Principled RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation

Chanwoo Park* Mingyang Liu Kaiqing Zhang Asuman Ozdaglar

May 2, 2024

Abstract

Reinforcement learning from human feedback (RLHF) has been an effective technique for aligning AI systems with human values, with remarkable successes in fine-tuning large-language models recently. Most existing RLHF paradigms make the underlying assumption that human preferences are relatively *homogeneous*, and can be encoded by a single reward model. In this paper, we focus on addressing the issues due to the inherent *heterogeneity* in human preferences, as well as their potential *strategic* behavior in providing feedback. Specifically, we propose two frameworks to address heterogeneous human feedback in principled ways: personalization-based one and preference-aggregation-based one. For the former, we propose two approaches based on representation learning and clustering, respectively, for learning *multiple* reward models that trades-off the bias (due to preference heterogeneity) and variance (due to the use of fewer data for learning each model by personalization). We then establish sample complexity guarantees for both approaches. For the latter, we aim to adhere to the single-model framework, as already deployed in the current RLHF paradigm, by carefully *aggregating* diverse and truthful preferences from humans. We propose two approaches based on reward and preference aggregation, respectively: the former utilizes both utilitarianism and Leximin approaches to aggregate individual reward models, with sample complexity guarantees; the latter directly aggregates the human feedback in the form of probabilistic opinions. Under the probabilistic-opinion-feedback model, we also develop an approach to handle strategic human labelers who may bias and manipulate the aggregated preferences with untruthful feedback. Based on the ideas in mechanism design, our approach ensures truthful preference reporting, with the induced aggregation rule maximizing social welfare functions.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) has been widely used in fine-tuning Large Language Models like ChatGPT and Claude 3, showing their significance and usefulness in aligning these models with human values/utilities that can be nuanced and complicated (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022). The underlying assumption of RLHF is that human utility functions are closely linked to human preferences, making it feasible to learn about human utility from preference data. Therefore, most of the RLHF use a *single* utility function (also

*C. Park, M. Liu, and A. Ozdaglar are with Massachusetts Institute of Technology, Cambridge, MA, 02139. K. Zhang is with the University of Maryland, College Park, MD, 20742. E-mails: {cpark97,liumy19, asuman}@mit.edu, kaiqing@umd.edu.

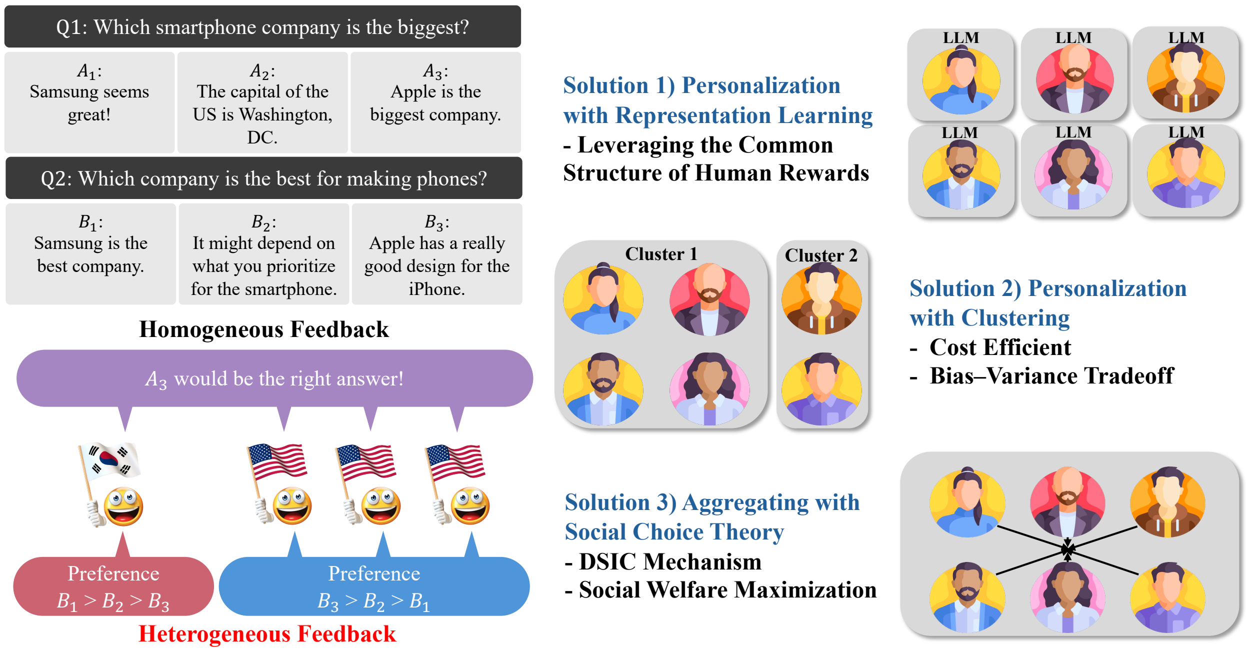

Figure 1: We demonstrate a setting where humans might have heterogeneous feedback. We provide a personalization-based framework and a human preference aggregation-based framework.

known as a reward function) to model the preferences of human labelers (Ziegler et al., 2019). Contrary to the assumption of "homogeneity" in reward valuation, humans assign "heterogeneous" reward values to the same question-and-answer pairs, especially for sensitive and open-ended questions, depending on their background. This diversity suggests that the probability of selecting a particular answer cannot be universally proportional to a presumed common reward value (Figure 1). Therefore, the assumption that having a homogeneous preference for humans, which is arguably an overly simplified model could potentially lead to reward models that prioritize the preferences of the majority group in the data while neglecting the preferences of minorities. This is concerning because it can result in biases and inequalities in the fine-tuned LLM systems, making it less fair and representative of diverse perspectives that are inherent in society.

In addressing the challenge of diverse human preferences, the notion of personalized reward models and LLMs emerges: Intuitively, one could consider creating tailored LLMs for each human user who *provides their own personalized data*. However, this method might be inadequate due to the *limited* data available for each individual, potentially leading to inaccurate reward estimations and reduced precision in LLMs due to high variance. Conversely, employing a single reward and associated language model for every human user by naively pooling all users' data often relies on an implicit assumption that humans have *homogeneous* preferences, which is barely the case in practice (Pollak and Wales, 1992; Boxall and Adamowicz, 2002).

On the other hand, even when acknowledging the heterogeneity of human preferences and aggregating them carefully to learn a *single* reward model, there is a challenge that has not received enough attention in the RLHF literature: humans are by nature rational (of certain degrees) and strategic, with their own objectives to optimize. They might manipulate their own feedback by not providing *truthful* one, to distort the aggregated results to be closer to their own preferences, and thus manipulate the output of LLMs fine-tuned over the aggregated preference data. For example, in online rating systems, users may provide extreme feedback to disproportionately influence the overall ratings toward their viewpoint.

Given the challenges above, we raise and attempt to address two questions: (1) how should we mitigate the variance of the personalized reward model and LLM model arising from the paucity of each human user's data, and (2) if we use a *single* LLM model, as already deployed in the current RLHF pipeline, what constitutes an effective method for aggregating diverse and *truthful* preferences for fine-tuning LLMs?

Contributions. We provide a systematic study to address the challenges above due to the heterogeneity of human preferences in RLHF, with provable efficiency guarantees. We develop two frameworks: one based on personalization, and one based on human preference aggregation. The personalization-based framework seeks to address the question (1) above, and includes two approaches as detailed below:

- **Representation-learning-based Personalization:** We propose a representation learning-based approach for learning the heterogeneous individual reward function (see Figure 1). By leveraging comparison data from a *diverse* set of humans, we can enhance the accuracy of the representation learning (Theorem 3.4). This improvement results in better sample complexity for estimating the personalized reward model and, consequently, more effective policy learning compared to approaches using non-diverse but heterogeneous data sets (Theorem 3.1). To improve the sample complexity, we develop a new proof technique to bound the reward estimation errors without accurate representations, leading to the use of summations of non-zero expectation random variables, which may be of independent interest (proof of Lemma 5).

- **Clustering-based Personalization:** We then consider personalization with human user clustering, with sample complexity guarantees. We introduce an approach for learning clustered reward functions, and personalizing users' reward model within each cluster by the learned model for that cluster.

The aggregation-based framework seeks to address question (2), which aggregates heterogeneous human preferences in a principled way via social choice theory, and includes two aggregation approaches as detailed below:

- **Reward Aggregation with Comparison Data:** We first estimate the parameters for each individual's reward model using the individual's preference comparison data. Then, we aggregate reward models using a family of reward aggregation rules, including those based on the utilitarianism and Leximin approaches. We then provide sample complexities of the policy induced from the single aggregated reward model.

- **Preference Aggregation with Probabilistic Opinion Data:** We then provide aggregation methods for RLHF with a new feedback form: probabilistic opinion (Dietrich and List, 2016), where human labelers provide feedback as a *probability vector* instead of binary comparison data, as in the existing RLHF pipeline. Without assuming the relationship between human reward and preference, we directly aggregate the diverse probabilistic preferences into a consensus one. Under this feedback model, we also develop a mechanism to handle *strategic* human labelers, who may benefit by reporting untruthful preferences, while biasing the aggregated preference and thus the fine-tuned LLM. Our mechanism guarantees that the labelers will report truthful preferences, while the induced aggregation rules maximize various social welfare functions.

1.1 Related Work

Reinforcement Learning from Human Feedback. Empirical evidence has demonstrated the efficacy of incorporating human preferences into reinforcement learning (RL) for enhancing robotics (Abramson et al., 2022; Hwang et al., 2023) and for refining large-scale language models (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022). These human inputs take various forms, such as rankings (Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022), demonstrations (Finn et al., 2016), and scalar ratings (Warnell et al., 2018). A few approaches have been explored empirically to personalize RLHF. For example, assigning fine-grained rewards to small text segments to enhance the training process (Wu et al., 2024), or training each human labeler's reward model with Multi-Objective Reinforcement Learning perspective (Jang et al., 2023; Hwang et al., 2023) have been proposed. Moreover, (Li et al., 2024) suggested the training of each human labeler's reward model directly using personalized feedback with human embedding obtained by the human model, and also an approach for the clustering with finding cluster embedding.

On the theory front, the studies of RLHF have received increasing research interest. The most related prior works are (Zhu et al., 2023; Zhan et al., 2023; Wang et al., 2024), where (Zhu et al., 2023) investigated the Bradley-Terry-Luce (BTL) model (Bradley and Terry, 1952) within the context of a linear reward framework; while (Zhan et al., 2023) generalized the results to encompass more general classes of reward functions. Both works concern the setting with offline preference data. (Xiong et al., 2023) provided a theoretical analysis for KL-regularized RLHF. In the online setting, (Wang et al., 2024) established a correlation between online preference learning and online RL through a *preference-to-reward* interface. Yet, to the best of our knowledge, there is no prior work that has analyzed RLHF with heterogeneous feedback with theoretical guarantees (except the recent independent works discussed in detail below).

Representation Learning. Early work of (Baxter, 2000) established a generalization bound that hinges on the concept of a task generative model within the representation learning framework. More recently, (Tripuraneni et al., 2021; Du et al., 2020) demonstrated that, in the setup with linear representations and squared loss functions, task diversity can significantly enhance the efficiency of learning representations. Moreover, (Tripuraneni et al., 2020) provided a representation learning with general representation and general loss functions. Representation learning has been extended to the reinforcement learning setting as well. For low-rank Markov Decision Processes, where both the reward function and the probability kernel are represented through the inner products of state and action representations with certain parameters, (Agarwal et al., 2020; Ren et al., 2022; Uehara et al., 2021) explored the theoretical foundations for learning these representations. Also, (Ishfaq et al., 2024; Bose et al., 2024) analyzed the sample complexity of multi-task offline RL.

Reward and Preference Aggregation. Preference aggregation is the process by which multiple humans' preference orderings of various social alternatives are combined into a single, collective preference or choice (List, 2013). Arrow's Impossibility Theorem demonstrates that no aggregation rule for preference orderings can simultaneously meet specific criteria essential for ensuring a fair and rational aggregation of each human user's preferences into a collective decision (Arrow, 1951). Therefore, people considered replacing preference orderings with assigning real numbers to social alternatives (Sen, 2018; Moulin, 2004), which is sometimes called a reward (welfare) function [♮] for each human user. (Skiadas, 2016; Moulin, 2004) provided reward (welfare) aggregation rules which satisfy several desirable properties. Furthermore, an alternative method to circumvent Arrow's impossibility theorem involved aggregating preferences via probabilistic opinion (Stone, 1961; Lehrer and Wagner, 2012). In this approach, opinions are represented as probability assignments to specific events or propositions of interest.

Comparison with Recent Works. While preparing the present work, we noticed two recent independent works that are closely related. Firstly, (Chakraborty et al., 2024) considered the aggregation of reward models with heterogeneous preference data, focusing on aligning with the Egalitarian principle in social choice theory. In contrast, we provide a framework with various aggregation rules and also prove that the aggregation rules we considered are also welfare-maximizing. More importantly, we design mechanisms for human feedback providers so that they can truthfully report their preferences even when they may be strategic. Moreover, we also develop another framework to handle heterogeneous preferences: the personalization-based one. Finally, we establish near-optimal sample complexity analyses for the frameworks we developed.

More recently, (Zhong et al., 2024) provided a theoretical analysis of reward aggregation in RLHF, focusing primarily on linear representations. Our work, in comparison, considers general representation functions and general relationships between reward function and preference. Unlike (Zhong et al., 2024), where they focused on reward aggregation, we focus on personalization for every human labeler and also employ clustering techniques for personalization. (Zhong et al., 2024) and our paper also both investigated the case that reward and preference are not related. Our paper suggested a probabilistic opinion pooling with a mechanism design to effectively elicit truthful human preferences, presuming human labelers may be strategic. In contrast, (Zhong et al., 2024) analyzed an algorithm for a von Neumann winner policy, where a von Neumann winner policy is a policy that has at least a 50% chance of being preferred compared to any other policy. Moreover, (Zhong et al., 2024) also explored the Pareto efficiency of the resulting policy.

Notation. The matrix $\mathbf{O}$ denotes an all-zero matrix, while I stands for an identity matrix, of proper dimensions. We use $A \succ \mathbf{O}$ to denote that matrix A is a positive definite matrix. The function σ represents the Sigmoid function, defined by $\sigma(x) = 1/(1+\exp(-x))$. The notation $[K]$ denotes the set $\{1, 2, \ldots, K\}$. $\Delta(\mathcal{A})$ refers to a probability vector in $\mathbb{R}^{|\mathcal{A}|}$. The term $\sigma_k^2(A)$ denotes the k-th largest singular value of matrix A. A function $f(x)$ is categorized based on the complexity notation as follows: $f(x) = O(g(x))$ if there exists $C > 0$ and x_0 such that $f(x) \leq Cg(x)$ holds for all $x \geq x_0$, $f(x) = \Omega(g(x))$ if there exists $C > 0$ and x_0 such that $f(x) \geq C \cdot g(x)$ for all $x \geq x_0$, $f(x) = o(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$, and $f(x) = \widetilde{O}(g(x))$ if $f(n) = O(g(x) \cdot \log^k(g(x)))$ for some finite k. The vector e_1 is defined as the standard basis vector of proper dimension with the first component being 1. For a finite-dimensional vector x, the norm $\|x\|_1$ refers to its ℓ_1-norm, while $\|x\|$ refers to the ℓ_2-norm, unless otherwise specified. We also define $\|x\|_\Sigma = \sqrt{x^\intercal \Sigma x}$ for a positive definite matrix Σ. For a

[♮] In our paper, we regard the reward function as a welfare function in social choice theory.

matrix M, the norm $\|M\|_F$ denotes the Frobenius norm of M. The multinomial distribution is denoted by Multinomial$(p_1, \ldots, p_n)$, where $p_1, \ldots, p_n$ are the probabilities of outcomes for each of the n categories, respectively, with $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for all $i \in [n]$. Kullback-Leibler (KL) divergence between two probability distributions $P, Q \in \Delta(X)$ is defined as $\sum_{x \in \mathrm{supp}(X)} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$.

2 Preliminaries

Most existing RLHF processes (for language model fine-tuning) consist of two main stages: (1) learning a model of human rewards (oftentimes from preference data), and (2) fine-tuning with the reference policy through Reinforcement Learning algorithms, e.g., Proximal Policy Optimization (PPO) (Schulman et al., 2017). It may also be possible to avoid the explicit learning of reward functions while fine-tuning the policy directly from preference data (Rafailov et al., 2024). We will introduce both approaches in detail below, and will seek to develop frameworks that will cover both cases.

Markov Decision Processes. We define the state s as an element of the set of possible prompts or questions, denoted by $\mathcal{S}$, and the set of actions a, contained in $\mathcal{A}$, as the potential answers or responses to these questions. Consider an RLHF setting with N human labelers (or users), each of whom has their own reward function. This setting can be characterized by a Markov Decision Process (MDP) with N reward functions, represented by the tuple $M = (\mathcal{S}, \mathcal{A}, H, (P_h)_{h \in [H]}, \boldsymbol{r} = (r_i)_{i \in [N]})$, where H denotes the length of the horizon, $P_h : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is the state transition probability at step $h \in [H]$, $\mathcal{T} := (\mathcal{S} \times \mathcal{A})^H$ denotes the set of all possible trajectories, and $r_i : \mathcal{T} \to \mathbb{R}$ is the reward function for individual i and trajectory $\tau \in \mathcal{T}$, representing the utility of human user i from a sequence of responses to a given prompt. We assume $-R_{\max} \leq r_i(\tau) \leq R_{\max}$ for every $\tau \in \mathcal{T}$ and $i \in [N]$, for some $R_{\max} > 0$. This reward model also covers the case that $r_i(\tau) = \sum_{h \in [H]} r_{h,i}(s_h, a_h)$, where $r_{h,i} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the state-action reward function for each step h and individual i, and $\tau = (s_1, a_1, s_2, a_2, \ldots, s_H, a_H)$. The MDP concludes at an absorbing termination state with zero reward after H steps.

A policy $\pi_h : (\mathcal{S} \times \mathcal{A})^{h-1} \times \mathcal{S} \to \Delta(\mathcal{A})$ is defined as a function mapping trajectories to distributions over actions for each step $h \in [H]$ within the horizon H. We define the history-dependent policy class as Π. The collection of these policies across all steps is denoted by π. The expected cumulative reward of a policy π is given by $J(\pi; r_i) := \mathbb{E}_{\tau, \pi}[r_i(\tau)]$ where the expectation in the formula is taken over the distribution of the trajectories under the policy π. Trajectory occupancy measures, denoted by $d_\pi : \mathcal{T} \to [0,1]$, are defined as $d_\pi(\tau) := \mathbb{P}_\pi(\tau)$, which denotes the probability of generating trajectory τ following policy π.

