

Provable Pluralistic Alignment: Multi-Party RLHF under Offline Human Feedback

Huiying Zhong¹, Tianwei Gao², Zhiwei Steven Wu³, Linjun Zhang⁴, Weijie J. Su⁵, and Zhun Deng²

¹Massachusetts Institute of Technology

²University of North Carolina at Chapel Hill

³Carnegie Mellon University

⁴Rutgers University

⁵University of Pennsylvania

Abstract

Pluralistic alignment requires learning from feedback that reflects persistent and potentially conflicting stakeholder preferences while ultimately selecting a single collective policy. We study this problem in offline reinforcement learning from human feedback (RLHF), where the party associated with each comparison is observed. Under a shared low-rank linear reward model, we jointly estimate party-specific rewards and perform pessimistic policy optimization under Nash, Utilitarian, and Egalitarian social-welfare objectives. We establish nonasymptotic bounds for party-specific reward estimation and the resulting policy suboptimality under offline coverage conditions. We further consider general pairwise preferences that need not admit a scalar reward representation and may exhibit cycles. In this setting, we construct a pessimistic von Neumann winner policy and derive corresponding performance guarantees. Under these models, our results provide a unified finite-sample solution to a central challenge in pluralistic alignment: learning from limited, heterogeneous, and potentially cyclic feedback, and producing a single policy with explicit collective-welfare guarantees. Our framework thereby makes preference aggregation an explicit and statistically analyzable design choice rather than an implicit consequence of pooling human feedback.

1 Introduction

Pluralistic alignment requires learning from human feedback that may reflect persistent and potentially conflicting preferences while ultimately selecting a collective policy. Reinforcement Learning with Human Feedback (RLHF) has become a widely used approach for this broader alignment problem, particularly in language-model post-training and sequential decision making [Christiano et al., 2017; Ouyang et al., 2022; Ziegler et al., 2019]. Standard RLHF methods typically learn a single reward model from preference data, often collected as pairwise comparisons, and then optimize a policy against that reward. This pipeline implicitly assumes that feedback from different users can be represented by a common reward function.

This *single-party* assumption is poorly suited to settings in which users have heterogeneous and potentially conflicting viewpoints. In open-ended social decision-making, variation in feedback may reflect genuine disagreement rather than noise around one latent objective [Casper et al.,

2023; Santurkar et al., 2023; Siththaranjan et al., 2024]. Pooling the resulting comparisons therefore discards information about which parties hold which preferences. It also makes a consequential normative decision implicitly: the statistical procedure used to fit the pooled reward determines how conflicting preferences are combined.

A preliminary version of this manuscript was publicly circulated in March 2024 and was among the first theoretical studies to formulate heterogeneous human feedback as an explicit multi-party RLHF problem. Our work inspired subsequent research that has generalized, operationalized, or directly analyzed components of this formulation. Specifically, rather than treating disagreement as statistical noise, resolving it solely through personalization, or assuming that the relevant rewards are already known, we study how to learn unknown party-specific preferences from offline comparisons and use them to select one collective policy. This formulation makes both the statistical cost of preserving heterogeneity and the normative choice of aggregation rule explicit.

The problem lies at the intersection of three areas that are commonly treated separately: offline preference learning, shared-representation learning across related parties, and social-choice aggregation. Existing offline RLHF theory largely analyzes the estimation and optimization of a single reward function [Zhan et al., 2024; Zhu et al., 2023]. Personalized approaches learn different rewards or policies for different users [Park et al., 2024; Liu et al., 2025; Lee et al., 2026], whereas social-choice and policy-aggregation approaches often begin with preferences, rewards, or policies that are already specified [Ge et al., 2024; Alamdari et al., 2024]. Our setting is distinct: party identities are observed, party-specific preferences are unknown and must be learned from limited offline comparisons, and the final output must be one collective policy rather than a collection of personalized policies.

To see why this distinction matters, consider the following example.

Example (2 parties and 3 alternatives). Consider three alternatives A, B, C . Half of the population prefers $A > B > C$, with rewards

$$R_1(A) = 1, \quad R_1(B) = 1/2, \quad R_1(C) = 0,$$

while the other half prefers $C > B > A$, with rewards

$$R_2(C) = 1, \quad R_2(B) = 1/2, \quad R_2(A) = 0.$$

When the party labels are discarded and the two sub-populations are pooled, the population-level pairwise comparisons cannot distinguish among A, B, C . A single reward model fitted to the pooled comparisons therefore cannot identify B as the compromise. Finite-sample variation or arbitrary tie-breaking may instead select A or C , leaving one party with reward 0. By retaining the party labels, learning the two rewards separately, and then applying Nash or Egalitarian aggregation, the collective policy selects B and gives both parties reward 1/2. Thus, the difficulty is not only estimating preferences accurately; it is preserving the structure of disagreement until an explicit aggregation rule is chosen.

This example motivates the central question of the paper:

How can one learn a single collective policy from offline, party-indexed comparisons without first collapsing heterogeneous feedback into one implicit reward model?

