlines the `P-ShareLoRA` algorithm with a warm-up phase. Notably, by setting the number of warm-up epochs $T_w = 0$, Algorithm 2 reduces to the vanilla `P-ShareLoRA` algorithm. Conversely, setting $T_w = T_{\text{Global}}$ transforms the algorithm into `global-P-ShareLoRA`. We define the per-sample function as $f(\mathbf{\Theta}; o, \tau_0, \tau_1) := \log P_{\mathbf{\Theta}}(o \mid \tau_0, \tau_1)$.

Algorithm 2 `P-ShareLoRA` for RLHF (with warm-up)

Input: Pre-trained model parameters W; Human preference dataset $\widehat{\mathcal{D}}$; Rank r; Scheduled learning rate η^t; Number of warm-up epochs T_w; Number of epochs T.
Initialize: Low-rank matrices $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times d}$, $B_i \in \mathbb{R}^{r\times d} \quad \forall i \in [N]$ (e.g., randomly or zeros).
Freeze pre-trained weights W.
Warm-up phase:
for each epoch $t = 1$ to T_w **do**
  for each $\{o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\} \in \widehat{\mathcal{D}}$ **do**
    Compute $f(W + A^t B^t; o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})$.
    Update $A^{t+1} \leftarrow A^t - \eta^t \nabla_{A^t} f$.
    Update $B^{t+1} \leftarrow B^t - \eta^t \nabla_{B^t} f$.
  end for
end for
Running `P-ShareLoRA`:
Set $A^1 \leftarrow A^{T_w}$, $B_i^1 \leftarrow B^{T_w} \quad \forall i \in [N]$.
for each epoch $t = 1$ to T **do**
  for each random sampled $\{o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\} \in \widehat{\mathcal{D}}$ **do**
    Compute $f(W + A^t B_i^t; o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})$.
    Update $A^{t+1} \leftarrow A^t - \eta^t \nabla_{A^t} f$.
    Update $B_i^{t+1} \leftarrow B_i^t - \eta^t \nabla_{B_i^t} f$.
  end for
end for
Policy optimization by `PPO-Clip` (Schulman et al., 2017):
Initialize policy parameters for each agent θ_i, $\forall i \in [N]$.
for each PPO iteration $k = 1$ to K **do**
  for each agent $i = 1$ to N **in parallel do**
    Collect a set of trajectories $\mathcal{D}_i$ by running policy $\pi_{\theta_i^t}$.
    Compute rewards $r_t^{(i)}$ and advantage estimates $\widehat{A}_t^{(i)}$ using GAE.
    Compute the PPO surrogate loss:

$$\mathcal{L}_i^{\text{CLIP}}(\theta_i) = \mathbb{E}_t \left[\min \left(\rho_t^{(i)}(\theta_i^t) \widehat{A}_t^{(i)}, \text{clip} \left(\rho_t^{(i)}(\theta_i^t), 1-\epsilon, 1+\epsilon \right) \widehat{A}_t^{(i)} \right) \right],$$

    where $\rho_t^{(i)}(\theta_i) = \dfrac{\pi_{\theta_i}(a_t^{(i)} | s_t^{(i)})}{\pi_{\theta_i^{\text{old}}}(a_t^{(i)} | s_t^{(i)})}$.
    Update $\theta_i^{t+1} \leftarrow \theta_i^t - \eta^t \nabla_\theta \mathcal{L}_i^{\text{CLIP}}$.
  end for
end for
Output: Fine-tuned model parameters for each reward model $A^T, \{B_i^T\}_{i=1}^N$; Fine-tuned model parameters for each local policy $\{\theta_i^K\}_{i=1}^N$.

We also detail the baseline algorithms `LoRA-global` and `LoRA-local` in Algorithm 3 and Algorithm 4 for comparison.