Relationship between Preference and Reward Function. For the MDP with $M = (\mathcal{S}, \mathcal{A}, H, (P_h)_{h \in [H]}, \boldsymbol{r} = (r_i)_{i \in [N]})$, if we compare two trajectories τ_0 and τ_1, we define some random variable o such that $o = 0$ if $\tau_0 \succ \tau_1$, and $o = 1$ if $\tau_0 \prec \tau_1$. Here, $\tau_0 \succ \tau_1$ indicates that τ_0 is preferred than τ_1. We assume that $P_{r_i}(o = 0 \mid \tau_0, \tau_1) = \Phi(r_i(\tau_0) - r_i(\tau_1))$ for all $i \in [N]$, where $\Phi : \mathbb{R} \to [0,1]$ is a monotonically increasing link function, which satisfy $\Phi(x) + \Phi(-x) = 1$ and $\log \Phi(x)$ is a strongly convex function. For example, $\Phi(x) = \sigma(x)$ indicates the BTL model (Definition 2.1 below), a frequently used model for the relationship between preference and reward. Also, we define $P_{\boldsymbol{r}}(\cdot \mid \tau_0, \tau_1) := (P_{r_1}(\cdot \mid \tau_0, \tau_1)^\intercal, \ldots, P_{r_N}(\cdot \mid \tau_0, \tau_1)^\intercal)^\intercal$. We call $P_{\boldsymbol{r}}$ and P_{r_i} a preference probability vector induced by the reward vector $\boldsymbol{r}$ and the reward r_i.

Definition 2.1. *The Plackett-Luce (PL) model (Plackett, 1975; Luce, 2005) quantifies the likelihood that a trajectory τ_k is preferred over all other pairs in the set $\{\tau_k\}_{k\in[K]}$ by assigning it a probability defined as*

$$P_r\left(\tau_k \succ \tau_{k'} \forall k' \neq k \,\middle|\, (\tau_k)_{k\in[K]}\right) = \frac{\exp(r(\tau_k))}{\sum_{k'\in[K]} \exp(r(\tau_{k'}))}$$

where r is the reward function for a human labeler. In the case where $k = 2$, this formulation simplifies to the Bradley-Terry-Luce (BTL) Model (Bradley and Terry, 1952).

Direct Preference Optimization (DPO) (Rafailov et al., 2024). Consider the case with Markovian reward and policy, i.e., the reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a function of state s and action a, and the policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is also depending only on the state s. Also, assume that we compare actions for each state rather than the whole trajectories. In the fine-tuning phase using RL, when KL-regularization with the reference policy π^{old} is employed, the optimal policy is given by:

$$\pi(a \mid s) = \frac{1}{Z(s)} \pi^{\mathrm{old}}(a \mid s) \exp\left(\frac{r(s,a)}{\beta}\right),$$

where $Z(s)$ serves as a normalization factor that is independent of the answer a, and β represents the coefficient for KL regularization. Integrating the BTL model into this framework yields:

$$\pi^{\mathrm{RLHF}} = \arg\min_{\pi} -\mathbb{E}_{(s,a_0)\succ(s,a_1)}\left[\log\sigma\left(\beta \log \frac{\pi(a_0 \mid s)}{\pi^{\mathrm{old}}(a_0 \mid s)} - \beta \log \frac{\pi(a_1 \mid s)}{\pi^{\mathrm{old}}(a_1 \mid s)}\right)\right],$$

where σ denotes the Sigmoid function (Rafailov et al., 2024). This formulation bypasses the step of explicitly estimating the reward function.

Fundamentals of Auction Theory. Consider the sealed-bid auction mechanism (Vickrey, 1961), where each participant $i \in [N]$ privately submits a bid $b_i(x)$ for every possible outcome $x \in X$, whose true value is $p_i(x) \in \mathbb{R}$. An auction is termed a Dominant Strategic Incentive-Compatible (DSIC) auction (Roughgarden, 2010) if revealing each participant's true valuation is a weakly dominant strategy, i.e., an individual's optimal strategy is to bid their true valuation of the item, $b_i(x) = p_i(x)$ for all $x \in X$, irrespective of the bids $b_{-i}(x)$ submitted by others for all $x \in X$. This mechanism is also called a *truthful* mechanism (Roughgarden, 2010). An auction has a social-welfare-maximizing allocation rule (Roughgarden, 2010) if the outcome x is $\arg\max_{x\in X} \sum_{i\in[N]} p_i(x)$.

3 Personalized RLHF via Representation Learning

In this section, we provide the first approach in the personalization-based framework, based on representation learning.

Reward Function Class. We will assume that we have access to a pre-trained feature function $\phi : \mathcal{T} \to \mathbb{R}^d$, which encodes a trajectory of states and actions (i.e., questions and answers) to a d-dimensional feature vector. This covers the case where feature $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is defined at each state-action pair, i.e., $\phi(\tau) := \sum_{h\in[H]} \phi_h(s_h, a_h)$ for trajectory $\tau = (s_1, a_1, \ldots, s_H, a_H)$. For example, it is common to use the penultimate layer of an existing pre-trained LLM or other pre-trained backbones to encode a long sentence to a feature vector (Donahue et al., 2014; Gulshan et al., 2016; Tang et al., 2015).

Our first goal is to learn reward models for each human user using preference datasets. First, we define the reward function class as

$$\mathcal{G}_{\boldsymbol{r}} = \left\{ (\langle \psi_\omega(\phi(\cdot)), \theta_i \rangle)_{i\in[N]} \,\middle|\, \psi_\omega \in \Psi, \theta_i \in \mathbb{R}^k \text{ and } \|\theta_i\|_2 \le B \text{ for all } i \in [N] \right\},$$

for some $B > 0$, where Ψ is the set of representation functions parameterized by $\omega \in \Omega$, i.e., $\Psi = \{\psi_\omega \mid \omega \in \Omega\}$, where $\psi_\omega : \mathbb{R}^d \to \mathbb{R}^k$. We assume that $d \gg k$. We denote $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$, and to emphasize the relationship between reward and $(\omega, \boldsymbol{\theta})$, we will write $r_{\omega,\theta_i}(\cdot) := \langle \psi_\omega(\phi(\cdot)), \theta_i \rangle$ for each individual $i \in [N]$ and $\boldsymbol{r}_{\omega,\boldsymbol{\theta}}(\cdot) := (r_{\omega,\theta_1}(\cdot), \cdots, r_{\omega,\theta_N}(\cdot))^\intercal \in \mathbb{R}^N$. From this section, we will write $\boldsymbol{r}^\star = (r_1^\star, \dots, r_N^\star)$ as the underlying human reward functions.

Assumption 3.1 (Realizability)**.** *We assume that the underlying true reward can be represented as $r_i^\star(\cdot) = \langle \psi^\star(\phi(\cdot)), \theta_i^\star \rangle$ for some representation function $\psi^\star \in \Psi$ (in other words, there exists some $\omega^\star \in \Omega$ such that $\psi_{\omega^\star} = \psi^\star$) and $\|\theta_i^\star\|_2 \le B$ for each individual $i \in [N]$.*

To emphasize $(\omega, \boldsymbol{\theta})$, we define shorthand notation $P_{\omega,\boldsymbol{\theta}} := P_{r_{\omega,\boldsymbol{\theta}}}$ as the preference probability induced by $r_{\omega,\boldsymbol{\theta}}$. We also write $P_{\omega,\theta} := P_{\langle \psi_\omega(\phi(\cdot)),\theta \rangle}$, which is the probability induced by $\langle \psi_\omega(\phi(\cdot)), \theta \rangle$.

3.1 Algorithms

We introduce our algorithms for learning personalized policy. Compared to traditional RLHF algorithms (Ziegler et al., 2019; Ouyang et al., 2022; Zhu et al., 2023), we consider personalized reward function by representation learning.

Algorithm 1 outputs a joint estimation of $\psi^\star$ and $\boldsymbol{\theta}^\star$ with maximum likelihood estimation (MLE), together with personalized policies. The input of the algorithm is $\widehat{\mathcal{D}} = \cup_{i\in[N]} \widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$. Here, $\tau_{i,t}^{(j)}$ is sampled from the distribution μ_t for $t = 0, 1$, and $o_i^{(j)} \sim P_{r_i^\star}(\cdot|\tau_0^{(j)}, \tau_1^{(j)})$. We regard the reward estimation of human users as a multi-task representation learning problem (Du et al., 2020; Tripuraneni et al., 2020). After estimating the reward functions, we construct two kinds of confidence sets for the reward function and outputs based on each confidence set:

- Confidence set (Equation (3.1)) for the MLE estimation as Liu et al. (2022), which is also used in (Liu et al., 2023; Zhan et al., 2023; Wang et al., 2024; Zhan et al., 2022), with $\zeta = C_1 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)$ for a constant $C_1 > 0$, which will be related to Theorem 3.1. the definition of bracketing number ($\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}$) is deferred to Appendix B.

- Confidence set (Equation (3.2)) with $\zeta' = C_8 \left(k \frac{\xi^2 \kappa^2 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)}{\eta^2 N N_p} + \frac{\xi^2 (k + \log(N/\delta))}{\eta^2 N_p} + \lambda B^2 \right)$, where $C_8, \lambda > 0$ are constants, $\xi := \max_{x\in[-2R_{\max}, 2R_{\max}]} \left| \frac{\Phi'(x)}{\Phi(x)} \right|$, $\kappa := (\min_{x\in[-2R_{\max}, 2R_{\max}]} \Phi'(x))^{-1}$, and $\eta := \min_{x\in[-2R_{\max}, 2R_{\max}]} \left(\frac{\Phi'(x)^2 - \Phi''(x)\Phi(x)}{\Phi(x)^2} \right)$. In the case that $\Phi(x) = \sigma(x)$ (i.e. Φ is a Sigmoid function), $\xi \le 1$ and $\kappa = \eta = \frac{1}{2 + \exp(-2R_{\max}) + \exp(2R_{\max})}$. This confidence set will be related to Theorem 3.4. With this confidence set, we can get sharper bound with Assumptions 3.2 - 3.4.

Lastly, we find the best policy based on the pessimistic expected value function. $\mu_{i,\text{ref}}$ in Algorithm 1 is a known reference trajectory distribution for individual $i \in [N]$, and it can be set as μ_1.

Algorithm 1 Personalized RLHF via Representation Learning

Input: Dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$ is the preference dataset for the ith individual.

Estimate $\omega^\star$ and $\boldsymbol{\theta}^\star$ by

$$(\widehat{\omega},\widehat{\boldsymbol{\theta}}) \leftarrow \mathop{\arg\min}_{\omega\in\Omega, \|\theta_i\|_2\le B \text{ for all } i\in[N]} \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\omega,\theta_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})$$

Construct a confidence set of the reward function by

$$\mathcal{R}(\widehat{\mathcal{D}}) \leftarrow \left\{ \boldsymbol{r}_{\omega,\boldsymbol{\theta}} \,\middle|\, \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\omega,\theta_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) \ge \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\widehat{\omega},\widehat{\theta}_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) - \zeta \right\} \tag{3.1}$$

or

$$\mathcal{R}'(\widehat{\mathcal{D}}) \leftarrow \cap_{i\in[N]}\left\{ \boldsymbol{r}_{\omega,\boldsymbol{\theta}} \,\middle|\, \frac{1}{N_p}\sum_{j\in[N_p]} \left|(r_{\widehat{\omega},\widehat{\theta}_i}(\tau_{i,0}^{(j)}) - r_{\widehat{\omega},\widehat{\theta}_i}(\tau_{i,1}^{(j)})) - (r_{\omega,\theta_i}(\tau_{i,0}^{(j)}) - r_{\omega,\theta_i}(\tau_{i,1}^{(j)}))\right|^2 \le \zeta' \right\} \tag{3.2}$$

Compute policy with respect to $\mathcal{R}(\widehat{\mathcal{D}})$ (or $\mathcal{R}'(\widehat{\mathcal{D}})$) for all $i \in [N]$ by

$$\widehat{\pi}_i \leftarrow \mathop{\arg\max}_{\pi\in\Pi} \min_{\boldsymbol{r}\in\mathcal{R}(\widehat{\mathcal{D}})} \Big(J(\pi; r_i) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i(\tau)]\Big) \tag{3.3}$$

or

$$\widehat{\pi}_i' \leftarrow \mathop{\arg\max}_{\pi\in\Pi} \min_{\boldsymbol{r}\in\mathcal{R}'(\widehat{\mathcal{D}})} \Big(J(\pi; r_i) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i(\tau)]\Big) \tag{3.4}$$

Output: $(\widehat{\omega},\widehat{\boldsymbol{\theta}},(\widehat{\pi}_i)_{i\in[N]})$ or $(\widehat{\omega},\widehat{\boldsymbol{\theta}},(\widehat{\pi}_i')_{i\in[N]})$.

Algorithm 2 addresses a scenario where a new human user, who was not a labeler before, aims to learn their own reward models using representations previously learned by other human users, focusing solely on learning $\theta_0^\star$. They leverage the learned representation $\psi_{\widehat{\omega}}$ from Algorithm 1. The input of the algorithm is $\widehat{\mathcal{D}}_0 = \{(o_0^{(j)}, \tau_{0,0}^{(j)}, \tau_{0,1}^{(j)})_{j\in[N_p]}\}$. Algorithm 2 provides an estimation of $\theta_0^\star$ with MLE using the frozen representation $\psi_{\widehat{\omega}}$. Similarly, after estimating the reward function, we construct confidence set for the MLE estimation with $\zeta = C_8\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2 NN_p} + \frac{\xi^2(k+\log(1/\delta))}{\eta^2 N_p} + \lambda B^2\right)$ for a constant $C_8 > 0$. Lastly, we find the best policy based on the pessimistic expected value function. $\mu_{0,\mathrm{ref}}$ in Algorithm 2 is a known reference trajectory distribution.

3.2 Results and Analyses

For ease of analysis, we consider the case where the sizes of preference datasets for each individual $i \in \{0\}\cup[N]$ are identical, i.e., $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$, satisfies $|\widehat{\mathcal{D}}_i| = N_p$ for all $i \in \{0\}\cup[N]$. The result in this section can also be extended to the case with $|\widehat{\mathcal{D}}_i| = N_{p,i}$ for each individual i. We defer all the proofs of this section to Appendix C.

Algorithm 2 Transferable RLHF for a New Human User via Representation Learning

Input: Dataset $\widehat{\mathcal{D}}_0 = \{(o_0^{(j)}, \tau_{0,0}^{(j)}, \tau_{0,1}^{(j)})_{j\in[N_p]}\}$ and $\widehat{\omega}$ from Algorithm 1.
Estimate $\theta_0^\star$ by

$$\widehat{\theta}_0 \leftarrow \underset{\|\theta_0\|_2\leq B}{\arg\min} \sum_{j\in[N_p]} \log P_{\widehat{\omega},\theta_0}(o_0^{(j)} \mid \tau_{0,0}^{(j)}, \tau_{0,1}^{(j)})$$

Construct a confidence set of the reward function by

$$\mathcal{R}(\widehat{\mathcal{D}}) \leftarrow \left\{ r_{\omega,\theta_0} \,\Big|\, \frac{1}{N_p} \sum_{j\in[N_p]} \left|(r_{\widehat{\omega},\widehat{\theta}_0}(\tau_{0,0}^{(j)}) - r_{\widehat{\omega},\widehat{\theta}_0}(\tau_{i,1}^{(j)})) - (r_{\omega,\theta_0}(\tau_{0,0}^{(j)}) - r_{\omega,\theta_0}(\tau_{0,1}^{(j)}))\right|^2 \leq \zeta \right\}$$

Compute policy with respect to $\mathcal{R}(\widehat{\mathcal{D}})$ by

$$\widehat{\pi}_0 \leftarrow \underset{\pi\in\Pi}{\arg\max} \min_{r_0\in\mathcal{R}(\widehat{\mathcal{D}}_0)} \Big(J(\pi; r_0) - \mathbb{E}_{\tau\sim\mu_{0,\mathrm{ref}}}[r_0(\tau)] \Big)$$

Output: $(\widehat{\pi}_i)_{i\in[N]}$.

Definition 3.1 (Concentrability Coefficient)**.** *The concentrability coefficient, with respect to a reward vector class $\mathcal{G}_{\boldsymbol{r}}$, human user i, a target policy π_{tar} (which policy to compete with, which potentially can be the optimal policy $\pi_i^\star$ corresponding to $r_i^\star$), and a reference policy μ_{ref}, is defined as follows:*

$$C_{\boldsymbol{r}}\Big(\mathcal{G}_{\boldsymbol{r}}, \pi_{tar}, \mu_{ref}, i\Big) := \max\left\{ 0, \sup_{r\in\mathcal{G}_r} \frac{\mathbb{E}_{\tau_0\sim\pi_{tar}, \tau_1\sim\mu_{ref}}\left[r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i(\tau_0) + r_i(\tau_1)\right]}{\sqrt{\mathbb{E}_{\tau_0\sim\mu_0,\tau_1\sim\mu_1}\left[\left|r_i^\star(\tau_0) - r_i^\star(\tau_1) - r_i(\tau_0) + r_i(\tau_1)\right|^2\right]}} \right\}.$$

We also define the concentrability coefficient of the reward scalar class in Appendix B.2, and we denote this as $C_r(\mathcal{G}_r, \pi_{tar}, \mu_{ref})$.

(Zhan et al., 2023) provides an interpretation of concentrability coefficient. For example, if $\mu_{\mathrm{ref}} = \mu_1$, the value of $C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{\mathrm{tar}}, \mu_1, i) \leq \sqrt{\max_{\tau\in\mathcal{T}} \frac{d_{\pi_{\mathrm{tar}}}(\tau)}{\mu_0(\tau)}}$, so this reflects the concept of "single-policy concentrability" (Rashidinejad et al., 2021; Zanette et al., 2021; Ozdaglar et al., 2023), which is commonly assumed to be bounded in the offline RL literature.

We present the gap of the expected value function between the target policy $\pi_{i,\mathrm{tar}}$ and the estimated policy $\widehat{\pi}_i$ for each individual $i \in [N]$. Here, $\pi_{i,\mathrm{tar}}$, which may be the optimal policy $\pi_i^\star$ over $r_i^\star$, serves as the policy that $\widehat{\pi}_i$ will compete.

Theorem 3.1. (Total Expected Value Function Gap). *Suppose Assumption 3.1 holds. For any $\delta \in (0,1]$, with probability at least $1-\delta$, the output $(\widehat{\pi}_i)_{i\in[N]}$ of Algorithm 1 satisfies*

$$\sum_{i\in[N]} \Big(J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \Big) \leq \sqrt{\frac{c\kappa^2 N C_{\max}^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/NN_p)/\delta)}{N_p}},$$

where $C_{\max} := \max_{i\in[N]} C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)$ and $c > 0$ is a constant.

Corollary 3.2. (Expected Value Function Gap). *Suppose Assumption 3.1 holds. For any $\delta \in (0,1]$ and all $i \in [N]$, with probability at least $1-\delta$, the output $\widehat{\pi}_i$ of Algorithm 1 satisfies*

$$J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \leq \sqrt{\frac{c\kappa^2 C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)^2 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/NN_p)/\delta)}{N_p}},$$

where $c > 0$ is a constant.

Note that the results above do not need any assumption on $(\theta_i^\star)_{i\in[N]}$. Still, as $N_p \to \infty$, $\widehat{\pi}_i$ has comparable or better performance than the comparator policy $\pi_{i,\text{tar}}$, which approaches the optimal policy if $\pi_{i,\text{tar}} = \pi_i^\star$. We defer the proofs of Theorem 3.1 and Corollary 3.2 to Appendix C.1.

Next, we want to improve the bound for Corollary 3.2, as the gap of the expected value function does not decay with N, which is the number of human users. Therefore, we consider the case that $(\theta_i)_{i\in[N]}$ are diverse (Assumption 3.2), which is critical for improving the sample complexity of Algorithm 1 by outputting $(\widehat{\pi}_i')_{i\in[N]}$. We will additionally assume the uniqueness of the representation up to the orthonormal linear transformation (Assumption 3.3), and uniform concentration of covariance (Assumption 3.4). These assumptions are commonly used in multi-task learning (Du et al., 2020; Tripuraneni et al., 2021; Lu et al., 2021)

Assumption 3.2 (Diversity). *The matrix $\Theta^\star = [\theta_1^\star, \cdots, \theta_N^\star] \in \mathbb{R}^{k\times N}$ satisfies $\sigma_k^2(\Theta^\star) \geq \Omega(N/k)$.*

Assumption 3.2 means that θ_i is evenly distributed in $\mathbb{R}^d$ space for $i \in [N]$, which indicates "diverse" human reward function.