Scope of pluralistic alignment. We use pluralistic alignment in a collective-decision sense. Our objective is not to require every model response to display multiple viewpoints, nor to produce a separate personalized policy for each user. Instead, we retain observed party-specific preferences during learning, quantify their statistical uncertainty separately, and select one collective policy through an explicitly specified social-welfare or pairwise social-choice rule. Our formulation therefore studies the statistical foundations of collective preference aggregation, complementing Overton, steerable, and distributional approaches that focus on the diversity, controllability, or population calibration of model outputs. We take the set of parties, their normative weights, and the admissibility of their preferences as externally specified.

Our approach and technical novelty. As one of the earliest theoretical frameworks for multi-party RLHF, our work connects offline preference learning, shared-representation learning, and social-welfare optimization within a single finite-sample framework. We explicitly separate two coupled problems. The first is statistical: learning party-specific preferences from limited and unevenly covered offline feedback while exploiting structure shared across parties. The second is normative: selecting a social-welfare rule that converts the learned preferences into a single collective policy. This separation makes preference aggregation an explicit design choice rather than an incidental consequence of pooling the data.

More specifically, to our knowledge, this work provides the first offline finite-sample analysis that jointly learns observed party-specific reward models through an unknown shared low-rank representation and performs pessimistic optimization of nonlinear social-welfare objectives. The novelty is therefore not merely the application of a welfare function after reward learning. The party rewards are themselves unknown and statistically coupled, and uncertainty from their estimation must be propagated through the nonlinear aggregation rule to the final policy. This creates a different statistical problem from both single-reward offline RLHF and social-welfare optimization with known rewards.

We first study a reward-representable regime. Each party follows a linear Bradley–Terry–Luce preference model, and the party-specific reward parameters share an unknown low-rank representation. This shared structure allows information from different parties to be used jointly without pooling away their distinct reward functions. We estimate the party rewards simultaneously, construct party-specific confidence sets, and optimize the resulting pessimistic social welfare. The same statistical framework supports Utilitarian, Nash, and Egalitarian objectives, thereby separating the choice of collective-welfare criterion from the statistical problem of reward estimation.

We then consider a substantially more general regime in which pairwise preferences need not admit any scalar reward representation and may be stochastic, nontransitive, or cyclic. In this setting, reward maximization is no longer an appropriate solution concept. We instead estimate party-indexed pairwise-preference matrices and construct a pessimistic von Neumann winner directly from the offline comparisons. The resulting framework therefore covers both utility-representable preferences and general pairwise preferences under a common principle: preserve heterogeneous feedback, quantify its statistical uncertainty, and apply an explicit collective-decision rule.

We formulate the main problem as a contextual bandit and extend the framework to trajectory-level preference feedback in Markov decision processes. Our results characterize how the

difficulty of collective alignment depends jointly on the shared reward structure, the number of parties, and the coverage of the offline comparison data.

Our main contributions are summarized as follows:

- **An end-to-end framework for pluralistic offline RLHF.** We formulate collective policy learning from party-indexed offline comparisons and explicitly separate preference learning from preference aggregation. Unlike personalization, our objective is a single collective policy; unlike conventional social-choice aggregation, the underlying party preferences are unknown and must be estimated from data.
- **Finite-sample theory for learning and aggregating heterogeneous rewards.** Under a shared low-rank linear reward model, we jointly estimate party-specific rewards, construct confidence sets, and optimize pessimistic Utilitarian, Nash, and Egalitarian welfare objectives. We establish nonasymptotic bounds for party-specific reward estimation and for the suboptimality of the resulting collective policy under realizability, diversity, and offline-coverage conditions. The analysis quantifies how preference-estimation uncertainty propagates through nonlinear social-welfare optimization.
- **Collective alignment beyond scalar reward models.** For general and potentially cyclic pairwise preferences, we construct a pessimistic von Neumann winner [Dudík et al., 2015; Fishburn, 1984] directly from offline party-indexed comparisons. We establish finite-sample performance guarantees and, under additional structural conditions, analyze ex-post efficiency. This extension addresses settings in which heterogeneous feedback cannot be reduced to the maximization of party-specific scalar rewards.

Historical positioning and subsequent developments. A preliminary version of this manuscript was publicly circulated in March 2024 and was among the first theoretical treatments of multi-party RLHF. At that time, only a small number of contemporaneous studies had begun to investigate heterogeneous preferences in RLHF [Chakraborty et al., 2024; Park et al., 2024]. Our broad historical claim is that this work was among the earliest contributions to the emerging theory of multi-party and pluralistic RLHF. Our more specific priority claim is that, to our knowledge, it was the first to provide finite-sample guarantees for the combination of shared low-rank party-specific reward learning, pessimistic nonlinear social-welfare optimization, and offline policy selection. It also introduced a multi-party offline von Neumann formulation for preferences that need not admit scalar reward representations.