Algorithm 3 Baseline algorithm 1: `LoRA-global`

Input: Pre-trained model parameters W; Human preference dataset $\widehat{\mathcal{D}}$; Rank r; Learning rate η; Number of epochs T_{Global}.
Initialize: Low-rank matrices $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times d}$ (e.g., randomly or zeros).
Freeze pre-trained weights W.
for each epoch $t = 1$ to T_{Global} **do**
  for each random sampled $\{o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\} \in \widehat{\mathcal{D}}$ **do**
    Compute $f(W + A^t B^t; o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})$.
    Update $A^{t+1} \leftarrow A^t - \eta \nabla_{A^t} f$.
    Update $B^{t+1} \leftarrow B^t - \eta \nabla_{B^t} f$.
  end for
end for
Output: Fine-tuned model parameters for a global reward model $A^{T_{\text{Global}}}, B^{T_{\text{Global}}}$.

Algorithm 4 Baseline algorithm 2: `LoRA-local`

Input: Pre-trained model parameters W; Human preference dataset $\widehat{\mathcal{D}}$; Rank r; Learning rate η; Number of epochs T_{local}.
Initialize: Low-rank matrices $A_i \in \mathbb{R}^{d\times r}$, $B_i \in \mathbb{R}^{r\times d} \quad \forall i \in [N]$ (e.g., randomly or zeros).
Freeze pre-trained weights W.
for each agent $i = 1$ to N **in parallel do**
  for each epoch $t = 1$ to T_{local} **do**
    for each random sampled $\{o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\} \in \widehat{\mathcal{D}}_i$ **do**
      Compute $f(W + A_i^t B_i^t; o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})$.
      Update $A_i^{t+1} \leftarrow A^t - \eta \nabla_{A^t} f$.
      Update $B_i^{t+1} \leftarrow B^t - \eta \nabla_{B^t} f$.
    end for
  end for
end for
Output: Fine-tuned model parameters for each reward model $\{A_i^{T_{\text{local}}}\}_{i=1}^N, \{B_i^{T_{\text{local}}}\}_{i=1}^N$.

D.2 Implementation Details

Hyperparamters. For all experiments conducted using both Vanilla LoRA (`LoRA-global` and `LoRA-local`) and `P-ShareLoRA` based algorithms, we employed a batch size of 128. The initial learning rate was set to $5 \cdot 10^{-5}$, with a linear scheduler applied to adjust the learning rate during training. For both GPT-J 6B and Llama3 8B models, the maximum token length was set to 2048. The rank k in all LoRA modules was fixed at 32, and the scaling factor α was set to 16. To simplify training, we applied LoRA only to the Q (query) and K (key) matrices for both models.

In the case of the `P-ShareLoRA(G)`, the initialization process was critical for ensuring effective fine-tuning. Specifically, the personalized A matrices and the shared B matrix were initialized using the A and B matrices obtained after two epochs of training with the `LoRA-global` method. Following this initialization, the PLAS model was fine-tuned for an additional epoch to refine the parameters further.

To maintain a fair comparison between `P-ShareLoRA(G)` and the other training methods, we adjusted the starting learning rate for `P-ShareLoRA(G)`. Given that a learning rate scheduler was used, the initial learning rate for PLAS-FT was set to one-third of the original learning rate, specifically $1.67 \cdot 10^{-5}$. This adjustment ensures that the fine-tuning process operates under comparable training dynamics as the baseline methods.

All experiments were implemented based on TRL[†], and additional hyperparameters were kept consistent across different methods.

Computational Resources. Our experiments were conducted using two NVIDIA A100 80GB GPUs. Training

[†]https://huggingface.co/docs/trl/en/index

`P-LoRAShare(SI)` on a single GPU took around six hours, but this time could be reduced with multi-GPU training.