Assumption 3.3 (Uniqueness of Representation (up to Orthonormal-Transformation)). *For any representation functions $\psi, \psi' \in \Psi$ and $\epsilon > 0$, if there exists $\{v_i\}_{i=1}^T, \{v_i'\}_{i=1}^T$, and a trajectory distribution μ that satisfy*

$$\frac{1}{T}\sum_{i\in[T]} \mathbb{E}_{\tau\sim\mu} \|\psi(\phi(\tau))^\top v_i - \psi'(\phi(\tau))^\top v_i'\|^2 \leq \epsilon,$$

$W = [v_1, v_2, \cdots, v_T] \in \mathbb{R}^{k\times T}$ satisfies $\sigma_k^2(W) \geq \Omega(T/k)$, and $\|v_i\|_2 \leq B$ for all $i \in [T]$. Then, there exists a constant orthonormal matrix P such that

$$\|\psi(\phi(\tau)) - P\psi'(\phi(\tau))\|^2 \leq ck\epsilon/B$$

for all trajectory τ where $c > 0$ is a constant.

This assumption posits that if two representation functions, ψ and ψ', yield sufficiently small differences in expected squared norms of their inner products with corresponding vectors over trajectory distributions, then they are related by a constant orthonormal transformation. If $\psi_\omega(\phi(s,a)) := \omega\phi(s,a)$ where ω is $k \times d$ orthonormal matrix, we can prove that Assumption 3.3 holds with non-degenerate $\phi(s,a)$ distribution (Appendix C.2.2).

Definition 3.2. *Given distributions μ_0, μ_1 and two representation functions $\psi, \psi' \in \Psi$, define the covariance between ψ and ψ' with respect to μ_0, μ_1 to be*

$$\Sigma_{\psi,\psi'}(\mu_0,\mu_1) := \mathbb{E}_{\tau_0\sim\mu_0,\tau_1\sim\mu_1}[(\psi(\phi(\tau_0)) - \psi(\phi(\tau_1)))(\psi'(\phi(\tau_0)) - \psi'(\phi(\tau_1)))^\intercal] \in \mathbb{R}^{k\times k}.$$

Define the symmetric covariance as

$$\Lambda_{\psi,\psi'}(\mu_0,\mu_1) = \begin{bmatrix} \Sigma_{\psi,\psi}(\mu_0,\mu_1) & \Sigma_{\psi,\psi'}(\mu_0,\mu_1) \\ \Sigma_{\psi',\psi}(\mu_0,\mu_1) & \Sigma_{\psi,\psi'}(\mu_0,\mu_1) \end{bmatrix}.$$

We make the following assumption on the concentration property of the representation covariances.

Assumption 3.4. (Uniform Concentrability). *For any $\delta \in (0,1]$, there exists a number $N_{unif}(\Psi, \mu_0, \mu_1, \delta)$ such that for any $n \geq N_{unif}(\Psi, \mu_0, \mu_1, \delta)$, the empirical estimation $\widehat{\Lambda}_{\psi,\psi'}(\mu_0, \mu_1)$ of $\Lambda_{\psi,\psi'}(\mu_0, \mu_1)$ based on n independent trajectory sample pairs from distributions (μ_0, μ_1), with probability at least $1-\delta$, will satisfy the following inequality:*

$$1.1\Lambda_{\psi,\psi'}(\mu_0, \mu_1) \geq \widehat{\Lambda}_{\psi,\psi'}(\mu_0, \mu_1) \geq 0.9\Lambda_{\psi,\psi'}(\mu_0, \mu_1),$$

for all $\psi, \psi' \in \Psi$.

Assumption 3.4 means that the empirical estimate $\widehat{\Lambda}_{\psi,\psi'}(\mu_0, \mu_1)$ closely approximates the true $\Lambda_{\psi,\psi'}(\mu_0, \mu_1)$ with high probability. Similarly, if $\psi_\omega(\phi(\tau)) := \omega\phi(\tau)$, $N_{\mathrm{point}}(\Psi, \mu_0, \mu_1, \delta) = \widetilde{\mathcal{O}}(d)$ (Du et al., 2020, Claim A.1). If distributions μ_0, μ_1 are clear from the context, we omit the notation μ_0, μ_1 for $\Sigma_{\psi,\psi'}(\mu_0, \mu_1)$ and $\Lambda_{\psi,\psi'}(\mu_0, \mu_1)$. Moreover, we also write $\Sigma_{\psi,\psi}$ as Σ_ψ for notational convenience.

With Assumption 3.2 and Assumption 3.3, $\psi^\star$ and ψ_ω are close up to an orthonormal matrix transformation, as asserted below:

Corollary 3.3. (Closeness between $\psi^\star$ and ψ_ω). *Suppose Assumptions 3.1, 3.2, and 3.3 hold. For any $\delta \in (0,1]$, with probability at least $1-\delta$, if $\boldsymbol{r}_{\omega,\boldsymbol{\theta}} \in \mathcal{R}'(\mathcal{D})$ as specified in Algorithm 1, then there exists an orthonormal matrix P_ω such that*

$$\left[\|\psi^\star(\phi(\tau_0)) - \psi^\star(\phi(\tau_1)) - P_\omega(\psi_\omega(\phi(\tau_0)) - \psi_\omega(\phi(\tau_1)))\|^2\right] \leq k\frac{c_{rep}\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_pB^2}$$

for all τ_0, τ_1, where $c_{rep} > 0$ is a constant.

We provide the proof of Corollary 3.3 in Appendix C.3. Using Corollary 3.3, now we have a better sample complexity compared with that in Theorem 3.1, as stated below:

Theorem 3.4. (Improved Expected Value Function Gap). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. For any $\delta \in (0,1]$, all $i \in [N]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $\widehat{\pi}_i'$ of Algorithm 1 satisfies*

$$\begin{aligned} &J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i'; r_i^\star) \\ &\leq \sqrt{cC_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)^2\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2NN_p} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2N_p} + \lambda B^2\right)} \end{aligned} \tag{3.5}$$

where $c > 0$ is a constant.

Lastly, we can also use the learned representation for a new human user as follows:

Theorem 3.5. (Expected Value Function Gap for a New Human User). *Suppose Assumptions 3.2, 3.3, and 3.4 hold. For any $\delta \in (0,1]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $\widehat{\pi}_0$ of Algorithm 2 satisfies*

$$\begin{aligned} &J(\pi_{0,tar}; r_0^\star) - J(\widehat{\pi}_0; r_0^\star) \\ &\leq \sqrt{cC_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)^2\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2NN_p} + \frac{\xi^2(k+\log(1/\delta))}{\eta^2N_p} + \lambda B^2\right)} \end{aligned}$$

where $c > 0$ is a constant.

We provide the proofs of Theorem 3.4 and Theorem 3.5 in Appendix C.4.

Remark 1 (Sample Complexity)**.** *For Theorem 3.4, if we naively learn the personalization model without representation learning,* $\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))$ *will be very large. For example, if we use linear representation* $\phi_\omega(x) = \omega x$ *and* ω *is a* $d \times k$ *orthonormal matrix, then* $\log(\mathcal{N}_{\mathcal{G}_r}(1/NN_p)/\delta) \le \mathcal{O}\Big((dk+Nk)\log\Big(R_{\max}NN_p/\delta\Big)\Big)$ *while naive personalization with*

$$\mathcal{G}_{\boldsymbol{r}}' = \Big\{(\langle\phi(\cdot),\theta_i\rangle)_{i\in[N]}\rangle \,\Big|\, \theta_i \in \mathbb{R}^d \text{ and } \|\theta_i\|_2 \le B \text{ for all } i \in [N]\Big\}$$

provides $\mathcal{N}_{\mathcal{G}_r'}(1/(NN_p)) \le \mathcal{O}\Big(Nd\log\Big(R_{\max}NN_p/\delta\Big)\Big)$*. Since* $d \gg k$*, the bound of Equation (3.5)'s right-hand side has a significant improvement when we use the representation learning. If the representation function class is an MLP class, we can use a known bracket number by Bartlett et al. (2017).*

We also point out that the existing technique from representation learning literature does not cover the case with general representation function learning with a log-likelihood loss function with $\mathcal{O}(1/N_p)$ rate, to the best of our knowledge.

Lastly, we examine the tightness of our analysis by the theoretical lower bound of the sub-optimality gap of personalization.

Theorem 3.6. (Lower Bound for the Sub-Optimality Gap of Personalization). *For any* $k > 6, N_p \ge Ck\Lambda^2$ *and* $\Lambda \ge 2$*, there exists a representation function* $\phi(\cdot)$ *so that*

$$\min_{i\in[N]}\inf_{\widehat{\pi}}\sup_{Q\in\mathrm{CB}(\Lambda)}\Big(\max_{\pi^*\in\Pi} J(\pi^*; r_{\omega,\theta_i}) - J(\widehat{\pi}; r_{\omega,\theta_i})\Big) \ge C\Lambda\cdot\sqrt{\frac{k}{N_p}},$$

where

$$\mathrm{CB}(\Lambda) := \Big\{Q := \Big(\{\mu_0,\mu_1\},\{\tau_{i,0}^{(j)},\tau_{i,1}^{(j)}\}_{i\in[N],j\in[N_p]},\omega,\boldsymbol{\theta}\Big)\,\Big|\, C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}},\pi^\star,\mu_1,i) \le \Lambda \text{ for all } i\in[N]\Big\}$$

is the family of MDP with N *reward functions and* $H = 1$ *instances.*

The proof is largely adapted from (Zhu et al., 2023, Theorem 3.10). By Theorem 3.6, for general representation function class, we establish that Algorithm 1 is near-optimal for the sub-optimality of the induced personalization policy, as $\log(\mathcal{N}_{\mathcal{G}_r}(1/NN_p))$ can be small so that $\sqrt{k\frac{\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)}{NN_p}}$ can be dominated by $\sqrt{\frac{k}{N_p}}$. Note that if Ψ is a linear representation class, our result for personalization (Theorem 3.4) still has a $\sqrt{k}$ gap compared to the lower bound (Theorem 3.6). This gap is also observed in (Tripuraneni et al., 2020). We will leave the sharpening of this $\sqrt{k}$ factor as a future work.

4 Personalized RLHF via Human User Clustering

We now provide the second approach in the personalization-based framework, through human user clustering. In particular, fine-tuning an LLM for each individual may be impractical. We thus propose an alternative approach that segments human users into clusters and fine-tunes an LLM for *each cluster*. This strategy entails deploying K clustered models, which can be smaller than the number of human users N. A critical aspect of this methodology is the way to generate clusters. This clustering-based personalization has also been studied in the federated (supervised) learning literature (Mansour et al., 2020; Ghosh et al., 2020; Sattler et al., 2020). We introduce our algorithm next, based on the algorithmic idea in Mansour et al. (2020).

4.1 Algorithms

We partition all the N human users into K clusters and find the best parameters for each cluster as follows:

$$\max_{(r_{(k)})_{k\in[K]}} \sum_{i\in[N]} \frac{1}{N} \max_{k\in[K]} \mathbb{E}_{\mathcal{D}_i}\left[\log P_{r_{(k)}}(o_i \mid \tau_{i,0}, \tau_{i,1})\right].$$

Since we only have access to the empirical data distribution, we instead solve the following problem:

$$\max_{(r_{(k)})_{k\in[K]}} \sum_{i\in[N]} \frac{1}{N} \max_{k\in[K]} \sum_{j\in[N_p]} \log P_{r_{(k)}}\left(o_i^{(j)} \,\middle|\, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\right). \tag{4.1}$$

Algorithm 3 outputs K clustered policies and a map from human users to clusters. The input of the algorithm is $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$, which is the same as Algorithm 1. After estimating the representation parameter $\widehat{\omega}$, the algorithm will estimate the reward function parameters $(\widehat{\theta}_{(k)})_{k\in[K]}$ with Equation (4.2). Lastly, we find the best policy based on expected value function.

We also provide a practical algorithm that uses DPO and EM (Moon, 1996) algorithms to solve Equation (4.2). Given the inherent complexity of this hierarchical optimization problem, which presents more challenges than standard optimization tasks (Anandalingam and Friesz, 1992), we propose a novel algorithm that circumvents the need for explicit reward function estimation in Algorithm 4. Our approach begins by randomly assigning each human user to a cluster. Subsequently, we reassign random human users to the cluster where the policy most effectively maximizes their empirical DPO loss (Equation (4.3)). Finally, we refine our solution by optimizing the DPO loss function for the selected human users within each cluster, thereby enhancing the overall policy effectiveness.

4.2 Results and Analyses

First, we introduce label discrepancy (Mohri and Muñoz Medina, 2012) for the preference dataset and reward function class. We defer all the proofs of this section to Appendix D.

Definition 4.1 (Label Discrepancy). *Label discrepancy for preference distribution $\boldsymbol{D}_i$ and $\boldsymbol{D}_j$, which are distributions of (o, τ_0, τ_1), with reward function class $\mathcal{G}_r$ is defined as follows:*

$$disc(\boldsymbol{D}_i, \boldsymbol{D}_j, \mathcal{G}_r) = \max_{r\in\mathcal{G}_r}\left|\mathbb{E}_{\boldsymbol{D}_i} \log P_r(o \mid \tau_1, \tau_0) - \mathbb{E}_{\boldsymbol{D}_j} \log P_r(o \mid \tau_1, \tau_0)\right|.$$

Algorithm 3 Personalized RLHF via Clustering

Input: Dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$ is the preference dataset for the ith individual, and $\widehat{\omega}$ form Algorithm 1.

Learn $\theta_{(i)}$ and the clustering map $f : [N] \to [K]$ by

$$(\widehat{\theta}_{(k)})_{k\in[K]} \leftarrow \underset{\|\theta_{(k)}\|_2\le B \text{ for all } k\in[K]}{\arg\max} \sum_{i\in[N]} \max_{k\in[K]} \sum_{j\in[N_p]} \log P_{\widehat{\omega},\theta_{(k)}}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) \tag{4.2}$$

$$\widehat{f}(i) \leftarrow \underset{k\in[K]}{\arg\max} \sum_{j\in[N_p]} \log P_{\widehat{\omega},\widehat{\theta}_{(k)}}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) \text{ for all } i\in[N]$$

For each $k \in [K]$,

$$\widehat{\pi}_{(k)} \leftarrow \underset{\pi\in\Pi}{\arg\max}\Big(J(\pi; r_{\widehat{\omega},\widehat{\theta}_{(k)}}) - \mathbb{E}_{\tau\sim\mu_1}[r_{\widehat{\omega},\widehat{\theta}_{(k)}}(\tau)]\Big).$$

Output: $((\widehat{\pi}_{(k)})_{k\in[K]}, (\widehat{\theta}_{(k)})_{k\in[K]}, \widehat{\omega}, \widehat{f})$.

Lemma 1 (Mansour et al. (2020)). *For any $\delta \in (0,1]$, with probability at least $1-\delta$, the output $((\widehat{\pi}_{(k)})_{k\in[K]}, (\widehat{\theta}_{(k)})_{k\in[K]}, \widehat{\omega}, \widehat{f})$ of Algorithm 3 satisfies*

$$\max_{\|\theta_i'\|\le B \text{ for all } i\in[N]} \sum_{i\in[N]} \sum_{j\in[N_{p,i}]} \log\left(\frac{P_{\widehat{\omega},\theta_i'}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})}{P_{\widehat{\omega},\widehat{\theta}_{\widehat{f}(i)}}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})}\right)$$
$$\le C_{cluster} N N_p\left(\sqrt{\frac{\log(2K/\delta)}{N_p}} + \sqrt{\frac{kK\log(N_p/K)}{N_p}} + \sum_{i\in[N]}\frac{1}{N} disc(\mathcal{D}_i, \mathcal{C}_{\widehat{f}(i)}, \mathcal{G}_{\psi_{\widehat{\omega}}})\right),$$

where $\mathcal{C}_k := \cup_{\widehat{f}(i)=k}\mathcal{D}_i$, $C_{cluster} > 0$ is a constant, and $\mathcal{G}_{\psi_\omega} := \{r_{\omega,\theta} \mid \|\theta\| \le B\}$ for all $\omega \in \Omega$.

Theorem 4.1. (Total Expected Value Function Gap). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. Also, assume that $C_r(\mathcal{G}_r, \pi, \mu_{i,ref}, i) \le C'_{max}$ for all policy π and $i \in [N]$. For any $\delta \in (0,1]$, all $i \in [N]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $((\widehat{\pi}_{(k)})_{k\in[K]}, \widehat{f})$ of Algorithm 3 satisfies*

$$\sum_{i\in[N]}\Big(J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_{\widehat{f}(i)}; r_i^\star)\Big)$$
$$\le cN\kappa\left(\frac{\log(2K/\delta)}{N_p} + \frac{kK\log(N_p/K)}{N_p} + \frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}\right.$$
$$\left. + \left(\sum_{i\in[N]}\frac{1}{N} disc(\mathcal{D}_i, \mathcal{C}_{\widehat{f}(i)}, \mathcal{G}_{\psi^\star}))\right)^2 + \left(\frac{\log(\mathcal{N}_{\mathcal{G}_{\psi^\star}}(1/NN_p)/\delta)}{NN_p}\right)^2\right)^{1/4},$$

where $c > 0$ is a constant.