Subsequent research has generalized, operationalized, or directly analyzed components of this formulation. Lee et al. [2026] extend the predefined-party model to contextual low-rank user heterogeneity; Xiong and Singh [2025] develop an efficient projection framework addressing the computational cost of nonlinear multi-objective and multi-group aggregation; and Kleine Buning et al. [2025] directly analyze the incentive properties of pluralistic RLHF procedures, including the pessimistic social-welfare rule studied here. Other work has developed axiomatic reward aggregation and policy-level social choice [Ge et al., 2024; Alamdari et al., 2024]. These developments demonstrate that our early formulation anticipated several statistical, computational, and strategic questions that have since become central to pluralistic alignment, while also clarifying the scope of our guarantees. In particular, our analysis assumes observed, persistent party identities and non-strategic feedback under the stated preference-model and offline-coverage conditions.

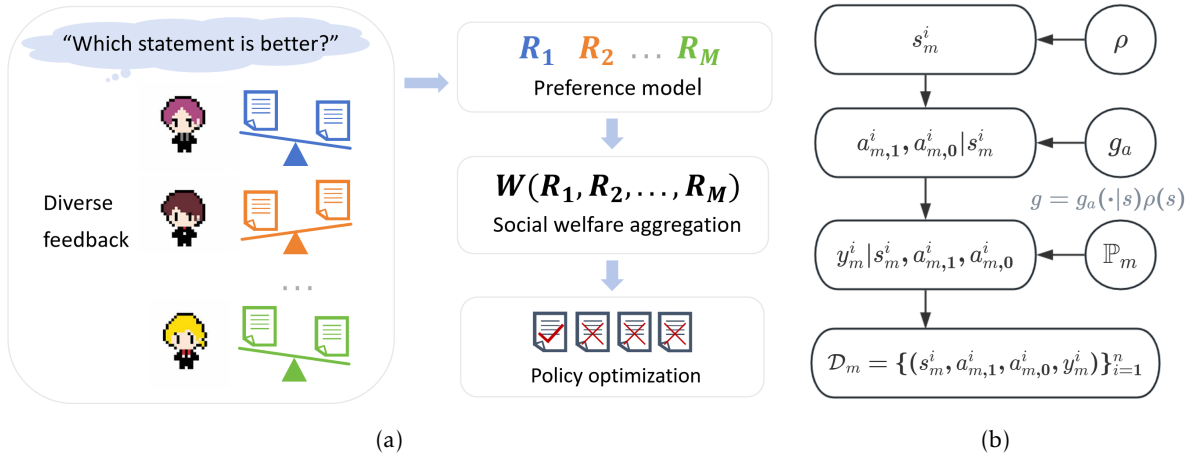


Figure 1: Multi-party RLHF: (a) party-specific preferences are learned and combined through an explicit social-welfare rule; (b) the offline feedback-collection process retains the party associated with each comparison.

1.1 Related Work

Offline RLHF theory. Preference comparisons are widely used in reinforcement learning because they are often easier for humans to provide than numerical rewards [Chen et al., 2022; Ouyang et al., 2022; Ziegler et al., 2019]. The theoretical foundations most closely related to our offline setting include Zhu et al. [2023], who study pessimistic offline RLHF under a linear Bradley–Terry–Luce (BTL) reward model, and Zhan et al. [2024], who extend the analysis to general function approximation. Wang et al. [2023] analyze reward-based and general pairwise-preference models in online single-party RLHF. These works primarily study feedback generated by a single latent preference model. In contrast, our analysis retains observed party identities, jointly estimates multiple statistically related reward models, and propagates their uncertainty through a collective social-welfare objective. The resulting problem combines preference estimation and preference aggregation rather than treating them as separate stages with a single reward.

Heterogeneous and pluralistic RLHF. A growing literature studies how RLHF systems should respond to heterogeneous human feedback. Empirical evidence demonstrates that human preferences can vary systematically across demographic, cultural, and contextual dimensions [Santurkar et al., 2023; Siththaranjan et al., 2024]. Bakker et al. [2022] use isoelastic social-welfare functions to fine-tune language models toward consensus among heterogeneous preferences. Chakraborty et al. [2024] learn a mixture of reward models and apply an Egalitarian objective to align a language model with diverse latent preferences. Park et al. [2024] study both personalization and collective aggregation, proposing representation-learning and clustering methods for personalized rewards as well as reward- and preference-level aggregation mechanisms.

Our work is distinguished by the specific combination of observed persistent parties, shared low-rank reward estimation, offline statistical uncertainty, nonlinear social-welfare optimization, and a single collective policy. Unlike personalization, we do not produce a different policy for each user. Unlike latent-mixture approaches, party membership is observed in the feedback

data and carries normative meaning in the subsequent aggregation step. More specifically, our contribution is an end-to-end finite-sample analysis of how uncertainty in jointly estimated party rewards propagates through pessimistic Utilitarian, Nash, and Egalitarian policy optimization.

Shared representation and personalized preference learning. Our reward-estimation method is related to multi-task and representation learning, where information from multiple related tasks is used to estimate a shared low-dimensional structure [Baxter, 2000; Maurer et al., 2016]. Finite-sample guarantees for shared-representation linear models have been developed by Du et al. [2021]; Tripuraneni et al. [2020]; Li and Zhang [2023], with related extensions to offline multi-task reinforcement learning [Bose et al., 2024]. In heterogeneous-feedback learning, Liu et al. [2025] use shared low-rank adaptation to learn personalized reward models, while Lee et al. [2026] model low-rank interactions between user contexts and query–response features and study distribution shift.