D.3 Additional Experiment Results

Individual Labeler Performance. In Section 5, we present the averaged preference estimation accuracy across all five labelers. In this section, we also provide the results of the separate estimation accuracy for each labeler in Figure 2. We observe that our proposed methods, `P-ShareLoRA(SI)`, `P-ShareLoRA(G)`, and `P-ShareLoRA(WU)`, consistently outperform the baseline methods `LoRA-global` and `LoRA-local` for most of the labelers. Specifically, `P-ShareLoRA(WU)` achieves the highest accuracy for most labelers, peaking at 0.7803 for Labeler 1. While `P-ShareLoRA(SI)` and `P-ShareLoRA(G)` also show significant improvements over the baseline methods for labelers 0,1 and 2.

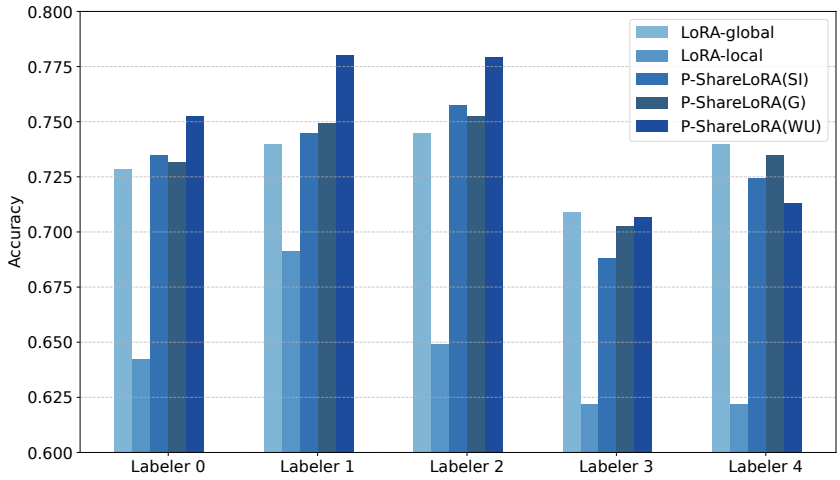

Figure 2: Accuracies of Different Methods Across Labelers (Llama3 8B)

Share Down-projection VS Share Up-projection. Previous works (Tian et al., 2024; Guo et al., 2024) have observed that the cosine similarity among down-projection matrices (A matrices) is significantly higher than that among up-projection matrices (B matrices). They interpret this as indicating that the down-projection matrices serve as a shared representation, mapping the input into a common representation space. Based on this observation, they introduce methods of sharing down-projection matrices among clients or experts. In contrast, our study finds that sharing the up-projection matrices (B matrices) yields better performance, as illustrated in Figure 3. Specifically, the approach of sharing B matrices consistently outperforms the method of sharing A matrices across all labelers and for both GPT-J 6B and Llama 3 8B models.

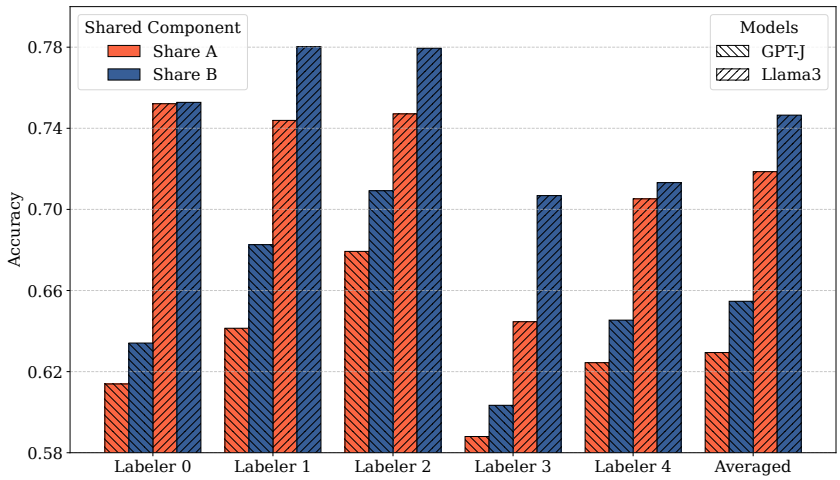

Figure 3: Compare Accuracy between Share A and Share B