We note that due to the $\sqrt{dk/N_p}$ order on the right-hand side of Lemma 1, we have a slower rate in Theorem 4.1 than Theorem 3.4. This gap is mainly due to the fact that the analysis of Lemma 1

Algorithm 4 ClusterDPO: Learning K clustered policies by DPO

Input: Dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{a_{i,0}^{(j)} \succ a_{i,1}^{(j)}, s_i^{(j)})_{j\in[N_p]}\}$ is the preference dataset for the ith individual, β is a parameter for DPO

Randomly select K human users $p_1,\dots,p_K$ and initialize $\pi_{(k)}^0$ for all $k \in [K]$ as

$$\pi_{(k)}^0 \leftarrow \arg\max_{\pi\in\Pi} \sum_{j\in[N_p]} \log\sigma\left(\beta\log\frac{\pi(a_{p_k,0}^{(j)} \mid s_{p_k}^{(j)})}{\pi^{\text{old}}(a_{p_k,0}^{(j)} \mid s_{p_k}^{(j)})} - \beta\log\frac{\pi(a_{p_k,1}^{(j)} \mid s_{p_k}^{(j)})}{\pi^{\text{old}}(a_{p_k,1}^{(j)} \mid s_{p_k}^{(j)})}\right)$$

Randomly initialize $f^0(i)$ for $i \notin \{p_1,\dots,p_K\}$

for $t \in [T]$ **do**

  Randomly select K human users $p_1,\dots,p_K$.

  for $i \in [N]$ **do**

    if $i \notin \{p_1,\dots,p_K\}$ **then**

      Define $f^t(i) \leftarrow f^{t-1}(i)$

    end if

  end for

  Assign $f^t(p_k)$ for all $k \in [K]$ as

$$f^t(p_k) \leftarrow \arg\max_{s\in[K]} \sum_{j\in[N_p]} \log\sigma\left(\beta\log\frac{\pi_{(s)}^{t-1}(a_{p_k,0}^{(j)} \mid s_{p_k}^{(j)})}{\pi^{\text{old}}(a_{p_k,0}^{(j)} \mid s_{p_k}^{(j)})} - \beta\log\frac{\pi_{(s)}^{t-1}(a_{p_k,1}^{(j)} \mid s_{p_k}^{(j)})}{\pi^{\text{old}}(a_{p_k,1}^{(j)} \mid s_{p_k}^{(j)})}\right) \tag{4.3}$$

  Run a few steps of optimization to update $\pi_{(s)}^{t-1}$ for all $s \in [K]$ (for example, gradient ascent or Adam) to maximize

$$\sum_{f(p_k)=s}\sum_{j\in[N_p]} \log\sigma\left(\beta\log\frac{\pi(a_{p_k,0}^{(j)} \mid s_{p_k}^{(j)})}{\pi^{\text{old}}(a_{p_k,0}^{(j)} \mid s_{p_k}^{(j)})} - \beta\log\frac{\pi(a_{p_k,1}^{(j)} \mid s_{p_k}^{(j)})}{\pi^{\text{old}}(a_{p_k,1}^{(j)} \mid s_{p_k}^{(j)})}\right)$$

  and obtain $\pi_{(s)}^t$ for all $s \in [K]$.

end for

Assign $f^{T+1}(i)$ for all $i \in [N]$ as

$$f^{T+1}(i) \leftarrow \arg\max_{s\in[K]} \sum_{j\in[N_p]} \log\sigma\left(\beta\log\frac{\pi_{(s)}^{T}(a_{i,0}^{(j)} \mid s_i^{(j)})}{\pi^{\text{old}}(a_{i,0}^{(j)} \mid s_i^{(j)})} - \beta\log\frac{\pi_{(s)}^{T}(a_{i,1}^{(j)} \mid s_i^{(j)})}{\pi^{\text{old}}(a_{i,1}^{(j)} \mid s_i^{(j)})}\right)$$

Output: $(\pi_{(k)}^T)_{k\in[K]}$ and f^{T+1}

should cover uniformly for arbitrary $\widehat{f}$, and also due to a difference between max and expectation of max, which is bounded using McDiarmid's inequality.

Remark 2. *In contrast to the results in Section 3, we additionally assume $C_r(\mathcal{G}_r, \pi, \mu_1, i) \le C'_{max}$ in*

Theorem 4.1. To adopt a pessimistic approach, constructing a confidence set for clustered reward functions across all clusters is necessary. However, the ambiguity of which human user belongs to which cluster complicates this analysis, as pessimism would need to be applied to every potential cluster. Consequently, defining a confidence set for every possible clustering scenario is required, significantly complicating the analysis of the algorithm.

5 Reward and Preference Aggregation

This section adheres to the RLHF setting with a single LLM, while handling the heterogeneous human feedback by reward/preference aggregation. For reward aggregation, we leverage comparison data similar to the methods discussed in the previous sections. We first estimate individual reward functions based on such data and then aggregate these functions to form a unified reward model. In comparison, for preference aggregation, we introduce a novel framework termed "probabilistic opinion pooling". Specifically, instead of relying on binary comparison data, human users provide feedback as *probability vectors*. This approach eliminates the need to aggregate heterogeneous preferences via reward functions, allowing for the direct aggregation of probabilistic opinions provided by users.

5.1 Reward Aggregation with Comparison Data

We introduce the following reward aggregation rules (Equations (5.1) and (5.2)), which are favorable as they satisfy several pivotal axioms in social choice theory. These axioms – monotonicity, symmetry, continuity, independence of unconcerned agents, translation independence, and the Pigou-Dalton transfer principle – are crucial for ensuring fairness and consistency in the decision-making process (List, 2013; Skiadas, 2009, 2016). We present the definition of these axioms in Appendix E.1 for completeness. The aggregation rules are presented as follows:

$$\mathrm{Agg}_\alpha(\boldsymbol{r}) = \begin{cases} \frac{1}{\alpha}\log\left(\frac{1}{N}\sum_{i\in[N]}\exp(\alpha r_i)\right) & \alpha \neq 0 \\ \frac{1}{N}\sum_{i\in[N]} r_i & \alpha = 0 \end{cases} \tag{5.1}$$

and

$$\mathrm{Agg}'_\alpha(\boldsymbol{r}) = \begin{cases} \frac{1}{N\alpha}\sum_{i\in[N]}(\exp(\alpha r_i)-1) & \alpha \neq 0 \\ \frac{1}{N}\sum_{i\in[N]} r_i & \alpha = 0 \end{cases} \tag{5.2}$$

where $\boldsymbol{r} = (r_1, \ldots, r_N)^\intercal$ is a reward vector with trajectory input. Note that Equation (5.1) and Equation (5.2) are equivalent in the sense of the associated optimal policy, as the log function is monotonically increasing. We can verify that $\lim_{\alpha\to-\infty}\mathrm{Agg}_\alpha(\boldsymbol{r}) = \min_{i\in[N]} r_i$ and $\lim_{\alpha\to\infty}\mathrm{Agg}_\alpha(\boldsymbol{r}) = \max_{i\in[N]} r_i$. This implies that when α is small, the reward aggregation rule emphasizes on $\min_{i\in[N]} r_i$, and when α is large, it emphasizes on $\max_{i\in[N]} r_i$. When $\alpha = 0$, Equation (5.1) represents utilitarianism, and when $\alpha \to -\infty$, Equation (5.1) represents a Leximin-based aggregation rule (List, 2013).

5.1.1 Algorithm and Analysis

Algorithm 5 outputs a joint estimation of $\psi^\star$ and $\boldsymbol{\theta}^\star$ with maximum likelihood estimation as Algorithm 1. The procedure is overall the same as Algorithm 1, except the last step for estimating the best policy for the pessimistic expected value function associated with the aggregated reward function.

Algorithm 5 RLHF with Reward Aggregation

Input: Dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$ is the preference dataset for the ith human, $\lambda > 0$, and $\widehat{\omega}$ from Algorithm 1. We also use Equation (3.2) for constructing a confidence set of reward function $\mathcal{R}'(\widehat{\mathcal{D}})$.
Compute policy with respect to $\mathcal{R}'(\widehat{\mathcal{D}})$ for all $i \in [N]$ by

$$\widehat{\pi} \leftarrow \operatorname*{arg\,max}_{\pi\in\Pi} \min_{\boldsymbol{r}\in\mathcal{R}'(\widehat{\mathcal{D}})} \Big(J(\pi; \mathrm{Agg}_\alpha(r_1,\ldots,r_N)) - \mathbb{E}_{\tau\sim\mu_{\mathrm{ref}}}[\mathrm{Agg}_\alpha(r_1,\ldots,r_N)(\tau)]\Big). \tag{5.3}$$

Output: $(\widehat{\omega}, \widehat{\boldsymbol{\theta}}, \widehat{\pi})$.

Similar to the results in Theorem 3.4, we have the following theorem for Algorithm 5:

Theorem 5.1. (Expected Value Function Gap). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. For any $\delta \in (0,1]$, all $i \in [N]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 5 satisfies*

$$\begin{aligned}&J(\pi_{tar}; Agg_\alpha(\boldsymbol{r}^\star)) - J(\widehat{\pi}; Agg_\alpha(\boldsymbol{r}^\star))\\ &\le \sqrt{c_\alpha C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{tar}, \mu_{ref})^2 \left(\frac{k\kappa^2 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/(\delta/N))}{NN_p} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2 N_p} + \lambda B^2\right)}\end{aligned}$$

where $c_\alpha > 0$ is a constant depending on α, and other constants are defined in Section 3.1.

We defer the proof of Theorem 5.1 to Appendix E.2. Lastly, we consider the tightness of our analysis by providing a theoretical lower bound of the sub-optimality gap of aggregation.

Theorem 5.2. (Lower Bound for the Sub-Optimality Gap of Aggregation). *For any $k > 6, N_p \ge Ck\Lambda^2, \Lambda \ge 2$, and $\alpha \in \mathbb{R}$ there exists a representation function $\phi(\cdot)$ so that*

$$\inf_{\widehat{\pi}} \sup_{Q\in\mathrm{CB}(\Lambda)} \left(\max_{\pi^*\in\Pi} J(\pi^*; Agg_\alpha(\boldsymbol{r}_{\omega,\boldsymbol{\theta}})) - J(\widehat{\pi}; Agg_\alpha(\boldsymbol{r}_{\omega,\boldsymbol{\theta}}))\right) \ge C\Lambda \cdot \sqrt{\frac{k}{N_p}},$$

where

$$\mathrm{CB}(\Lambda) := \left\{Q := \left(\{\mu_0, \mu_1\}, \{\tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\}_{i\in[N], j\in[N_p]}, \omega, \boldsymbol{\theta}\right) \,\middle|\, C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi^\star, \mu_1, i) \le \Lambda \text{ for all } i \in [N]\right\}$$

is the family of MDP with N reward functions and $H = 1$ instances.

By Theorem 5.2, for general representation function class, we establish that Algorithm 5 is near-optimal for the sub-optimality of the induced personalization policy.

Remark 3 (Comparison with Recent Independent Work (Zhong et al., 2024))**.** *(Zhong et al., 2024) also considered reward aggregation rules, adhering to the axiom of scale independence but not translation independence (Moulin, 2004). Here, scale/translation independence of the aggregation rule means that the aggregation rule should yield the same choice even if the reward functions are scaled/translated,* *respectively. Our theoretical results can also be extended to the reward aggregation rule in (Zhong et al., 2024). However, since we also consider the relationship of the reward aggregation rule and preference aggregation rule by probabilistic opinion pooling, which will be presented in the next section, we only present Equation (5.1) for the reward aggregation rule.*

(Zhong et al., 2024) also considered Nash's bargaining (Nash, 1953), which maximizes $\prod_{i\in[N]}(r_i - \min r_i)$ rather than $\prod_{i\in[N]} r_i$. In this case, they can also consider translation independence. Ours can also be extended by substituting r_i to $r_i/\min r_i$ in the aggregation function (Equation (5.1)), but we decided not to include it in our paper since we did not have this result in our initial draft.

5.2 Preference Aggregation with Probabilistic Opinion Data

5.2.1 RLHF with Probabilistic Opinion Feedback

Consider a set of questions $\{s^{(j)}\}_{j\in[N_p]}$, and for each question $s^{(j)}$, there are K potential answers denoted by $\mathcal{A}^{(j)} := \{a_k^{(j)}\}_{k\in[K]}$. Traditional RLHF methods involve human labelers $i \in [N]$ selecting a preferred answer from $\mathcal{A}^{(j)}$. This approach limits the human feedback to a singular choice, which, though being simple, restricts the expressiveness of human preferences.

To address this, we introduce a new setting whereby human labelers provide feedback as a probability vector $q_i^{(j)} \in \Delta(\mathcal{A}^{(j)})$, which is also called *probabilistic opinion* in social choice theory (Stone, 1961; Lehrer and Wagner, 2012). Here, $\Delta(\mathcal{A}^{(j)})$ represents the set of all possible distributions over the answers in $\mathcal{A}^{(j)}$. This allows labelers to quantify their preferences across multiple answers rather than selecting only one, and can be implemented in practice without increasing too much of overload for feedback collection.

Our setup does not assume a predefined relationship between each reward function for every human labeler and their preferences. Instead, we aggregate the diverse probabilistic preferences of multiple labelers into a consensus probability distribution over the answers. We define an aggregation function (or a *probabilistic opinion pooling function*), $\text{Agg-p}_\alpha(\boldsymbol{P})$, which takes a tuple of human preference distributions $\boldsymbol{P} = (P_1, \dots, P_N) \in (\Delta(\mathcal{A}))^N$ and maps it to a single probability distribution in $\Delta(\mathcal{A})$ where $\mathcal{A}$ is the potential answer set. This aggregated distribution represents a probabilistic opinion pooling among human labelers for each $a \in \mathcal{A}$:

$$\text{Agg-p}_\alpha(\mathbf{P})(a) = \begin{cases} \dfrac{\left(\sum_{i\in[N]}(P_i(a))^\alpha\right)^{1/\alpha}}{\sum_{a'\in\mathcal{A}}\left(\sum_{i\in[N]}(P_i(a'))^\alpha\right)^{1/\alpha}} & \alpha \neq 0 \\ \dfrac{\left(\prod_{i\in[N]} P_i(a)\right)^{1/N}}{\sum_{a'\in\mathcal{A}}\left(\prod_{i\in[N]} P_i(a')\right)^{1/N}} & \alpha = 0 \end{cases}. \tag{5.4}$$

Remark 4. *The case where $\alpha = 0$ is referred to as the geometric pooling function (McConway, 1978). This function is known for preserving unanimity and not being eventwise independent, while it does satisfy external Bayesianity (Madansky, 1964; Dietrich and List, 2016). External Bayesianity mandates that updating the probabilities with new information should yield consistent results regardless of whether the update occurs before or after the aggregation process (Genest, 1984).*

Interestingly, Equation (5.4), which describes the aggregation of probabilistic preferences, has a connection to Equation (5.2), concerning reward aggregation, under the assumption of the Plackett-Luce model for the relationship between reward functions and preference models (Definition 2.1). We then formalize the connection between the probabilistic opinion pooling in Equation (5.4) and the reward aggregation rule in Equation (5.1). We defer the proof of Theorem 5.3 to Appendix E.4.

Theorem 5.3. (Relationship between Reward Aggregation and Preference Aggregation). *Suppose human preferences are modeled by the PL model, and all human labelers share a common lower bound on their reward functions. Let* $(R_i(a))_{a\in\mathcal{A}}$ *represent the reward function associated with action* $a \in \mathcal{A}$ *and* $P_i \in \Delta(\mathcal{A})$ *denote the corresponding probabilistic opinion for individual* $i \in [N]$*. Then, the preference aggregation* $\textit{Agg-p}_\alpha(\mathbf{P})$*, is equivalent to the preference derived under the PL model with the aggregated rewards* $(\textit{Agg}_\alpha(\mathbf{R}(a)))_{a\in\mathcal{A}}$ *for any* $\alpha \in [-\infty, \infty]$.

While we generally do not presuppose any specific relationship between probabilistic opinions and reward functions, Theorem 5.3 shows that under the classical choice model of Plackett-Luce, these two aggregation rules can coincide (while the probabilistic aggregation framework may potentially handle other cases).

5.2.2 Algorithm

Now, we provide an algorithm that uses the feedback in the form of probabilistic opinions (Algorithm 6). The only difference from the DPO algorithm (Rafailov et al., 2024) is to change the deterministic answer a_i to the a_i sampled based on the probabilistic opinion pooling, which is in the second line in the for loop of Algorithm 6.

Algorithm 6 Probabilistic Opinion Pooling DPO (POP-DPO)

Input: Dataset $\widehat{\mathcal{D}} = \cup_{i\in[N]}\widehat{\mathcal{D}}_i$ where $\widehat{\mathcal{D}}_i = \{q_i^{(j)}(s_i^{(j)}), s^{(j)}, i)\}_{j\in[N_p]}$ is the probabilistic opinion dataset for the ith individual, $q_i^{(j)} \in \Delta(\mathcal{A})$ with $|\mathcal{A}| = 2$, β is a parameter for DPO, α is a parameter for aggregation

for every epoch **do**

For every question $s^{(j)}$ where j is in the batch, $q^{(j)} := \text{Agg-p}_\alpha(\boldsymbol{q}^{(j)})$.

Sample $a_0^{(j)} \sim \text{Multinomial}(q^{(j)})$ and define $a_1^{(j)}$ as non-selected answer.

Run a few steps of optimization to update π (for example, gradient ascent or Adam) to maximize

$$\sum_{j\in\text{batch}} \log\sigma\left(\beta\log\frac{\pi(a_0^{(j)} \mid s^{(j)})}{\pi^{\text{old}}(a_0^{(j)} \mid s^{(j)})} - \beta\log\frac{\pi(a_1^{(j)} \mid s^{(j)})}{\pi^{\text{old}}(a_1^{(j)} \mid s^{(j)})}\right)$$

end for

Output: π

5.3 Mechanism Design for Preference Aggregation

Suppose that human labeler i ($i \in [N]$) provides preference data by probabilistic opinion $P_i \in \Delta(\mathcal{A})$. We now consider the natural scenario where the labelers may be *strategic* – given they are human beings with (certain degree of) rationality. In particular, knowing the form of preference aggregation (and the fact that they may affect the process), human labelers may provide *untruthful* feedback of their preference, in order to benefit more in terms of their *actual* utility/preference. In particular, the untruthful preference may *bias* the aggregated preference (that LLM will be fine-tuned over) towards their own preference, and thus manipulates the LLM output. We demonstrate the scenario more quantitatively in the following example.

An Example with Untruthful Feedback. Consider a set of N labelers evaluating two answers, where each labeler expresses a probabilistic opinion on the answers (a_1, a_2). Specifically, suppose labeler N believes that a_1 is slightly preferable to a_2, represented by the probability vector $P_N = (0.6, 0.4)^\intercal$. Conversely, all other labelers $i \in [N-1]$ have probabilistic opinion favoring the second answer, represented by $P_i = (0.2, 0.8)^\intercal$.

We assume that the aggregation of these opinions employs the Agg-p$_{-\infty}$ rule, defined as $\text{Agg-p}_{-\infty}(\boldsymbol{P})(a_t) = \frac{\min_{i\in[N]} P_i(a_t)}{\min_{i\in[N]} P_i(a_1) + \min_{i\in[N]} P_i(a_2)}$ for $t = 1, 2$, where $\boldsymbol{P}$ represents the matrix of probabilistic opinions across all labelers and answers. Under truthful reporting, the aggregated result would be calculated as $\text{Agg-p}_{-\infty}(\boldsymbol{P}) = (1/3, 2/3)^\intercal$. However, labeler N can strategically provide an untruthful probabilistic opinion to distort the aggregated result toward his original view: If labeler N reports a distorted opinion of $P'_N = (13/15, 2/15)^\intercal$ instead of $(0.6, 0.4)^\intercal$, the new aggregated opinion becomes $\text{Agg-p}_{-\infty}(\boldsymbol{P}') = (0.6, 0.4)^\intercal$, where $\boldsymbol{P}' = (P_1, \ldots, P_{N-1}, P'_N)$, which aligns exactly with labeler N's probabilistic opinion, while further deviating from other labelers' actual preference. This example underscores the potential of strategic behavior in the aggregation of probabilistic opinions, and thus highlights the importance of incentivizing truthful preference reporting.

5.3.1 Setup

To address the untruthful feedback issue, we resort to the ideas in mechanism design (Nisan and Ronen, 1999; Börgers, 2015; Roughgarden, 2010). Specifically, we will develop mechanisms that can impose some *cost* on human labelers, so that they do not have the incentive to report untruthful preferences.

Imposing Cost for Human Feedback Collection. Though not being enforced in most existing RLHF frameworks, we believe it is reasonable and possible to incorporate it in the feedback collection, especially in scenarios where a single reward model (and thus a single LLM) is mandated. For example, the future large models may be regulated by some administrative agency, e.g., the government. These agencies' objective is for social good, despite the heterogeneity in human preferences, and also possess the power to enforce cost to human labelers, e.g., via taxing. It may also be possible for big technology companies who train LLMs, e.g., OpenAI, to incentivize truthful feedback through personalized and strategic (negative) payment (which corresponds to the cost here) to human labelers.