These approaches primarily target personalization or continuously varying user contexts. Our predefined parties can be viewed as a discrete and observed form of heterogeneity, but our decision problem is different: the objective is to learn one collective policy under a specified social-welfare rule. Accordingly, our analysis must control not only party-specific estimation error but also its effect on the nonlinear aggregation of party values.

Social choice, welfare aggregation, and multi-objective RL. Social choice theory studies how individual preferences or utilities should be combined into collective decisions. Relevant solution concepts include social-welfare functions [Moulin, 2004], Pareto efficiency [Aumann and Dombb, 2010; Barman et al., 2018], von Neumann winners or maximal lotteries [Brandl et al., 2016; Dudík et al., 2015; Rivest and Shen, 2010], and ex-post efficiency [Fishburn, 1984; Zeng and Psomas, 2020]. Fish et al. [2024] connect generative AI with approval-based social choice. Ge et al. [2024] study axioms for aggregating pairwise human feedback into a collective linear reward, showing that standard statistical reward-learning procedures can violate natural social-choice properties. Alamdari et al. [2024] instead aggregate already-defined policies through the geometry of state–action occupancy measures.

Related work in multi-agent and multi-objective RL studies collective policies when the relevant reward functions or objectives are already specified. Ogryczak et al. [2013]; Siddique et al. [2020] optimize generalized-Gini objectives, while Mandal and Gan [2022]; Ju et al. [2024] study broader families of social-welfare functions, including Utilitarian, Nash, and Egalitarian objectives, with online or offline guarantees. Multi-objective alignment methods likewise balance several specified objectives through reward, parameter, or policy combinations [Siddique et al., 2023; Wang et al., 2024; Zhou et al., 2023; Rame et al., 2023].

Our setting differs from these formulations because the party-specific rewards are not given: they must first be inferred from offline comparisons. Consequently, the confidence sets in our analysis capture uncertainty arising from preference learning, and the final policy must remain reliable after this uncertainty is propagated through the chosen social-welfare function. The novelty lies in coupling statistical reward learning with social-welfare optimization, rather than introducing either component in isolation.

Preferences beyond scalar reward models. Several recent approaches avoid assuming that pairwise preferences can be represented by a scalar reward. Munos et al. [2024] formulate

Nash Learning from Human Feedback as a two-player preference game, while [Swamy et al. \[2024\]](#) develop a self-play method based on a minimax winner that accommodates stochastic, nontransitive, and non-Markovian preferences. These methods demonstrate the importance of solution concepts defined directly from pairwise comparisons.

Our general pairwise-preference setting shares this motivation but addresses a different statistical problem. We observe offline, party-indexed comparisons and seek a single collective decision under uncertainty. We aggregate the party-specific pairwise preference matrices and construct a pessimistic von Neumann winner, obtaining finite-sample guarantees without requiring each party’s preferences to admit a scalar reward representation.

Subsequent extensions and limitations. Subsequent work has generalized or operationalized specific components of our formulation. [Lee et al. \[2026\]](#) extend predefined parties to contextual low-rank user heterogeneity. [Xiong and Singh \[2025\]](#) develop a computationally efficient projection framework for nonlinear multi-objective and multi-group aggregation, addressing the cost of repeatedly optimizing different nonlinear welfare functions. [Wang et al. \[2025\]](#) propose an efficient post-processing approach that combines policies already trained for different preferences. These methods complement our focus on reward estimation, offline uncertainty, and finite-sample welfare guarantees.

Other work clarifies limitations of the underlying modeling assumptions. [Kleine Buening et al. \[2025\]](#) show that pluralistic RLHF procedures, including pessimistic social-welfare optimization, need not be strategyproof and analyze the trade-off between truthful elicitation and collective welfare. [Xiao et al. \[2025\]](#) show that downstream KL-regularized policy optimization can produce preference collapse even when reward information is available. Active data-collection approaches select informative prompts or feedback providers rather than treating the offline dataset as fixed [[Liu et al., 2024](#)]. These developments clarify the scope of our results: we study the offline statistical problem with observed party identities and non-strategic feedback, while incentive compatibility, active elicitation, contextual personalization, and scalable policy optimization constitute complementary directions.

2 Preliminaries

Notation. Let $[n] = \{1, 2, \dots, n\}$. Let $a \vee b = \max\{a, b\}$. We use $\|\cdot\|_2$ or $\|\cdot\|$ to denote the ℓ_2 norm of a vector or the spectral norm of a matrix, and $\|x\|_\Sigma = \sqrt{x^\top \Sigma x}$ to denote the matrix-induced norm for a positive-semidefinite matrix Σ . We use $\mathcal{O}_{d_1 \times d_2}$ where $d_1 > d_2$ to denote the set of orthogonal matrices (i.e., the columns are orthonormal). Let $\langle \cdot, \cdot \rangle$ be the Euclidean inner product between vectors or matrices. For a matrix $A \in \mathbb{R}^{m \times n}$, let $\sigma_i(A)$ be its i -th largest singular value. Our use of $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ follows the standard notation.