In this setup, we will first prove the existence of a cost function $c_i : \Delta(\mathcal{A})^N \to \mathbb{R}$ for all $i \in [N]$ that induces truthful reporting of probabilistic opinions from human labelers. Here, the input of c_i is the probabilistic opinion of every human labeler. This is also called the dominant strategy incentive-compatible (DSIC) mechanism (Nisan and Ronen, 1999; Börgers, 2015; Roughgarden, 2010). Then, we prove that there exists an aggregation rule and cost function that induce DSIC, and also maximize social welfare. We denote each human labeler's underlying (true) probabilistic opinion as $p_i\left(s^{(j)}\right)$ for each question $s^{(j)}$. Accounting for such cost, we define the *utility function* of individual i for question $s^{(j)}$ as

$$u_i^{(j)}\left(p_i\left(s^{(j)}\right), \left(P_i\left(s^{(j)}\right)\right)_{i\in[N]}\right) = -d\left(p_i\left(s^{(j)}\right), \text{Agg-p}\left(\left(P_i\left(s^{(j)}\right)\right)_{i\in[N]}\right)\right) - c_i\left(\left(P_i\left(s^{(j)}\right)\right)_{i\in[N]}\right).$$

Here, $d : \Delta(\mathcal{A}) \times \Delta(\mathcal{A}) \to \mathbb{R}$ represents the distance between the underlying true probabilistic opinion and the aggregated preference. Moreover, we define the *welfare function* of individual i from addressing question $s^{(j)}$ as $\mathrm{Wel}_i^{(j)}(O) = -d(p_i(s^{(j)}), O)$ for any $O \in \Delta(\mathcal{A})$.

Remark 5 (Examples of Distance Function d)**.** *We can instantiate $d(p,q)$ as the KL-divergence. Also, we may instantiate $d_\alpha(p,q) = sgn(\alpha)\frac{1}{1-\alpha}\sum_{j\in\mathcal{A}}\left(1 - p_j^\alpha q_j^{1-\alpha}\right)$, which is a variant of the α-Renyi divergence for $\alpha \neq 0$. One can easily check that $d_\alpha(p,q) \geq 0$. In fact, one can also prove that $\lim_{\alpha\to 1} d_\alpha(p,q) = d(p,q)$ with $d(p,q)$ being the KL-divergence (Appendix E.5).*

5.3.2 Mechanism and Guarantees

We design a mechanism inspired by the Vickery-Clarke-Groves mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1973), as defined below.

Definition 5.1 (VCG Mechanism)**.** *Assume that there are n strategic agents and a finite set X of outcome, and each individual i has a private valuation v_i for each outcome $x \in X$. The bidding $\boldsymbol{b} = (b_1, \ldots, b_N)^\intercal \in (\mathbb{R}^{|X|})^N$ where $b_i \in \mathbb{R}^{|X|}$ is bidding for all outcome of individual $i \in [N]$. Define their utility function as $v_i(\boldsymbol{x}(\boldsymbol{b})) - c_i(\boldsymbol{b})$, where $\boldsymbol{x} : (\mathbb{R}^{|X|})^N \to X$ is the allocation rule and $c_i : (\mathbb{R}^{|X|})^N \to \mathbb{R}$ is the cost function. The summation of welfare function of all agents is defined as $Wel(x) = \sum_{i\in[N]} v_i(x)$ for all $x \in X$. The goal is to design $\boldsymbol{x}$ and $(c_i)_{i\in[N]}$ functions to make a DSIC and welfare-maximizing mechanism. The following $\boldsymbol{x}$ and c_i for $i \in [N]$ is DSIC welfare maximizing mechanism:*

$$\boldsymbol{x}(\boldsymbol{b}) = \arg\max_{x\in X} \sum_{i\in[N]} b_i(x), \qquad c_i(\boldsymbol{b}) = \max_{x\in X} \sum_{j\neq i} b_j(x) - \sum_{j\neq i} b_j(\boldsymbol{x}(\boldsymbol{b})) \textit{ for all } i \in [N].$$

Unfortunately, the classical VCG mechanism presents certain limitations such as it cannot be solved in polynomial time in general (Nisan and Ronen, 1999; Börgers, 2015; Roughgarden, 2010). We here adopt certain forms of allocation rule (which corresponds to the aggregation rule in our RLHF setting) and cost functions as follows, which allow the outcome set to be a simplex (with infinitely many outcomes):

$$\text{Agg-p}(\boldsymbol{P}) = \arg\min_{p\in\Delta(\mathcal{A})} \sum_{i\in[N]} d(\boldsymbol{P}, p), \qquad c_i(\boldsymbol{P}) = \sum_{j\neq i} d(P_i, \text{Agg-p}(\boldsymbol{P})) - \min_{p\in\Delta(\mathcal{A})} \sum_{j\neq i} d(P_i, p). \tag{5.5}$$

Theorem 5.4. (DSIC Welfare-Maximizing Mechanism). *The aggregation rule and the cost function as in Equation* (5.5) *provide a DSIC welfare-maximizing mechanism.*

Due to the modeling, we have an advantage compared to the original VCG mechanism. The minimization in the aggregation function can be achieved using a simple optimization method such as gradient descent, which makes our aggregation rule and cost function computation easy, which is in contrast with the original VCG mechanism.

Now, we connect our mechanism design with pre-defined preference aggregation function (Agg-p$_\alpha$ in Equation (5.4)). Theorem 5.5 implies that Equation (5.4) is maximizing social welfare and also we are available to construct the cost function to make human feedback truthful.

Theorem 5.5. *If we set d as a variant of the α-Renyi distance for $\alpha \neq 0$ (Remark 5) and define d as KL-divergence for $\alpha = 0$, the DSIC welfare-maximizing aggregation rule is Equation* (5.4)*. Therefore, aggregation rule Equation* (5.4) *is also welfare-maximizing with appropriate cost function.*

If we assume the relationship between reward and preference follows the PL model (Definition 2.1), then Equation (5.1) implies a welfare-maximizing aggregation rule, which connects reward aggregation and mechanism design. We defer all proofs for the results in Section 5.3.2 to Appendix E.6.

Acknowledgement

The authors would like to thank Dingwen Kong for the discussions, and thank Jisu Jang for providing helpful feedback for Figure 1.

References

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

Agarwal, A., Kakade, S., Krishnamurthy, A. and Sun, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, **33** 20095–20107.

Anandalingam, G. and Friesz, T. L. (1992). Hierarchical optimization: An introduction. *Annals of Operations Research*, **34** 1–11.

Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica: Journal of the Econometric Society* 404–437.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T. et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bartlett, P. L., Foster, D. J. and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. *NeurIPS*.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, **12** 149–198.

Börgers, T. (2015). *An introduction to the theory of mechanism design*. Oxford University Press, USA.

Bose, A., Du, S. S. and Fazel, M. (2024). Offline multi-task transfer rl with representational penalization. *arXiv preprint arXiv:2402.12570*.

Boxall, P. C. and Adamowicz, W. L. (2002). Understanding heterogeneous preferences in random utility models: a latent class approach. *Environmental and resource economics*, **23** 421–446.

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

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

Clarke, E. H. (1971). Multipart pricing of public goods. *Public choice* 17–33.

Dietrich, F. and List, C. (2016). Probabilistic opinion pooling.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*. PMLR.

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

Finn, C., Levine, S. and Abbeel, P. (2016). Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*. PMLR.

Genest, C. (1984). A characterization theorem for externally bayesian groups. *The Annals of Statistics* 1100–1105.

Ghosh, A., Chung, J., Yin, D. and Ramchandran, K. (2020). An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, **33** 19586–19597.

Groves, T. (1973). Incentives in teams. *Econometrica: Journal of the Econometric Society* 617–631.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, **316** 2402–2410.

Hsu, D., Kakade, S. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors.

Hwang, M., Weihs, L., Park, C., Lee, K., Kembhavi, A. and Ehsani, K. (2023). Promptable behaviors: Personalizing multi-objective rewards from human preferences. *arXiv preprint arXiv:2312.09337*.

Ishfaq, H., Nguyen-Tang, T., Feng, S., Arora, R., Wang, M., Yin, M. and Precup, D. (2024). Offline multitask representation learning for reinforcement learning. *arXiv preprint arXiv:2403.11574*.

Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y. and Ammanabrolu, P. (2023). Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Lehrer, K. and Wagner, C. (2012). *Rational consensus in science and society: A philosophical and mathematical study*, vol. 24. Springer Science & Business Media.

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

List, C. (2013). Social choice theory.

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

Liu, Q., Netrapalli, P., Szepesvari, C. and Jin, C. (2023). Optimistic mle: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*.

Lu, R., Huang, G. and Du, S. S. (2021). On the power of multitask representation learning in linear mdp. *arXiv preprint arXiv:2106.08053*.

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Madansky, A. (1964). *Externally bayesian groups*. Rand Corporation.

Mansour, Y., Mohri, M., Ro, J. and Suresh, A. T. (2020). Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*.

McConway, K. J. (1978). *The combination of experts' opinions in probability assessment: some theoretical considerations*. Ph.D. thesis, University College London (University of London).

Mohri, M. and Muñoz Medina, A. (2012). New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*. Springer.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, **13** 47–60.

Moulin, H. (2004). *Fair division and collective welfare*. MIT press.

Nash, J. (1953). Two-person cooperative games. *Econometrica: Journal of the Econometric Society* 128–140.

Nisan, N. and Ronen, A. (1999). Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, **35** 27730–27744.

Ozdaglar, A. E., Pattathil, S., Zhang, J. and Zhang, K. (2023). Revisiting the linear-programming framework for offline rl with general function approximation. In *International Conference on Machine Learning*. PMLR.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **24** 193–202.

Pollak, R. A. and Wales, T. J. (1992). *Demand system specification and estimation*. Oxford University Press, USA.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S. and Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, **36**.

RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, **34** 11702–11716.

REN, T., ZHANG, T., SZEPESVÁRI, C. and DAI, B. (2022). A free lunch from the noise: Provable and practical exploration for representation learning. In *Uncertainty in Artificial Intelligence*. PMLR.

ROUGHGARDEN, T. (2010). Algorithmic game theory. *Communications of the ACM*, **53** 78–86.

SATTLER, F., MÜLLER, K.-R. and SAMEK, W. (2020). Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, **32** 3710–3722.

SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

SEN, A. (2018). *Collective Choice and Social Welfare*. Harvard University Press.

SKIADAS, C. (2009). *Asset pricing theory*. Princeton University Press.

SKIADAS, C. (2016). Scale or translation invariant additive preferences. *Unpublished manuscript*.

STONE, M. (1961). The opinion pool. *The Annals of Mathematical Statistics* 1339–1342.

TANG, D., QIN, B., FENG, X. and LIU, T. (2015). Effective lstms for target-dependent sentiment classification. *arXiv preprint arXiv:1512.01100*.

TRIPURANENI, N., JIN, C. and JORDAN, M. (2021). Provable meta-learning of linear representations. In *International Conference on Machine Learning*. PMLR.

TRIPURANENI, N., JORDAN, M. and JIN, C. (2020). On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, **33** 7852–7862.

UEHARA, M., ZHANG, X. and SUN, W. (2021). Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*.

VICKREY, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, **16** 8–37.

WANG, Y., LIU, Q. and JIN, C. (2024). Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, **36**.

WARNELL, G., WAYTOWICH, N., LAWHERN, V. and STONE, P. (2018). Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32.

WU, Z., HU, Y., SHI, W., DZIRI, N., SUHR, A., AMMANABROLU, P., SMITH, N. A., OSTENDORF, M. and HAJISHIRZI, H. (2024). Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, **36**.

XIONG, W., DONG, H., YE, C., ZHONG, H., JIANG, N. and ZHANG, T. (2023). Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*.

Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, **102** 315–323.

Zanette, A., Wainwright, M. J. and Brunskill, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, **34** 13626–13640.

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

Zhan, W., Uehara, M., Sun, W. and Lee, J. D. (2022). Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*.

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

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

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Supplementary Materials for

"Principled RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation"

Contents

1 Introduction **1**
1.1 Related Work 4

2 Preliminaries **6**

3 Personalized RLHF via Representation Learning **7**
3.1 Algorithms 8
3.2 Results and Analyses 9

4 Personalized RLHF via Human User Clustering **13**
4.1 Algorithms 14
4.2 Results and Analyses 14

5 Reward and Preference Aggregation **17**
5.1 Reward Aggregation with Comparison Data 17
5.1.1 Algorithm and Analysis 17
5.2 Preference Aggregation with Probabilistic Opinion Data 19
5.2.1 RLHF with Probabilistic Opinion Feedback 19
5.2.2 Algorithm 20
5.3 Mechanism Design for Preference Aggregation 20
5.3.1 Setup 21
5.3.2 Mechanism and Guarantees 22

A Table of Notation **30**

B Deferred Definition **30**
B.1 Bracketing Number. 30
B.2 Concentrability Coefficient for a Reward Scalar Class 31

C Deferred Proofs in Section 3 **31**
C.1 Proof of Theorem 3.1 and Corollary 3.2 31
C.2 Discussion on Assumption 3.3 34
C.2.1 Comparing with (Lu et al., 2021, Assumption 6.4) 34
C.2.2 Case Study (Linear Representation): $\psi_\omega(x) = \omega x$ and ω is an Orthonormal Matrix 35
C.3 Proof of Corollary 3.3 35
C.4 Proof of Theorem 3.4 36
C.5 Proof of Theorem 3.6 40

D Proof of Section 4 **41**

E Proof of Section 5 **43**

E.1 Six Pivotal Axioms for Reward Aggregation 43
E.2 Proof of Theorem 5.1 44
E.3 Proof of Theorem 5.2 45
E.4 Proof of Theorem 5.3 45
E.5 Relationship between KL divergence and variant of α-Renyi divergence. 46
E.6 Proof of Section 5.3.2 46

A Table of Notation

Notation	Definition
N	Number of Individuals
$\mathcal{S}$	State Space
$\mathcal{A}$	Action Set
H	Horizon Length
P_h	Transition Probability at Horizon h
$\boldsymbol{r}$	Reward
$\mathcal{T}$	Trajectory Set
τ	Trajectory
$J(\pi; r_i)$	$\mathbb{E}_{\tau,\pi}[r_i(\tau)]$
$d_\pi(\tau)$	Occupancy Measure: $\mathbb{P}_\pi(\tau)$
$\Phi: \mathbb{R} \to [0,1]$	Strongly Convex Function Mapping Reward to Preference
$\sigma(x)$	Sigmoid Function: $\frac{e^x}{1+e^x}$
$P_{r_i}(o = 0 \mid \tau_0, \tau_1)$	$\Phi(r_i(\tau_0) - r_i(\tau_1))$
$\psi_\omega : \mathbb{R}^d \to \mathbb{R}^k$	Representation Function
Ψ	$\{\psi_\omega \mid \omega \in \Omega\}$
$\mathcal{G}_{\boldsymbol{r}}$	Set of Reward Functions: $\{(\langle \psi_\omega(\phi(\cdot)), \theta_i \rangle)_{i\in[N]} \mid \psi_\omega \in \Psi, \theta_i \in \mathbb{R}^k$ and $\|\theta_i\|_2 \le B$ for all $i \in [N]\}$
$\mathcal{N}_{\mathcal{G}_r}(\epsilon)$	Bracket Number of $\mathcal{G}_{\boldsymbol{r}}$ Associated with ϵ
$r_{\omega,\theta_j}(\cdot)$	$\langle \psi_\omega(\phi(\cdot)), \theta_j \rangle$
$\boldsymbol{r}_{\omega,\boldsymbol{\theta}}(\cdot)$	$(r_{\omega,\theta_1}(\cdot), \cdots, r_{\omega,\theta_N}(\cdot)) \in \mathbb{R}^N$
$r_i^\star(\cdot)$	Ground-truth Reward: $\langle \psi^\star(\phi(\cdot)), \theta_i^\star \rangle$
$\psi^\star (= \psi_{\omega^\star})$	Ground-truth Representation Function
$R_{\max}$	$-R_{\max} \le r_i^\star(\tau) \le R_{\max}$
$\widehat{\mathcal{D}}$	$\cup_{i\in[N]} \widehat{\mathcal{D}}_i$
$\widehat{\mathcal{D}}_i$	$\{(o_i^{(j)}, \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})_{j\in[N_p]}\}$
N_p	$N_p = \lvert\widehat{\mathcal{D}}_1\rvert = \lvert\widehat{\mathcal{D}}_2\rvert = \ldots = \lvert\widehat{\mathcal{D}}_N\rvert$
$C_r(\mathcal{G}_{\boldsymbol{r}}, \pi_{\text{tar}}, \mu_{\text{ref}}, i)$	Defined in Definition 3.1
$\text{Agg}_\alpha(\boldsymbol{r})$	Defined in Equation (5.2)
$\text{Agg-p}_\alpha(\boldsymbol{p})(a)$	Defined in Equation (5.4)

B Deferred Definition

B.1 Bracketing Number.

We modify and adopt the definition of the bracketing number of preferences introduced by (Zhan et al., 2023), with some adjustments. Consider $\mathcal{G}_{\boldsymbol{r}}$ as the class of functions representing sets of reward vectors, where each reward vector is denoted by $(r_i)_{i\in[N]}$. Assume g_1 and g_2 maps $(\tau_0, \tau_1) \in \mathcal{T} \times \mathcal{T}$ to $2N$-dimensional vectors. A pair (g_1, g_2) constitutes an ϵ-bracket if for every pair of trajectories (τ_0, τ_1) and for each $i \in [N]$, it holds that $g_1(\cdot \mid \tau_0, \tau_1) \le g_2(\cdot \mid \tau_0, \tau_1)$ and $\|g_1(\cdot \mid \tau_0, \tau_1) -$

$g_2(\cdot \mid \tau_0, \tau_1)\|_1 \le \epsilon$. The ϵ-bracketing number of $\mathcal{G}_{\boldsymbol{r}}$, denoted by $\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(\epsilon)$, is defined as the minimum number of ϵ-brackets $(g_{b,1}, g_{b,2})_{b\in[\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}(\epsilon)}]}$ required such that for any reward vector $\boldsymbol{r} \in \mathcal{G}_{\boldsymbol{r}}$, there exists at least one bracket $b \in [\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}(\epsilon)}]$ such that for all pairs of trajectories (τ_0, τ_1), $g_{b,1}(\cdot \mid \tau_0, \tau_1) \le P_{\boldsymbol{r}}(\cdot \mid \tau_0, \tau_1) \le g_{b,2}(\cdot \mid \tau_0, \tau_1)$ holds.

B.2 Concentrability Coefficient for a Reward Scalar Class

This definition is exactly the same with the concentrability coefficient of preference as outlined by (Zhan et al., 2023).

Definition B.1 (Zhan et al. (2023))**.** *The concentrability coefficient, with a reward vector class $\mathcal{G}_r$, a target policy π_{tar} (which policy to compete with (potentially optimal policy $\pi^\star$)), and a reference policy μ_{ref}, is defined as follows:*

$$C_r\left(\mathcal{G}_r, \pi_{tar}, \mu_{ref}\right) := \max\left\{0, \sup_{r\in\mathcal{G}_r} \frac{\mathbb{E}_{\tau_0\sim\pi_{tar}, \tau_1\sim\mu_{ref}}\left[r^\star(\tau_0) - r^\star(\tau_1) - r(\tau_0) + r(\tau_1)\right]}{\sqrt{\mathbb{E}_{\tau_0\sim\mu_0, \tau_1\sim\mu_1}\left[\left|r^\star(\tau_0) - r^\star(\tau_1) - r(\tau_0) + r(\tau_1)\right|^2\right]}}\right\}.$$

C Deferred Proofs in Section 3

C.1 Proof of Theorem 3.1 and Corollary 3.2

Theorem 3.1. (Total Expected Value Function Gap). *Suppose Assumption 3.1 holds. For any $\delta \in (0,1]$, with probability at least $1-\delta$, the output $(\widehat{\pi}_i)_{i\in[N]}$ of Algorithm 1 satisfies*

$$\sum_{i\in[N]}\left(J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)\right) \le \sqrt{\frac{c\kappa^2 N C_{\max}^2 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/NN_p)/\delta)}{N_p}},$$

where $C_{\max} := \max_{i\in[N]} C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)$ and $c > 0$ is a constant.