2.1 Offline Data Collection and Comparison Model

We ground our theoretical investigation of RLHF in the contextual bandit (CB) framework [[Xiong et al., 2023](#); [Zhu et al., 2023](#)]. This formulation naturally models the reinforcement learning dynamics of LLMs, where the generation of a completion $a \in \mathcal{A}$ given a prompt $s \in \mathcal{S}$ is evaluated by a reward model $R_\theta(s, a)$ [[Chakraborty et al., 2024](#); [Ouyang et al., 2022](#); [Xiong et al., 2023](#); [Zhu et al., 2023](#); [Ziegler et al., 2019](#)]. By operating within the CB setting, which is a

foundational case of MDPs with a horizon $H = 1$ and no transitions, we can isolate and rigorously analyze the fundamental complexities of multi-party preference aggregation. We subsequently extend our theoretical framework to general MDPs, demonstrating that these insights naturally generalize to sequential settings. Formally, we define a multi-party CB environment by the tuple $E = (\mathcal{S}, \mathcal{A}, \{R_m\}_{m=1}^M, \rho)$, where $\mathcal{S}$ is the state space (e.g., prompts), $\mathcal{A}$ is the action space (e.g., responses), M is the total number of parties, $R_m : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ denotes the reward function specific to the m -th party, and $\rho \in \Delta(\mathcal{S})$ specifies the initial state distribution. A policy maps a state to an action, taking the form of either a deterministic mapping $\pi : \mathcal{S} \rightarrow \mathcal{A}$ or a randomized distribution $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. Although the reward functions are indexed by party, the learner’s output is a single collective policy π aligned with respect to all parties’ preference, rather than a collection of personalized policies $\{\pi_m\}_{m=1}^M$.

We focus on the offline learning paradigm, which is an approach extensively explored in previous research on offline RLHF [Zhu et al., 2023; Zhan et al., 2024], driven by the practical constraints of LLM alignment: Collecting high-quality human feedback is resource-intensive and time-consuming, which typically precludes online, interactive human queries during the training phase. Consequently, models have to be aligned using pre-collected datasets. We assume access to an offline dataset $\mathcal{D} = \bigcup_{m=1}^M \mathcal{D}_m$, where $\mathcal{D}_m = \{(s_m^i, a_{m,1}^i, a_{m,0}^i, y_m^i)\}_{i=1}^n$ represents the pairwise comparison data specific to the m -th party. Specifically, the state s_m^i represents a prompt and is drawn from the initial distribution ρ , the action pair $(a_{m,1}^i, a_{m,0}^i)$ represents two candidate responses, and the outcome $y_m^i \in \{0, 1\}$ is the party’s binary preference label provided by an annotator. As illustrated in Figure 1b, the data generation process follows a structured sequence. In the language-model setting, the index m identifies a *party*: a persistent preference source whose preferences are observed across prompts. A party may be an individual or a stakeholder group represented by multiple feedback providers. For each party m , the dataset $\mathcal{D}_m$ contains comparison observations indexed by i . Specifically, each time a prompt s_m^i is drawn from the prompt distribution ρ , and two candidate completions, $a_{m,1}^i$ and $a_{m,0}^i$, are generated for that prompt. An *annotator*, or *feedback provider*, supplies comparison label y_m^i recording which completion is preferred by this party. When a party represents a stakeholder group, multiple feedback providers within this group may contribute observations to the same party-specific dataset $\mathcal{D}_m$.

2.2 Social Welfare Function

Having specified how party-indexed preference comparison data are collected, we next describe how the learned party preferences are used to select a collective policy. In the reward-representable setting, each party is associated with a reward function, and a policy may provide different values to different parties. Social welfare functions provide a systematic way to aggregate these party-specific values into a single objective for collective policy selection. Specifically, for a policy π , let $V_m(\pi) = \mathbb{E}_{s \sim \rho, a \sim \pi(\cdot|s)} [R_m(s, a)]$ denote its expected value to party m . Each policy therefore induces a vector of party values

$$V(\pi) = (V_1(\pi), \dots, V_M(\pi)).$$

A social welfare function maps this vector to a scalar measure of collective welfare. Different welfare functions express different approaches to balancing the interests of the parties and may therefore lead to different collective policies. In our framework, the offline comparison dataset

is used to learn the party-specific reward and thus estimate the expected values for each party, and the selected welfare function determines how those values are aggregated. This structure separates the statistical task of preference learning from the normative choice of collective aggregation.