Corollary 3.2. (Expected Value Function Gap). *Suppose Assumption 3.1 holds. For any $\delta \in (0,1]$ and all $i \in [N]$, with probability at least $1-\delta$, the output $\widehat{\pi}_i$ of Algorithm 1 satisfies*

$$J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \le \sqrt{\frac{c\kappa^2 C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)^2 \log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/NN_p)/\delta)}{N_p}},$$

where $c > 0$ is a constant.

Before having a proof of Theorem 3.1 and Corollary 3.2, we provide general two properties of MLE estimates, which is a slightly modified version of (Zhan et al., 2023) and (Liu et al., 2022).

Lemma 2 ((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 3 ((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 $o^{(j)} \sim P_{r^\star}(\cdot|\tau_0^{(j)}, \tau_1^{(j)})$,*

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

holds where $C_2 > 0$ is a constant.

Lemma 4 ((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 $o_i^{(j)} \sim P_{r_i^\star}(\cdot|\tau_0^{(j)}, \tau_1^{(j)})$,*

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

holds where $C_2 > 0$ is a constant.

Note that $\boldsymbol{r}^\star$ do not need to be in $\mathcal{G}_{\boldsymbol{r}}'$ for the above lemmas. Lemma 2 states that the log-likelihood $\log P_{\boldsymbol{r}}$ for a preference dataset generated by the reward model $\boldsymbol{r}^\star$ cannot exceed the log-likelihood $\log P_{\boldsymbol{r}^\star}$ for a preference dataset generated by the reward model $\boldsymbol{r}^\star$, with a gap related to the bracket number of $\mathcal{G}_{\boldsymbol{r}}$. Lemma 4 states that the ℓ_1 distance between likelihood function $P_{\boldsymbol{r}^\star}$ and $P_{\boldsymbol{r}}$ for all $\boldsymbol{r} \in \mathcal{G}_{\boldsymbol{r}}'$ can be bounded with the difference between log-likelihood $\log P_{\boldsymbol{r}^\star}$ and $\log P_{\boldsymbol{r}}$ for a preference dataset generated by the reward model $\boldsymbol{r}^\star$ with a gap related to the bracket number of $\mathcal{G}_{\boldsymbol{r}}'$.

Proof of Theorem 3.1 and Corollary 3.2. We define the event $\mathcal{E}_1, \mathcal{E}_2$ as satisfying (Lemma 2, Lemma 4) 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 $\mathcal{E}_1 \cap \mathcal{E}_2$. Then, we can guarantee that

$$\begin{aligned}\sum_{i\in[N]}\sum_{j\in[N_p]} &\log P_{\widehat{\omega},\widehat{\theta}_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) \\ &\le \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\omega^\star,\theta_i^\star}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) + C_1 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta),\end{aligned}$$

which indicates that $\boldsymbol{r}^\star (= \boldsymbol{r}_{\omega^\star,\boldsymbol{\theta}^\star}) \in \mathcal{R}(\widehat{\mathcal{D}})$. Moreover, by the definition of Equation (3.1), if $\boldsymbol{r}_{\omega,\boldsymbol{\theta}}, \boldsymbol{r}_{\omega',\boldsymbol{\theta}'} \in \mathcal{R}(\widehat{\mathcal{D}})$,

$$\begin{aligned}\Big|\sum_{i\in[N]}\sum_{j\in[N_p]} &\log P_{\omega,\theta_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}) - \sum_{i\in[N]}\sum_{j\in[N_p]} \log P_{\omega',\theta_i'}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)})\Big| \\ &\le C_1 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)\end{aligned}$$

holds, since $\sum_{i\in[N]}\sum_{j\in[N_p]}\log P_{\omega,\theta_i}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})$ is bounded by $\sum_{i\in[N]}\sum_{j\in[N_p]}\log P_{\widehat{\omega},\widehat{\theta}_i}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})$ by definition of $\widehat{\omega},\widehat{\boldsymbol{\theta}}$ if $r_{\omega,\boldsymbol{\theta}}\in\mathcal{G}_{\boldsymbol{r}}$. Therefore, by Lemma 4, we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i\in[N]}\mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\omega,\theta_i}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})-P_{\omega^\star,\theta_i^\star}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})\|_1^2\right]\\
&\leq\frac{C_2}{NN_p}\left(\sum_{i\in[N]}\sum_{j\in[N_p]}\log\left(\frac{P_{\omega^\star,\theta_i^\star}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}{P_{\omega,\theta_i}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}\right)+\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)\right)\\
&\leq\frac{C_2}{NN_p}\Big(C_1\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)+\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)\Big)\\
&=\frac{C_3}{NN_p}\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta)
\end{aligned}
$$

for any $\boldsymbol{r}_{\omega,\boldsymbol{\theta}}\in\mathcal{R}(\widehat{\mathcal{D}})$, where $C_3=C_2(C_1+1)$. Then, by the mean value theorem, for any $\boldsymbol{r}_{\omega,\boldsymbol{\theta}}\in\mathcal{R}(\widehat{\mathcal{D}})$, we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i\in[N]}\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_{\omega,\theta_i}(\tau_{i,0})-r_{\omega,\theta_i}(\tau_{i,1}))-(r_i^\star(\tau_{i,0})-r_i^\star(\tau_{i,1}))\right|^2\right]\\
&\leq\frac{\kappa^2}{N}\sum_{i\in[N]}\mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\omega,\boldsymbol{\theta}}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)},i)-P_{\omega^\star,\boldsymbol{\theta}^\star}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)},i)\|_1^2\right] \qquad (C.1)\\
&\leq\frac{C_3\kappa^2}{NN_p}\log(\mathcal{N}_{\mathcal{G}_{\boldsymbol{r}}}(1/(NN_p))/\delta).
\end{aligned}
$$

Now, we define for all policy π,

$$
r_\pi^{i,\inf}:=\operatorname*{arg\,min}_{\boldsymbol{r}\in\mathcal{R}(\mathcal{D})}\Big(J(\pi,r_i)-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i(\tau)]\Big).
$$

Then, we can bound the difference of the expected cumulative reward of a policy $\pi_{i,\mathrm{tar}}$ and $\widehat{\pi}_i$ by

$$
\begin{aligned}
&J(\pi_{i,\mathrm{tar}};r_i^\star)-J(\widehat{\pi}_i;r_i^\star)\\
&=(J(\pi_{i,\mathrm{tar}};r_i^\star)-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i^\star(\tau)])-(J(\widehat{\pi}_i;r_i^\star)-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i^\star(\tau)])\\
&\underset{(i)}{\leq}(J(\pi_{i,\mathrm{tar}};r_i^\star)-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i^\star(\tau)])\\
&\quad-(J(\pi_{i,\mathrm{tar}};r_{\pi_{i,\mathrm{tar}}}^{i,\inf})-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_{\pi_{i,\mathrm{tar}}}^{i,\inf}(\tau)])+(J(\widehat{\pi}_j;r_{\widehat{\pi}_i}^{i,\inf})-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(r_{\widehat{\pi}_i}^{i,\inf}(\tau)))\\
&\quad-(J(\widehat{\pi}_i;r_i^\star)-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i^\star(\tau)]) \qquad (C.2)\\
&\underset{(ii)}{\leq}(J(\pi_{i,\mathrm{tar}};r_i^\star)-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_i^\star(\tau)])-(J(\pi_{i,\mathrm{tar}};r_{\pi_{i,\mathrm{tar}}}^{i,\inf})-\mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}[r_{\pi_{i,\mathrm{tar}}}^{i,\inf}(\tau)])\\
&=\mathbb{E}_{\tau_{i,0}\sim\pi_{i,\mathrm{tar}},\tau_{i,1}\sim\mu_{i,\mathrm{ref}}}[(r_i^\star(\tau_{i,1})-r_i^\star(\tau_{i,0}))-(r_{\pi_{i,\mathrm{tar}}}^{i,\inf}(\tau_{i,1})-r_{\pi_{i,\mathrm{tar}}}^{i,\inf}(\tau_{i,0}))]\\
&\leq C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}},\pi_{i,\mathrm{tar}},\mu_{i,\mathrm{ref}},i)\sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1})-r_i^\star(\tau_{i,0}))-(r_{\pi_{i,\mathrm{tar}}}^{i,\inf}(\tau_{i,1})-r_{\pi_{i,\mathrm{tar}}}^{i,\inf}(\tau_{i,0}))\right|^2\right]}
\end{aligned}
$$

Here, (i) holds since $\widehat{\pi}_j$ is a distributional robust policy for $\mathcal{R}(\widehat{\mathcal{D}})$ (Equation (3.1)) and (ii) holds due to the definition of $r^{i,\inf}_{\widehat{\pi}_i}$. Therefore, if we sum Equation (C.2) over $i \in [N]$, we have

$$
\begin{aligned}
&\sum_{i\in[N]} \Big(J(\pi_{i,\mathrm{tar}}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)\Big) \\
&\le C_{\max} \sum_{i\in[N]} \sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,1}) - r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,0}))\right|^2\right]} \\
&\le C_{\max} \sqrt{N \sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,1}) - r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,0}))\right|^2\right]} \\
&\le C_{\max} \sqrt{\frac{C_3 N \kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/NN_p)/\delta)}{N_p}},
\end{aligned}
$$

which proves Theorem 3.1. Moreover, we have

$$
\begin{aligned}
&J(\pi_{i,\mathrm{tar}}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star) \\
&\le C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,\mathrm{tar}}, \mu_{i,\mathrm{ref}}, i) \sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,1}) - r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,0}))\right|^2\right]} \\
&\le C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,\mathrm{tar}}, \mu_{i,\mathrm{ref}}, i) \sqrt{\sum_{i\in[N]} \mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,1}) - r^{i,\inf}_{\pi_{i,\mathrm{tar}}}(\tau_{i,0}))\right|^2\right]} \\
&\le C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,\mathrm{tar}}, \mu_{i,\mathrm{ref}}, i) \sqrt{\frac{C_3 \kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/NN_p)/\delta)}{N_p}}
\end{aligned}
$$

which proves Corollary 3.2. □

C.2 Discussion on Assumption 3.3

C.2.1 Comparing with (Lu et al., 2021, Assumption 6.4)

Assumption C.1 ((Lu et al. (2021), Assumption 6.4)). *For any representation functions $\psi, \psi' \in \Psi$ and $\epsilon > 0$, if there exists $v, v' \in \mathbb{R}^d$ that satisfy*

$$\mathbb{E}\|\psi(x)^\top v - \psi'(x)^\top v'\|^2 \le \epsilon$$

Then there exists a constant invertible matrix P such that

$$\|\psi(x) - P\psi'(x)\|^2 \le o(\epsilon/\|v\|^2) = o(\epsilon/\|v'\|^2).$$

for all x.

Assumption 3.3 bears similarity to Assumption C.1; however, the latter is notably more stringent. For instance, consider the case where $v = v' = e_1$ without loss of generality. If it holds that $\mathbb{E}\|\psi_1(x) - \psi_1'(x)\|^2 \le \epsilon$, then it implies $\psi \sim P\psi'$. In this context, ψ_1 and ψ_1' represent the first coordinates of ψ and ψ', respectively. The assumption that similarity in the first coordinate necessitates equivalence of the entire representations ($\psi \sim P\psi'$) is a strong assumption.

C.2.2 Case Study (Linear Representation): $\psi_\omega(x) = \omega x$ and ω is an Orthonormal Matrix

Proposition 1. *Assume that $\psi_\omega(\phi(\tau)) = \omega\phi(\tau)$ where ω is a $k \times d$ orthornormal matrix. For any representation functions $\psi_\omega, \psi_{\omega'} \in \Psi$ and $\epsilon > 0$, if there exists $\{v_i\}_{i=1}^T, \{v_i'\}_{i=1}^T$, and a trajectory distribution μ that satisfy*

$$\frac{1}{T}\sum_{i\in[T]} \mathbb{E}_{\tau\sim\mu}\|\psi_\omega(\phi(\tau))^\top v_i - \psi_{\omega'}(\phi(\tau))^\top v_i'\|^2 \le \epsilon \tag{C.3}$$

dand $V = [v_1, v_2, \cdots, v_T] \in \mathbb{R}^{k\times T}$ satisfies $\sigma_k^2(W) \ge \Omega(T/k)$, and $\|v_i\|_2 \le B$ for all $i \in [T]$. If $\Sigma := \mathbb{E}_\mu[\phi(\tau)\phi(\tau)^\intercal] \succ \mathbf{O}$, then there exists a constant invertible matrix P such that

$$\|\psi_\omega(\phi(\tau)) - P\psi_{\omega'}(\phi(\tau))\|^2 \le ck\epsilon/B$$

where $c > 0$ is a constant.

Proof. By Equation (C.3), we have

$$(\omega^\intercal V - (\omega')^\intercal V')^\intercal \Sigma(\omega^\intercal V - (\omega')^\intercal V') \le T\epsilon,$$

where $V' = [v_1', \ldots, v_T'] \in \mathbb{R}^{k\times T}$. Since $\Sigma \succ \mathbf{O}$, we have

$$\|\omega^\intercal V - (\omega')^\intercal V'\|^2 \le T\epsilon.$$

By (Yu et al., 2015, Theorem 4), there exist an orthonormal matrix P such that

$$\|\omega - P(\omega')^\intercal\|^2 \le ck\epsilon$$

where $c > 0$ is a constant, which concludes Proposition 1. □

C.3 Proof of Corollary 3.3

Corollary 3.3. (Closeness between $\psi^\star$ and ψ_ω). *Suppose Assumptions 3.1, 3.2, and 3.3 hold. For any $\delta \in (0,1]$, with probability at least $1-\delta$, if $\mathbf{r}_{\omega,\boldsymbol{\theta}} \in \mathcal{R}'(\mathcal{D})$ as specified in Algorithm 1, then there exists an orthonormal matrix P_ω such that*

$$\left[\|\psi^\star(\phi(\tau_0)) - \psi^\star(\phi(\tau_1)) - P_\omega(\psi_\omega(\phi(\tau_0)) - \psi_\omega(\phi(\tau_1)))\|^2\right] \le k\frac{c_{rep}\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_pB^2}$$

for all τ_0, τ_1, where $c_{rep} > 0$ is a constant.

Proof. By Equation (C.1), if we use Assumption 3.3 with $\Theta^\star/B$, we can find an orthonormal matrix P_ω such that

$$\left[\|\psi^\star(\phi(\tau_0)) - \psi^\star(\phi(\tau_1)) - P_\omega(\psi_\omega(\phi(\tau_0)) - \psi_\omega(\phi(\tau_1)))\|^2\right] \le k\frac{c_{\mathrm{rep}}\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_pB^2}$$

for all τ_0, τ_1, where $c_{\mathrm{rep}} > 0$ is a constant. □

C.4 Proof of Theorem 3.4

Lemma 5. *Suppose Assumptions 3.1, 3.2 and 3.3 hold. For any $\delta \in (0,1]$ and $\lambda > 0$, with probability at least $1-\delta$, $\boldsymbol{r}^\star \in \mathcal{R}'(\widehat{\mathcal{D}})$, i.e., the underlying reward functions are an element of Equation (3.2).*

Proof. Assume that Corollary 3.3 holds with probability $1-\delta/2$ for $\widehat{\omega}$, i.e.,

$$\left[\|\psi^\star(\phi(\tau_0)) - \psi^\star(\phi(\tau_1)) - P_{\widehat{\omega}}(\psi_{\widehat{\omega}}(\phi(\tau_0)) - \psi_{\widehat{\omega}}(\phi(\tau_1)))\|^2\right] \le k\frac{c_{\text{rep}}\kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_pB^2}. \tag{C.4}$$

We only consider the event that Equation (C.4) holds. We will use this $P_{\widehat{\omega}}$ for the proof of Theorem 3.4. We will approach similarly with the proof of (Zhu et al., 2023). Consider the following optimization problem:

$$\underset{\|\theta\|_i \le B}{\text{maximize}} f(\theta_i) := \frac{1}{N_p}\sum_{j\in[N_p]} \log P_{\widehat{\omega},\theta_i}(o_i^{(j)} \mid \tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}).$$

Then, we have $\widehat{\theta}_i = \underset{\|\theta\|_i \le B}{\arg\max} f(\theta_i)$ and

$$\begin{aligned}
\nabla f(\theta_i) = \frac{1}{N_p}\sum_{j\in[N_p]} \Bigg(& \frac{\Phi'(\langle \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})), \theta_i\rangle)}{\Phi(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})), \theta_i\rangle)}\mathbf{1}(o_i^{(j)} = 0) \\
& - \frac{\Phi'(\langle \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})), \theta_i\rangle)}{\Phi(\psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})), \theta_i\rangle)}\mathbf{1}(o_i^{(j)} = 1)\Bigg)\left(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))\right)
\end{aligned}$$

$$\nabla^2 f(\theta_i) = \frac{1}{N_p}\sum_{j\in[N_p]} \frac{\Phi''(x_i^{(j)})\Phi(x_i^{(j)}) - \Phi'(x_i^{(j)})^2}{\Phi(x_i^{(j)})^2}\left(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))\right)\left(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))\right)^\intercal$$

where $x_i^{(j)} = \langle \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})), \theta_i\rangle$. Here, we also define $\psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i) \in \mathbb{R}^{N_p\times k}$ such as every $j \in [N_p]$th row is $\left(\psi_\omega(\phi(\tau_{i,0}^{(j)})) - \psi_\omega(\phi(\tau_{i,1}^{(j)}))\right)$.

Then, we have

$$\nabla^2 f(\theta_i) \preceq -\eta\widehat{\Sigma}_{\psi_{\widehat{\omega}}} := -\frac{\eta}{N_p}\sum_{j\in[N_p]}\left(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))\right)\left(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))\right)^\intercal$$

where $\eta := \min_{x\in[-2R_{\max}, 2R_{\max}]}\left(\frac{\Phi'(x)^2 - \Phi''(x)\Phi(x)}{\Phi(x)^2}\right)$. For example, if $\Phi(x) = \sigma(x)$, then $\eta = \frac{1}{2+\exp(-2R_{\max})+\exp(2R_{\max})}$.

Then, by the Taylor expansion of f, we have

$$f(\widehat{\theta}_i) - f(P_{\widehat{\omega}}^\intercal\theta_i^\star) - \langle \nabla f(P_{\widehat{\omega}}^\intercal\theta_i^\star), \widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\rangle \le -\frac{\eta}{2}\|\widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\|^2_{\widehat{\Sigma}_{\psi_{\widehat{\omega}}}}.$$

Since $\widehat{\theta}_i = \underset{\|\theta\|_i \le B}{\arg\max} f(\theta_i)$, for any $\lambda > 0$, we have

$$\|\nabla f(P_{\widehat{\omega}}^\intercal\theta_i^\star)\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}}+\lambda I)^{-1}}\|\widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\|_{\widehat{\Sigma}_{\psi_{\widehat{\omega}}}+\lambda I} \ge \langle \nabla f(P_{\widehat{\omega}}^\intercal\theta_i^\star), \widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\rangle \ge \frac{\eta}{2}\|\widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\|^2_{\widehat{\Sigma}_{\psi_{\widehat{\omega}}}}. \tag{C.5}$$

We define a random vector $V \in \mathbb{R}^{N_p}$ as follows:

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

for all $j \in [N_p]$. Define $\xi = \max_{x \in [-2R_{\max}, 2R_{\max}]} \left| \frac{\Phi'(x)}{\Phi(x)} \right|$. If $\Phi(x) = \sigma(x)$, $\xi \le 1$. Then, we can verify that $\mathbb{E}[V] = 0$ and $|V_j| \le \xi$ for all $j \in [N_p]$.

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

$$V_j' = \begin{cases} \frac{\Phi'(\langle \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})), P_{\widehat{\omega}}^\intercal \theta_i^\star \rangle)}{\Phi(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})), P_{\widehat{\omega}}^\intercal \theta_i^\star \rangle)} & \text{w.p.} \quad \Phi(\psi^\star(\phi(\tau_{i,0}^{(j)})) - \psi^\star(\phi(\tau_{i,1}^{(j)})), \theta_i^\star \rangle) \\ -\frac{\Phi'(\langle \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})), P_{\widehat{\omega}}^\intercal \theta_i^\star \rangle)}{\Phi(\psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})), P_{\widehat{\omega}}^\intercal \theta_i^\star \rangle)} & \text{w.p.} \quad \Phi(\psi^\star(\phi(\tau_{i,1}^{(j)})) - \psi^\star(\phi(\tau_{i,0}^{(j)})), \theta_i^\star \rangle) \end{cases}$$

for all $j \in [N_p]$. $\nabla f(P_{\widehat{\omega}}^\intercal \theta_i^\star)$ can be written as

$$\nabla f(P_{\widehat{\omega}}^\intercal \theta_i^\star) = \frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal V_i' = \frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal V_i + \frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal (V_i' - V_i).$$

Therefore, we can bound $\|\nabla f(P_{\widehat{\omega}}^\intercal \theta_i^\star)\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1}}$ by

$$\|\nabla f(P_{\widehat{\omega}}^\intercal \theta_i^\star)\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1}} \le \underbrace{\|\frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal V_i\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1}}}_{(i)} + \underbrace{\|\frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal (V_i' - V_i)\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1}}}_{(ii)}.$$

Step 1: Bounding (i).

Define $M = \frac{1}{N_p^2} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal$, then we have

$$\|\frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal V_i\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1}} = V^\intercal M V.$$

We can check

$$\mathrm{Tr}(M) \le \frac{k}{N_p}, \qquad \mathrm{Tr}\left(M^2\right) \le \frac{k}{N_p^2}, \qquad \|M\|_F = \sigma_1(M) \le \frac{1}{N_p}$$

in the same way with (Zhu et al., 2023, Page 19). Therefore, as V's components are bounded, independent, and $\mathbb{E}V = \mathbf{0}$, we can use Bernstein's inequality in quadratic form (for example, (Hsu et al., 2012, Theorem 2.1) and (Zhu et al., 2023, Page 19)), so we have

$$\|\frac{1}{N_p} \psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal V_i\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}} + \lambda I)^{-1}} \le \xi C_4 \sqrt{\frac{k + \log(N/\delta)}{N_p}} \tag{C.6}$$

for a constant $C_4 > 0$ with probability at least $1 - \delta/(2N)$.

Step 2: Bounding (ii).
We have $\left|\frac{\Phi'(x)}{\Phi(x)} - \frac{\Phi'(y)}{\Phi(y)}\right| \le \xi|x-y|$ by the mean value theorem if $x, y \in [-2R_{\max}, 2R_{\max}]$, so

$$
\begin{aligned}
|V_i - V_i'| &\le \max_{\tau_0,\tau_1} \xi |\langle (\psi^\star(\phi(\tau_0)) - \psi^\star(\phi(\tau_1))) - (P_{\widehat{\omega}}\psi_{\widehat{\omega}}(\phi(\tau_0)) - P_{\widehat{\omega}}\psi_{\widehat{\omega}}(\phi(\tau_1))), \theta_i^\star\rangle| \\
&\le \xi\sqrt{k\frac{c_{\mathrm{rep}}\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}.
\end{aligned}
$$

Therefore, we have

$$
\|\frac{1}{N_p}\psi_{\widehat{\omega}}(\widehat{\mathcal{D}}_i)^\intercal (V_i' - V_i)\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}}+\lambda I)^{-1}} \le \frac{\xi C_5}{\sqrt{N_p}}\sqrt{k\frac{\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}} \tag{C.7}
$$

where $C_5 > 0$ is a constant.

Step 3: Combining (i) and (ii).
Combining Equation (C.6) and Equation (C.7), we have

$$
\begin{aligned}
\|\nabla f(P_{\widehat{\omega}}^\intercal\theta_i^\star)\|_{(\widehat{\Sigma}_{\psi_{\widehat{\omega}}}+\lambda I)^{-1}} &\le \frac{\xi C_5}{\sqrt{N_p}}\sqrt{k\frac{\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}} + \xi C_4\sqrt{\frac{k+\log(N/\delta)}{N_p}} \\
&\le C_6\sqrt{k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p} + \frac{\xi^2(k+\log(N/\delta))}{N_p}}
\end{aligned}
$$

for a constant $C_6 > 0$ with probability at least $1-\delta/N$ and Equation (C.5) provides

$$
\|\widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\|_{\widehat{\Sigma}_{\psi_{\widehat{\omega}}}} \le C_7\sqrt{k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2 NN_p} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2 N_p} + \lambda B^2},
$$

which is equivalent to

$$
\begin{aligned}
&\frac{1}{N_p}\sum_{j\in[N_p]}\left|\langle(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)})) - \psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))), \widehat{\theta}_i - P_{\widehat{\omega}}^\intercal\theta_i^\star\rangle\right|^2 \\
&\qquad \le C_7^2\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2 NN_p} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2 N_p} + \lambda B^2\right),
\end{aligned}
$$

with probability at least $1-\delta/N$.

Now, we will bound $\frac{1}{N_p}\sum_{j\in[N_p]}\left|(r_{\widehat{\omega},\widehat{\theta}_i}(\tau_{i,0}^{(j)})-r_{\widehat{\omega},\widehat{\theta}_i}(\tau_{i,1}^{(j)}))-(r_i^\star(\tau_{i,0}^{(j)})-r_i^\star(\tau_{i,1}^{(j)}))\right|^2$:

$$
\begin{aligned}
&\frac{1}{N_p}\sum_{j\in[N_p]}\left|(r_{\widehat{\omega},\widehat{\theta}_i}(\tau_{i,0}^{(j)})-r_{\widehat{\omega},\widehat{\theta}_i}(\tau_{i,1}^{(j)}))-(r_i^\star(\tau_{i,0}^{(j)})-r_i^\star(\tau_{i,1}^{(j)}))\right|^2\\
&=\frac{1}{N_p}\sum_{j\in[N_p]}\left|\langle\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)}))-\psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)})),\widehat{\theta}_i\rangle-\langle\psi^\star(\phi(\tau_{i,0}^{(j)}))-\psi^\star(\phi(\tau_{i,1}^{(j)})),\theta_i^\star\rangle\right|^2\\
&\le\frac{2}{N_p}\sum_{j\in[N_p]}\left|\langle(\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)}))-\psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))),\widehat{\theta}_i-P_{\widehat{\omega}}^\intercal\theta_i^\star\rangle\right|^2\\
&\quad+\frac{2}{N_p}\sum_{j\in[N_p]}\left|\langle\psi_{\widehat{\omega}}(\phi(\tau_{i,0}^{(j)}))-\psi_{\widehat{\omega}}(\phi(\tau_{i,1}^{(j)}))-P_{\widehat{\omega}}(\psi^\star(\phi(\tau_{i,0}^{(j)}))-\psi^\star(\phi(\tau_{i,1}^{(j)}))),\theta_i^\star\rangle\right|^2\\
&\le 2C_7\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2NN_p}+\frac{\xi^2(k+\log(N/\delta))}{\eta^2N_p}+\lambda B^2\right)\\
&\quad+\frac{2}{N_p}N_pk\frac{c_{\text{rep}}\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}\\
&\le C_8\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2NN_p}+\frac{\xi^2(k+\log(N/\delta))}{\eta^2N_p}+\lambda B^2\right)
\end{aligned}
$$

for a constant $C_8>0$. Combining this result for all $i\in[N]$, Lemma 5 holds. □

Lemma 6. *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. For any $\delta\in(0,1]$, with probability at least $1-\delta$, for any $\boldsymbol{r}_{\omega,\boldsymbol{\theta}}\in\mathcal{R}'(\widehat{\mathcal{D}})$,*

$$
\begin{aligned}
&\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_{\omega,\theta_i}(\tau_{i,0})-r_{\omega,\theta_i}(\tau_{i,1}))-(r_i^\star(\tau_{i,0})-r_i^\star(\tau_{i,1}))\right|^2\right]\\
&\qquad\le C_9\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2NN_p}+\frac{\xi^2(k+\log(N/\delta))}{\eta^2N_p}+\lambda B^2\right)
\end{aligned}
$$

where $C_9>0$ is a constant.

Proof. For any τ_0,τ_1, by Assumption 3.4, with large $N_p\ge N_{\text{unif}}(\Psi,\mu_0,\mu_1,\delta)$, we have the analog of Equation (C.1):

$$
\begin{aligned}
&\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_{\omega,\theta_i}(\tau_{i,0})-r_{\omega,\theta_i}(\tau_{i,1}))-(r_i^\star(\tau_{i,0})-r_i^\star(\tau_{i,1}))\right|^2\right]\\
&=\begin{bmatrix}\theta_i\\-\theta_i^\star\end{bmatrix}^\intercal\Lambda_{\phi_\omega,\phi_{\psi^\star}}(\mu_0,\mu_1)\begin{bmatrix}\theta_i\\-\theta_i^\star\end{bmatrix}\le 1.1\begin{bmatrix}\theta_i\\-\theta_i^\star\end{bmatrix}^\intercal\widehat{\Lambda}_{\phi_\omega,\phi_{\psi^\star}}(\mu_0,\mu_1)\begin{bmatrix}\theta_i\\-\theta_i^\star\end{bmatrix}\\
&\le 1.1C_8\left(k\frac{\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2NN_p}+\frac{\xi^2(k+\log(N/\delta))}{\eta^2N_p}+\lambda B^2\right)
\end{aligned}
$$

which concludes the proof. □

Theorem 3.4. (Improved Expected Value Function Gap). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. For any $\delta \in (0,1]$, all $i \in [N]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $\widehat{\pi}'_i$ of Algorithm 1 satisfies*

$$J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}'_i; r_i^\star) \\ \leq \sqrt{cC_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,tar}, \mu_{i,ref}, i)^2 \left(k\frac{\xi^2\kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{\eta^2 NN_p} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2 N_p} + \lambda B^2\right)} \tag{3.5}$$

where $c > 0$ is a constant.

Proof. We have

$$J(\pi_{i,\text{tar}}; r_i^\star) - J(\widehat{\pi}'_i; r_i^\star) \\ \leq C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,\text{tar}}, \mu_{i,\text{ref}}, i)\sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,1}) - r_i^\star(\tau_{i,0})) - (r^{i,\inf}_{\pi_{i,\text{tar}}}(\tau_{i,1}) - r^{i,\inf}_{\pi_{i,\text{tar}}}(\tau_{i,0}))\right|^2\right]} \\ \leq \sqrt{cC_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{i,\text{tar}}, \mu_{i,\text{ref}}, i)^2 \left(\frac{k\kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2 N_p} + \lambda B^2\right)}$$

where $c > 0$ is a constant, which is similar to the proof of Theorem 3.1. □

In the exactly same way, we can prove Theorem 3.5, so we omit the proof of Theorem 3.5

C.5 Proof of Theorem 3.6

Theorem 3.6. (Lower Bound for the Sub-Optimality Gap of Personalization). *For any $k > 6, N_p \geq Ck\Lambda^2$ and $\Lambda \geq 2$, there exists a representation function $\phi(\cdot)$ so that*

$$\min_{i\in[N]} \inf_{\widehat{\pi}} \sup_{Q\in\text{CB}(\Lambda)} \left(\max_{\pi^*\in\Pi} J(\pi^*; r_{\omega,\theta_i}) - J(\widehat{\pi}; r_{\omega,\theta_i})\right) \geq C\Lambda \cdot \sqrt{\frac{k}{N_p}},$$

where

$$\text{CB}(\Lambda) := \left\{Q := \left(\{\mu_0, \mu_1\}, \{\tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\}_{i\in[N], j\in[N_p]}, \omega, \boldsymbol{\theta}\right) \middle| C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi^\star, \mu_1, i) \leq \Lambda \textit{ for all } i \in [N]\right\}$$

is the family of MDP with N reward functions and $H = 1$ instances.

Proof of Theorem 3.6. We follow the construction in Theorem 3.10 of Zhu et al. (2023).

We will only consider $H = 1$ case. Assume k can be divided by 3 without loss of generality. Let $\mathcal{S} := \{0, 1, ..., k/3 - 1\}$ and $\mathcal{A} := \{a_1, a_2, a_3\}$. Let $\psi_\omega(\phi(s, a_1)) = e_{3s}$, $\psi_\omega(\phi(s, a_2)) = e_{3s+1}$, and $\psi_\omega(\phi(s, a_3)) = 0$. Also, let $v_{-1} := \{1/d, 1/d + \Delta, -2/d - \Delta\}$ and $v_1 := \{1/d + 2\Delta, 1/d + \Delta, -2/d - 3\Delta\}$. We construct $2^{|\mathcal{S}|}$ instances in CB. Let $w \in \{\pm 1\}^{|\mathcal{S}|}$ and $\theta_w := [v_{w_1}, v_{w_2}, ..., v_{w_{|\mathcal{S}|}}]$. Let $\mu_0(s, a_1) = \frac{1-2\Lambda^2}{|\mathcal{S}|}$, $\mu_0(s, a_2) = \frac{2\Lambda^2}{|\mathcal{S}|}$, and $\mu_1(s, a_3) = 1$ for any $s \in \mathcal{S}$.

According to Zhu et al. (2023), $\left\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}\left[\psi_\omega(\phi(s, \pi^\star(s)))\right]\right\|_2 \leq \Lambda$, where ρ is the uniform distribution over $\mathcal{S}$. At the same time, for any θ_w we have $\|\theta_w\|_2 \in \Theta_B$ when taking $B = 1$, $d > 6$ and $\Delta < 1/(6d)$.

Next, we will show that $C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi^\star, \mu_1, i) \leq \Lambda$. By definition, we have

$$\left\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}\left[\psi_\omega(\phi(s,\pi^\star(s)))\right]\right\|_2 = \left\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho,a\sim\pi^\star(\cdot|s),(s',a')\sim\mu_1}\left[\psi_\omega(\phi(s,\pi^\star(s))) - \psi_\omega(\phi(s',a'))\right]\right\|_2,$$

since $a' \equiv a_3$ by definition of μ_1 and $\psi_\omega(\phi(\cdot, a_3) \equiv 0$. Then, by Section D.1. of Zhan et al. (2023), we have $C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi^\star, \mu_1, i) \leq \left\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}\left[\psi_\omega(\phi(s,\pi^\star(s)))\right]\right\|_2 \leq \Lambda$. Therefore, combined with Theorem 3.10 of Zhu et al. (2023), we finished the proof. □

D Proof of Section 4

Corollary 3.3 holds with probability $1-\delta/3$, so we have

$$\max_{\tau_0,\tau_1} \|(\psi^\star(\phi(\tau_0)) - \psi^\star(\phi(\tau_1))) - P_{\widehat{\omega}}(\psi_{\widehat{\omega}}(\phi(\tau_0)) - \psi_{\widehat{\omega}}(\phi(\tau_1)))\|^2 \leq k\frac{C_3\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_pB^2}$$

Claim 1. *For any $\delta \in (0,1]$, with probability at least $1-\delta$, for arbitrary $\mathcal{D}_i$ and $\mathcal{D}_j$, the gap between label discrepency with reward function class $\mathcal{G}_{\psi_{\widehat{\omega}}}$ and $\mathcal{G}_{\psi^\star}$ is bounded as follows:*

$$\left|\boldsymbol{disc}(\boldsymbol{D}_i, \boldsymbol{D}_j, \mathcal{G}_{\psi_{\widehat{\omega}}}) - \boldsymbol{disc}(\boldsymbol{D}_i, \boldsymbol{D}_j, \mathcal{G}_{\psi^\star})\right| \leq 2C_{10}\sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}$$

for $i,j \in [N]$ where $C_{10} > 0$ is a constant. We recall the definition of $\mathcal{G}_{\psi_w} = \{\langle\psi_w, \theta\rangle \mid \|\theta\|_2 \leq B\}$.

Proof.

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{D}_i}\log P_{\widehat{\omega},P_{\widehat{\omega}}^\intercal\theta}(o \mid \tau_1,\tau_0) - \mathbb{E}_{\boldsymbol{D}_i}\log P_{\omega^\star,\theta}(o \mid \tau_1,\tau_0)\right| \\
&\leq \mathbb{E}_{\boldsymbol{D}_i}\left|\log P_{\widehat{\omega},P_{\widehat{\omega}}^\intercal\theta}(o \mid \tau_1,\tau_0) - \log P_{\omega^\star,\theta}(o \mid \tau_1,\tau_0)\right| \\
&\leq \xi\mathbb{E}_{\boldsymbol{D}_i}\left|\langle P_{\widehat{\omega}}(\psi_{\widehat{\omega}}(\phi(\tau_1)) - \psi_{\widehat{\omega}}(\phi(\tau_0))) - (\psi^\star(\phi(\tau_1)) - \psi^\star(\phi(\tau_0))), \theta\rangle\right| \\
&\leq C_{10}\sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}
\end{aligned}$$

where $\xi := \max_{x\in[-R_{\max},R_{\max}]}\left|\frac{\Phi'(x)}{\Phi(x)}\right|$, which is also defined in Appendix C. □

Theorem 4.1. (Total Expected Value Function Gap). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. Also, assume that $C_r(\mathcal{G}_r, \pi, \mu_{i,ref}, i) \leq C'_{max}$ for all policy π and $i \in [N]$. For any $\delta \in (0,1]$, all $i \in [N]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $((\widehat{\pi}_{(k)})_{k\in[K]}, \widehat{f})$ of Algorithm 3 satisfies*

$$\begin{aligned}
&\sum_{i\in[N]}\left(J(\pi_{i,tar}; r_i^\star) - J(\widehat{\pi}_{\widehat{f}(i)}; r_i^\star)\right) \\
&\leq cN\kappa\left(\frac{\log(2K/\delta)}{N_p} + \frac{kK\log(N_p/K)}{N_p} + \frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}\right. \\
&\qquad\left. + \left(\sum_{i\in[N]}\frac{1}{N}\boldsymbol{disc}(\mathcal{D}_i, \mathcal{C}_{\widehat{f}(i)}, \mathcal{G}_{\psi^\star})\right)^2 + \left(\frac{\log(\mathcal{N}_{\mathcal{G}_{\psi^\star}}(1/NN_p)/\delta)}{NN_p}\right)^2\right)^{1/4},
\end{aligned}$$

where $c > 0$ is a constant.