We consider the isoelastic family of social welfare functions [Moulin, 2004; Sen, 2018]. For a positive value vector $v = (v_1, \dots, v_M) \in \mathbb{R}_+^M$, this family is defined as

$$W_\alpha(v) = \begin{cases} \left(\frac{1}{M} \sum_{m=1}^M v_m^\alpha \right)^{1/\alpha}, & \alpha \leq 1, \quad \alpha \neq 0, \\ \left(\prod_{m=1}^M v_m \right)^{1/M}, & \alpha = 0. \end{cases} \quad (1)$$

Its axiomatic characterization and additional properties are reviewed in Appendix A.1. In this article, we primarily focus on three representative members of this family: Utilitarian, Egalitarian, and Nash welfare.

Utilitarian welfare. The Utilitarian objective corresponds to $\alpha = 1$ and maximizes the average value across parties:

$$J_U(\pi) = \frac{1}{M} \sum_{m=1}^M V_m(\pi).$$

Under this objective, an improvement to any party contributes equally to collective welfare. The Utilitarian objective therefore emphasizes aggregate party value, although a high average value may coexist with a low value for one or more parties.

Egalitarian welfare. The Egalitarian, or max-min, objective is obtained in the limit as $\alpha \rightarrow -\infty$:

$$J_E(\pi) = \min_{m \in [M]} V_m(\pi).$$

This objective evaluates a policy according to the value it provides to the worst-off party. Maximizing it therefore prioritizes improving the lowest party value.

Nash welfare. The Nash objective corresponds to $\alpha = 0$ and maximizes the geometric mean of the party values:

$$J_N(\pi) = \left(\prod_{m=1}^M V_m(\pi) \right)^{1/M}.$$

The Nash objective rewards increases in aggregate welfare while remaining sensitive to parties with relatively low values. It therefore balances aggregate welfare and inequality across parties rather than focusing exclusively on either the average value or the worst-off party [Kelly et al., 1998; Zhang et al., 2022].

These objectives assign equal normative weight to the M parties. This equal weighting is a modeling and normative choice. Other externally specified party weights are possible, but we do not analyze them here.

In Section 3, we use Nash welfare as the primary example and subsequently apply the same reward-estimation and pessimistic-optimization framework to the Utilitarian and Egalitarian objectives in Section 3.4. In the general pairwise-preference setting considered in Section 4, scalar party values may not exist, so we instead define the collective solution directly from the party-specific pairwise-preference matrices.

3 Reward-Representable Regime

Having introduced the collective welfare objectives, we now study how the party-specific rewards entering these objectives can be learned from offline preference comparisons. We instantiate the social welfare perspective from Section 2 in a reward-representable regime, where each party’s preferences are represented by a scalar reward function. This reward-based formulation is the standard modeling approach in RLHF for LLMs [Ouyang et al., 2022; Ziegler et al., 2019; Christiano et al., 2017]. In this regime, collective alignment requires learning the party-specific rewards and using them to select a single collective policy according to an explicit social-welfare objective.

Reward-based Model We begin by specifying the reward-based model and how each party’s latent reward function determines the pairwise comparison data observed by the learner. Suppose that the preference distribution of party m is induced by a reward function:

$$\mathbb{P}_m(y = 1 | a_1, a_0, s) = \Phi(R_m(s, a_1) - R_m(s, a_0)), \quad \forall m \in [M], \quad (2)$$

where $y = 1$ indicates that action a_1 is preferred to action a_0 at state s , and Φ is the sigmoid function $\Phi(x) = 1/(1 + \exp(-x))$. This gives the well-known Bradley–Terry–Luce (BTL) pairwise comparison model [Bradley and Terry, 1952], which is widely used in RLHF frameworks [Christiano et al., 2017; Ouyang et al., 2022].

Fix a constant $B > 0$ and define $\Theta_B := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}$. Following Zhu et al. [2023], we assume that the underlying reward R_m for each party belongs to the family of linear functions $R_\theta(s, a) = \theta^\top \phi(s, a) + C_0$, where $\phi(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$ is a known feature mapping with $\max_{s,a} \|\phi(s, a)\|_2 \leq L$ and $C_0 > BL$ is a fixed positive constant. Specifically, we assume the true reward of party m is $R_m = R_{\theta_m^*}$, where $\theta_m^* \in \Theta_B$ denotes the ground truth parameter for the m -th party reward function. The constant C_0 cancels from every BTL comparison and therefore does not affect the induced preference probabilities. Its role is to specify the fixed reward normalization used by the welfare objective. In particular, for every $\theta \in \Theta_B$ and every policy π ,

$$\mathbb{E}_{s \sim \rho, a \sim \pi(s)}[R_\theta(s, a)] \geq C_0 - BL > 0.$$

Thus, all policy value functions appearing in welfare objective are uniformly positive. We remark that different choices of C_0 may induce different welfare-maximizing policies, and all guarantees below are conditional on this fixed normalization. In LLM applications, $\phi(s, a)$ may be instantiated as the representation of a prompt-response pair produced by a pretrained model before its final scalar output layer. The parameter θ_m^* then serves as the m -th party linear reward head applied to this shared representation.