Proof. By Claim 1 with Lemma 1, we have

$$
\begin{aligned}
&\sum_{i\in[N]}\sum_{j\in[N_{p,i}]}\log\left(\frac{P_{\omega^\star,\theta_i^\star}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}{P_{\widehat{\omega},\widehat{\theta}_{\widehat{f}(i)}}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}\right)\\
&\le \max_{\|\theta_i'\|\le B \text{ for all } i\in[N]}\sum_{i\in[N]}\sum_{j\in[N_{p,i}]}\log\left(\frac{P_{\omega^\star,\theta_i'}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}{P_{\widehat{\omega},\widehat{\theta}_{\widehat{f}(i)}}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}\right)\\
&\underset{(i)}{\le} \max_{\|\theta_i'\|\le B \text{ for all } i\in[N]}\sum_{i\in[N]}\sum_{j\in[N_{p,i}]}\log\left(\frac{P_{\widehat{\omega},\theta_i'}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}{P_{\widehat{\omega},\widehat{\theta}_{\widehat{f}(i)}}(o_i^{(j)}\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})}\right)\\
&\qquad + NN_pC_{10}\sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}\\
&\le C_{\text{cluster}}NN_p\Bigg(\sqrt{\frac{\log(2K/\delta)}{N_p}}+\sqrt{\frac{kK\log(N_p/K)}{N_p}}+\sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}\\
&\qquad\qquad + \sum_{i\in[N]}\frac{1}{N}\mathtt{disc}(\mathcal{D}_i,\mathcal{C}_{\widehat{f}(i)},\mathcal{G}_{\psi_{\widehat{w}}})\Bigg)\\
&\le C_{11}NN_p\Bigg(\sqrt{\frac{\log(2K/\delta)}{N_p}}+\sqrt{\frac{kK\log(N_p/K)}{N_p}}+\sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}\\
&\qquad\qquad + \sum_{i\in[N]}\frac{1}{N}\mathtt{disc}(\mathcal{D}_i,\mathcal{C}_{\widehat{f}(i)},\mathcal{G}_{\psi^\star}))\Bigg),
\end{aligned}
$$

where $\widehat{\omega}$ is a learned parameter from the representation learning, and $C_{11} > 0$ is a constant. Here, (i) came from the same reason with Claim 1. Therefore, by Lemma 3, we have

$$
\begin{aligned}
&\mathbb{E}_{\mu_0,\mu_1}\left[\|P_{\widehat{\omega},\widehat{\theta}_{f(i)}}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})-P_{w^\star,\theta^\star}(\cdot\mid\tau_{i,0}^{(j)},\tau_{i,1}^{(j)})\|_1^2\right]\\
&\le C_{11}\Bigg(\sqrt{\frac{\log(2K/\delta)}{N_p}}+\sqrt{\frac{kK\log(N_p/K)}{N_p}}+\sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}\\
&\qquad\qquad + \sum_{i\in[N]}\frac{1}{N}\mathtt{disc}(\mathcal{D}_i,\mathcal{C}_{\widehat{f}(i)},\mathcal{G}_{\psi^\star}))+\frac{\log(\mathcal{N}_{\mathcal{G}_{\psi_{\widehat{\omega}}}}(1/NN_p)/\delta)}{NN_p}\Bigg).
\end{aligned}
$$

Here, we used $\mathcal{N}_{\mathcal{G}_{\psi_{\widehat{\omega}}}}(1/NN_p) = \mathcal{N}_{\mathcal{G}_{\psi^\star}}(1/NN_p)$. Now, we get the similar bound with Equation (C.1):

$$\frac{1}{N}\sum_{i\in[N]}\mathbb{E}_{\mathcal{D}_i}\left[\left|(r_{\widehat{\omega},\widehat{\theta}_{\widehat{f}(i)}}(\tau_{i,0}) - r_{\widehat{\omega},\widehat{\theta}_{\widehat{f}(i)}}(\tau_{i,1})) - (r_i^\star(\tau_{i,0}) - r_i^\star(\tau_{i,1}))\right|^2\right]$$
$$\le C_{11}\kappa^2\Bigg(\sqrt{\frac{\log(2K/\delta)}{N_p}} + \sqrt{\frac{kK\log(N_p/K)}{N_p}} + \sqrt{\frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}}$$
$$+\sum_{i\in[N]}\frac{1}{N}\mathtt{disc}(\mathcal{D}_i, \mathcal{C}_{\widehat{f}(i)}, \mathcal{G}_{\psi^\star}) + \frac{\log(\mathcal{N}_{\mathcal{G}_{\psi^\star}}(1/NN_p)/\delta)}{NN_p}\Bigg).$$

Lastly, we use the following:

$$\begin{aligned}
&J(\pi_{i,\mathrm{tar}}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)\\
&= (J(\pi_{i,\mathrm{tar}}; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(r_i^\star(\tau))) - (J(\widehat{\pi}_i; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(r_i^\star(\tau)))\\
&= (J(\pi_{i,\mathrm{tar}}; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(r_i^\star(\tau))) - (J(\pi_{i,\mathrm{tar}}; \widehat{r}_i) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(\widehat{r}_i(\tau)))\\
&\quad + (J(\pi_{i,\mathrm{tar}}; \widehat{r}_i) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(\widehat{r}_i(\tau))) - (J(\widehat{\pi}_j; \widehat{r}_i) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(\widehat{r}_i(\tau)))\\
&\quad + (J(\widehat{\pi}_i; \widehat{r}_i) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(\widehat{r}_i(\tau))) - (J(\widehat{\pi}_i; r_i^\star) - \mathbb{E}_{\tau\sim\mu_{i,\mathrm{ref}}}(r_i^\star(\tau)))\\
&\le 2C'_{\max}\sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(r_i^\star(\tau_{i,0}) - r_i^\star(\tau_{i,1})) - (\widehat{r}_i(\tau_{i,0}) - \widehat{r}_i(\tau_{i,1}))\right|^2\right]}
\end{aligned}$$

where the last inequality came from the fact that $\widehat{\pi}_i$ is the best policy with respect to $\widehat{r}_{f(i)}$. Therefore, summing the above relationship with $i \in [N]$ provides

$$\sum_{i\in[N]}\Big(J(\pi_{i,\mathrm{tar}}; r_i^\star) - J(\widehat{\pi}_i; r_i^\star)\Big)$$
$$\le C_{12}N\kappa\Bigg(\frac{\log(2K/\delta)}{N_p} + \frac{kK\log(N_p/K)}{N_p} + \frac{k\xi^2\kappa^2\log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/\delta)}{NN_p}$$
$$+\Bigg(\sum_{i\in[N]}\frac{1}{N}\mathtt{disc}(\mathcal{D}_i, \mathcal{C}_{\widehat{f}(i)}, \mathcal{G}_{\psi^\star})\Bigg)^2 + \left(\frac{\log(\mathcal{N}_{\mathcal{G}_{\psi^\star}}(1/NN_p)/\delta)}{NN_p}\right)^2\Bigg)^{1/4}.$$

□

E Proof of Section 5

E.1 Six Pivotal Axioms for Reward Aggregation

For the completeness of the paper, we introduce six pivotal axioms for reward aggregation (Moulin, 2004).

- **Monotonicity**: For two reward vectors, $\boldsymbol{r} = (r_1, \ldots, r_N)^\top$ and $\boldsymbol{r}' = (r_1', \ldots, r_N')^\top$ such that $r_i = r_i'$ for $i \neq j$ and $r_j > r_j'$ for some $j \in [N]$, then $\boldsymbol{r} \succ \boldsymbol{r}'$. This is related to Pareto optimality, indicating that if one vector is strictly better than another in at least one dimension and no worse in any other, it is considered superior.

- **Symmetry**: The reward aggregation function should treat all individuals equally. The outcome should not depend on the identities of the individuals but only on their rewards.
- **Independence of Unconcerned Agents**: If for an individual $j \in [N]$, $r_j = r'_j$, then the magnitude of r_j does not influence the comparison between $\boldsymbol{r}$ and $\boldsymbol{r}'$.
- **The Pigou-Dalton Transfer Principle**: If $r_i < r_j$ and $r'_i + r_j = r'_j + r_i$ for a pair $(i,j) \in [N] \times [N]$ and $r_k = r'_k$ for all $k \neq i, j \in [N]$, then $\boldsymbol{r}' \succ \boldsymbol{r}$. This condition implies that, all else being equal, a social welfare function should favor allocations that are more equitable, reflecting a preference for balancing the rewards between individuals i and j.
- **Translation Independence**: If $\boldsymbol{r} \succ \boldsymbol{r}'$, then $\boldsymbol{r} + c \succ \boldsymbol{r}' + c$ for $c \in \mathbb{R}^N$.
- **Continuity**: In the context of social choice with a continuous preference scale, continuity means that small changes in the individual preferences should not lead to abrupt changes in the collective decision.

Equation (5.2) and its monotonically increasing transformation is only reward aggregation that satisfying the above six axioms. In (Zhong et al., 2024), the consider *Scale Independence* rather than *Translation Independence*, which is defined as follows:

- **Scale Independence**: If $\boldsymbol{r} \succ \boldsymbol{r}'$, then $\lambda \cdot \boldsymbol{r} \succ \lambda \cdot \boldsymbol{r}'$ for $\lambda > 0$.

In this case, the reward aggregations that satisfying six axioms are

$$\mathrm{Agg}_\alpha(\boldsymbol{r}) = \begin{cases} \frac{1}{N\alpha} \sum_{i \in [N]} r_i^\alpha & \alpha \neq 0 \\ \prod_{i \in [N]} r_i & \alpha = 0 \end{cases}$$

for $\alpha \in [-\infty, \infty]$.

E.2 Proof of Theorem 5.1

Theorem 5.1. (Expected Value Function Gap). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 hold. For any $\delta \in (0,1]$, all $i \in [N]$ and $\lambda > 0$, with probability at least $1-\delta$, the output $\widehat{\pi}$ of Algorithm 5 satisfies*

$$\begin{aligned} & J(\pi_{tar}; Agg_\alpha(\boldsymbol{r}^\star)) - J(\widehat{\pi}; Agg_\alpha(\boldsymbol{r}^\star)) \\ & \leq \sqrt{c_\alpha C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{tar}, \mu_{ref})^2 \left(\frac{k\kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/(\delta/N))}{NN_p} + \frac{\xi^2(k + \log(N/\delta))}{\eta^2 N_p} + \lambda B^2 \right)} \end{aligned}$$

where $c_\alpha > 0$ is a constant depending on α, and other constants are defined in Section 3.1.

Proof. Define $C_\alpha := \max_{x,y,z,w \in [-R_{\max}, R_{\max}]} \frac{|(\exp(\alpha x) - \exp(\alpha y)) - (\exp(\alpha z) - \exp(\alpha w))|}{\alpha |(x-y)-(z-w)|}$ for $\alpha \neq 0$ and $C_\alpha = 1$ for

$\alpha = 0$. Then we know that $C_\alpha < \infty$. Now, in the same way of proof of Theorem 3.4, we have

$$
\begin{aligned}
& J(\pi_{\text{tar}}; \text{Agg}_\alpha(r_1^\star, \ldots, r_N^\star)) - J(\widehat{\pi}; \text{Agg}_\alpha(r_1^\star, \ldots, r_N^\star)) \\
& \leq C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{\text{tar}}, \mu_{\text{ref}}) \sqrt{\mathbb{E}_{\mu_0,\mu_1}\left[\left|(\text{Agg}_\alpha(\boldsymbol{r}^\star(\tau_1)) - \text{Agg}_\alpha(\boldsymbol{r}^\star(\tau_0))) - (\text{Agg}_\alpha(\boldsymbol{r}^{\inf}_{\pi_{\text{tar}}}(\tau_1)) - \text{Agg}_\alpha(\boldsymbol{r}^{\inf}_{\pi_{\text{tar}}}(\tau_0)))\right|^2\right]} \\
& \leq C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi_{\text{tar}}, \mu_{\text{ref}}) \sqrt{C_\alpha^2 \mathbb{E}_{\mu_0,\mu_1}\left[\frac{1}{N}\sum_{i\in[N]} \left|(r_i^\star(\tau_1) - r_i^\star(\tau_0)) - (r^{\inf}_{\pi_{\text{tar}}}(\tau_1) - r^{\inf}_{\pi_{\text{tar}}}(\tau_0))\right|^2\right]} \\
& \leq \sqrt{c_\alpha\left(\frac{k\kappa^2 \log(\mathcal{N}_{\mathcal{G}_r}(1/(NN_p))/(\delta/N))}{NN_p^2} + \frac{\xi^2(k+\log(N/\delta))}{\eta^2 N_p} + \lambda B^2\right)}.
\end{aligned}
$$

where the last line is from Lemma 6, which conclude the proof. □

E.3 Proof of Theorem 5.2

Theorem 5.2. (Lower Bound for the Sub-Optimality Gap of Aggregation). *For any $k > 6, N_p \geq Ck\Lambda^2, \Lambda \geq 2$, and $\alpha \in \mathbb{R}$ there exists a representation function $\phi(\cdot)$ so that*

$$
\inf_{\widehat{\pi}} \sup_{Q\in \text{CB}(\Lambda)} \left(\max_{\pi^*\in\Pi} J(\pi^*; Agg_\alpha(\boldsymbol{r}_{\omega,\boldsymbol{\theta}})) - J(\widehat{\pi}; Agg_\alpha(\boldsymbol{r}_{\omega,\boldsymbol{\theta}}))\right) \geq C\Lambda \cdot \sqrt{\frac{k}{N_p}},
$$

where

$$
\text{CB}(\Lambda) := \left\{ Q := \left(\{\mu_0, \mu_1\}, \{\tau_{i,0}^{(j)}, \tau_{i,1}^{(j)}\}_{i\in[N], j\in[N_p]}, \omega, \boldsymbol{\theta}\right) \,\middle|\, C_{\boldsymbol{r}}(\mathcal{G}_{\boldsymbol{r}}, \pi^\star, \mu_1, i) \leq \Lambda \text{ for all } i \in [N] \right\}
$$

is the family of MDP with N reward functions and $H = 1$ instances.

Proof. We start with the same setting and the same instances that achieve the lower bounds with Theorem 3.6. Since

$$
\mathbb{E}_s[\text{Agg}_\alpha(\boldsymbol{r})(s,\pi^\star) - \text{Agg}_\alpha(\boldsymbol{r})(s,\pi')] \geq \Omega\left(\mathbb{E}_s[\sum_{i\in[N]}(r_i(s,\pi^\star) - r_i(s,\pi'))]\right) \geq \Omega\left(C\Lambda \cdot \sqrt{\frac{k}{N_p}}\right)
$$

We can finish the proof for all $\alpha \in \mathbb{R}$. The first inequality holds by definition when $\alpha = 0$. When $\alpha \neq 0$, for any $i \in [N]$, we have $\exp(r_i(s,\pi^\star)) - \exp(r_i(s,\pi')) \geq \exp(-R_{\max}) \left|r_i(s,\pi^\star) - r_i(s,\pi')\right| \geq \Omega\left(C\Lambda \cdot \sqrt{\frac{k}{N_p}}\right)$. □

E.4 Proof of Theorem 5.3

Theorem 5.3. (Relationship between Reward Aggregation and Preference Aggregation). *Suppose human preferences are modeled by the PL model, and all human labelers share a common lower bound on their reward functions. Let $(R_i(a))_{a\in\mathcal{A}}$ represent the reward function associated with action $a \in \mathcal{A}$ and $P_i \in \Delta(\mathcal{A})$ denote the corresponding probabilistic opinion for individual $i \in [N]$. Then, the preference aggregation Agg-$p_\alpha(\mathbf{P})$, is equivalent to the preference derived under the PL model with the aggregated rewards $(Agg_\alpha(\mathbf{R}(a)))_{a\in\mathcal{A}}$ for any $\alpha \in [-\infty, \infty]$.*

Proof. By the PL modeling, we have

$$P_i(a) = \frac{\exp(R_i(a))}{\sum_{a' \in \mathcal{A}} \exp(R_i(a'))}. \tag{E.1}$$

We divide Equation (E.1) by $P_i(a_{\mathrm{fix}})$, we have

$$R_i(a) = \log P_i(a) - (\log P_i(s, a_{\mathrm{fix}}) - R_i(a_{\mathrm{fix}})) := \log P_i(a) - C_i \tag{E.2}$$

where $C_i := \log P_i(a_{\mathrm{fix}}) - R_i(a_{\mathrm{fix}})$. Since $R_i(a)$ have upper bound as C_i, and we assumed that every reward $R_i(a)$ have the same upper bound, we can assume $C_i = C$ for every i. Therefore, plugging Equation (E.2) provides the equivalence between $\mathrm{Agg}_\alpha(\boldsymbol{R})$ and $\text{Agg-p}_\alpha(\boldsymbol{P})$. □

E.5 Relationship between KL divergence and variant of α-Renyi divergence.

By L'Hôpital's rule, we have

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha}\left(1 - \sum_{j \in \mathcal{A}} p_{ij} \left(\frac{q_{ij}}{p_{ij}}\right)^{1-\alpha}\right) = \sum_{j \in \mathcal{A}} \lim_{\beta \to 0} \left(-p_{ij} \log\left(\frac{q_{ij}}{p_{ij}}\right)\left(\frac{q_{ij}}{p_{ij}}\right)^{\beta}\right) = \mathsf{KL}(p, q).$$

E.6 Proof of Section 5.3.2

The proof of Theorem 5.4 is exactly the same as the proof of the fact that the VCG mechanism is DSIC welfare-maximizing. The difference with the proof of the original VCG mechanism's property is the parametrization of bidding, which will be explained in this section.

Theorem 5.4. (DSIC Welfare-Maximizing Mechanism). *The aggregation rule and the cost function as in Equation* (5.5) *provide a DSIC welfare-maximizing mechanism.*

Proof. The aggregated result space $\Delta(\mathcal{A})$ corresponds to the output space X of Definition 5.1. We can interpret the bidding part, $b_j(x)$, of Definition 5.1 as $-d(P_j, p)$. So, instead of bidding on every output without any rule, we can interpret the bidding as the minus distance function between their own probabilistic opinion and aggregated probabilistic opinion. The underlying value function therefore corresponds to $-d(p_j, p)$. This interpretation provides the same line of proof of the VCG mechanism's property. □

By good parametrization of the VCG mechanism, we can also achieve the computational efficiency of our cost function computation.

Theorem 5.5. *If we set d as a variant of the α-Renyi distance for $\alpha \neq 0$ (Remark 5) and define d as KL-divergence for $\alpha = 0$, the DSIC welfare-maximizing aggregation rule is Equation* (5.4)*. Therefore, aggregation rule Equation* (5.4) *is also welfare-maximizing with appropriate cost function.*

Proof. We solve the optimization problem as follows:

$$\underset{p \in \Delta(\mathcal{A})}{\arg\min} \sum_{i \in [N]} d_\alpha(P_i, p) \tag{E.3}$$

where $d_\alpha(p,q) = \mathrm{sgn}(\alpha)\frac{1}{1-\alpha}\sum_{j\in\mathcal{A}}\left(1-p_j^\alpha q_j^{1-\alpha}\right)$. We can check that $d_\alpha(p,q)$ is a convex function with respect to q, as

$$\frac{d^2}{dq_j^2}d_\alpha(p,q) = \alpha\mathrm{sgn}(\alpha)q_j^{-\alpha-1} \geq 0.$$

Therefore, Equation (E.3) can be solved with first-order condition:

$$\sum_{i\in[N]}\left(\frac{P_{ij}}{p_j}\right)^\alpha = \lambda \qquad \text{for all } j \in \mathcal{A}$$

which provides Equation (5.4). □