We further assume that the party-specific reward parameters share a common low-rank structure:

$$\theta_m^* = U^* \alpha_m^*,$$

where $U^* \in \mathcal{O}_{d \times r}$, $d \gg r$, has orthonormal columns spanning a shared r -dimensional subspace of the original parameter space, and $\alpha_m^* \in \mathbb{R}^r$ is the party-specific coefficient vector satisfying $\|\alpha_m^*\|_2 = \Theta(1)$. Equivalently,

$$R_m(s, a) = \alpha_m^{*\top} U^{*\top} \phi(s, a) + C_0.$$

Thus, $U^{*\top} \phi(s, a)$ provides a shared low-dimensional representation of the state-action features, while α_m^* captures how party m weights the resulting reward-relevant directions. This shared structure allows information from different parties to be used jointly without pooling their distinct reward functions into a single reward model. This low-rank structure is a standard assumption in shared-representation learning and underlies the estimation error reduction established later in this section [Du et al., 2021; Tripuraneni et al., 2020].

Under this model, our framework proceeds in two stages. First, we jointly estimate the shared representation and the party-specific reward parameters from the offline comparisons and construct party-specific confidence sets that quantify their statistical uncertainty. Second, we select a single collective policy by optimizing a pessimistic Nash welfare objective over these confidence sets. The remainder of this section analyzes the resulting estimation error, policy suboptimality, and welfare guarantees before extending the framework to other social-welfare objectives and trajectory-level preference feedback in MDPs. Throughout this section, we restrict attention to the deterministic policy class. Accordingly, all optimality, suboptimality, and efficiency guarantees in this section are relative to the best deterministic policy.

3.1 Pessimistic Nash Welfare Algorithm

Building on the reward-based model introduced above, we now present a two-step procedure for learning a deterministic collective policy from offline party-indexed comparisons. In the first step, we jointly estimate the shared representation and the party-specific reward parameters and construct confidence sets that quantify their statistical uncertainty. In the second step, we optimize a pessimistic Nash welfare objective over these confidence sets to obtain the final deterministic policy. Specifically, the pessimistic objective evaluates each candidate policy under the least favorable combination of party-specific reward parameters that remains plausible given the observed comparisons. This construction guards policy selection against estimation error arising from limited offline coverage. The complete procedure is summarized in Algorithm 1.

Algorithm 1 Pessimistic Nash Welfare

Input: The datasets $\mathcal{D} = \bigcup_{m=1}^M \mathcal{D}_m$, a failure probability δ , and the initial distribution ρ .

Output: The deterministic collective policy $\hat{\pi}$.

- 1: Jointly compute the MLE $(\hat{U}, \hat{\alpha})$ in Equation 3 and set $\hat{\theta}_m = \hat{U} \hat{\alpha}_m$ for every $m \in [M]$.
 - 2: Construct the party-specific confidence sets $\Theta_B(\hat{\theta}_m) = \{\theta \in \Theta_B : \|\theta - \hat{\theta}_m\|_{\Sigma_m} \leq \Gamma(\delta, n, M, d, r)\}$ according to the bound Γ on the estimation error established in Theorem 3.3.
 - 3: Establish the pessimistic Nash welfare function in (4).
 - 4: Calculate $\hat{\pi} = \arg \max_{\pi} \hat{J}(\pi)$.
-

Step 1: MLE and confidence set construction. The objective of this first step is to estimate the party-specific reward functions from the observed comparisons while quantifying their statistical uncertainty. We jointly estimate the shared representation and the party-specific coefficients using all party-indexed datasets. To simplify the notation, denote $\alpha := (\alpha_1, \dots, \alpha_M) \in \mathbb{R}^{r \times M}$ as the low-dimensional parameter matrix, and $X_{m,i} = \phi(s_m^i, a_{m,1}^i) - \phi(s_m^i, a_{m,0}^i)$ as the feature

difference vectors for $\forall i \in [n], m \in [M]$. Using the comparison dataset $\mathcal{D} = \bigcup_{m=1}^M \mathcal{D}_m$, we minimize the negative log-likelihood based on (2), which is given by

$$\widehat{U}, \widehat{\alpha} \in \arg \min_{U \in \mathcal{O}_{d \times r}, \|\alpha_m\|_2 \leq B} l_{\mathcal{D}}(U, \alpha), \quad (3)$$

$$l_{\mathcal{D}}(U, \alpha) = -\frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \left\{ \mathbf{1}(y_m^i = 1) \log(\Phi(\langle U \alpha_m, X_{m,i} \rangle)) + \mathbf{1}(y_m^i = 0) \log(\Phi(\langle U \alpha_m, -X_{m,i} \rangle)) \right\}.$$

When the minimizer is not unique, we take any of the solutions $(\widehat{U}, \widehat{\alpha})$ that achieves the minimum. The resulting estimator of party m 's reward parameter is $\widehat{\theta}_m = \widehat{U} \widehat{\alpha}_m$. Joint estimation allows comparisons from all parties to inform the shared representation $\widehat{U}$, while retaining a distinct coefficient vector $\widehat{\alpha}_m$ for each party.

We next construct a confidence set around each estimated reward parameter $\widehat{\theta}_m$ to quantify uncertainty about the corresponding true parameter θ_m^* . Specifically, we define:

$$\Theta_B(\widehat{\theta}_m) = \{\theta \in \Theta_B : \|\theta - \widehat{\theta}_m\|_{\Sigma_m} \leq \Gamma(\delta, n, M, d, r)\},$$

where Γ is a confidence radius defined in the next subsection, δ is a prescribed failure probability, and $\Sigma_m \in \mathbb{R}^{d \times d}$ is the empirical covariance matrix for the m -th party, defined as $\Sigma_m := \sum_{i=1}^n X_{m,i} X_{m,i}^\top / n$. For simplicity, we define $\Theta = (\theta_1, \dots, \theta_M)$, $\widehat{\Theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_M)$, and use $\Theta \in \Theta_B(\widehat{\Theta})$ to denote the event of $\theta_m \in \Theta_B(\widehat{\theta}_m)$ for any $m \in [M]$.

Step 2: Policy optimization under Nash welfare. Building on the confidence sets constructed in Step 1, we now select a deterministic collective policy by optimizing the Nash welfare objective while accounting for uncertainty in the learned rewards. For a candidate parameter θ_m , party m 's value under policy π is $\mathbb{E}_{s \sim \rho} [R_{\theta_m}(s, \pi(s))]$. We aggregate these M party-specific values using the Nash welfare objective introduced in Section 2.2.

Specifically, we account for reward-estimation uncertainty by defining the following pessimistic Nash welfare objective:

$$\widehat{J}(\pi) = \min_{\Theta \in \Theta_B(\widehat{\Theta})} \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{\theta_m}(s, \pi(s))]^{1/M}. \quad (4)$$

For each candidate policy, the inner minimization evaluates its Nash welfare under the least favorable combination of party-specific reward parameters contained in the confidence sets. It therefore prevents the policy optimizer from relying on optimistic reward estimates that are weakly supported by the offline comparisons. In particular, policies whose values depend strongly on poorly covered feature directions can receive a lower pessimistic welfare value. We obtain the final deterministic collective policy by maximizing this pessimistic objective:

$$\widehat{\pi} = \arg \max_{\pi} \widehat{J}(\pi).$$

Here, pessimism specifically addresses statistical uncertainty arising from finite and imperfectly covered offline data [Li et al., 2022; Xie et al., 2021; Zhan et al., 2024; Zhu et al., 2023].

3.2 Sub-optimality

We now establish finite-sample guarantees for the deterministic collective policy returned by Algorithm 1. Our analysis proceeds in three parts. First, we bound the error of the joint shared-representation estimator for each party’s reward parameter. Second, we propagate this estimation error through the pessimistic Nash welfare objective to obtain a policy-suboptimality bound. Finally, we provide a minimax lower bound that identifies an unavoidable dependence on the representation dimension and clarifies the remaining gap in our upper bound. Together, these results characterize how shared reward structure and offline data coverage affect the quality of the learned collective policy. Formally, we define the true Nash welfare function J and the sub-optimality of a deterministic policy π as follows:

$$J(\pi) := \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi(s))]^{1/M}, \quad \text{SubOpt}(\pi) := J(\pi^*) - J(\pi),$$

where $\pi^* = \arg \max_{\pi} J(\pi)$ is the optimal deterministic policy. Here, $J(\pi)$ is the true Nash welfare evaluated using the ground-truth party rewards, and π^* is an optimal deterministic policy under this objective. The suboptimality measures the welfare loss of a candidate policy relative to this benchmark. Our goal is to show that the pessimistic policy $\hat{\pi}$ has small suboptimality when the shared reward structure is identifiable and the offline comparisons provide sufficient coverage. We introduce several assumptions first.

Assumption 3.1 (Diversity design). Let $\Theta^* = (\theta_1^*, \dots, \theta_M^*) \in \mathbb{R}^{d \times M}$ be the original parameter matrix, $\nu = \sigma_r(\frac{\Theta^{*\top} \Theta^*}{M}) = \sigma_r(\frac{\alpha^{*\top} \alpha^*}{M})$ be the smallest singular value, and $\kappa = \sigma_1(\frac{\Theta^{*\top} \Theta^*}{M})/\nu = \sigma_1(\frac{\alpha^{*\top} \alpha^*}{M})/\nu$ be the condition number. We assume that $\nu > 0$ and $\kappa \leq O(1)$.

Assumption 3.2 (Feature design). Let $\Sigma_* = \mathbb{E}_g(\phi(s, a_1) - \phi(s, a_0))(\phi(s, a_1) - \phi(s, a_0))^\top$ be the covariance matrix, where the randomness is induced by the generation distribution g for the triplet (s, a_0, a_1) . We assume that Σ_* has bounded positive eigenvalues, i.e., $C_{\max} \geq \lambda_1 \geq \dots \geq \lambda_d \geq C_{\min} > 0$.

Assumption 3.1 is a standard diversity condition on the underlying reward parameters $\{\theta_m^*\}_{m=1}^M$ in shared-representation learning and multi-task learning; see Assumption 2 in [Tripuraneni et al. \[2020\]](#) and Assumption 4.3 in [Du et al. \[2021\]](#). In our setting, it requires the coefficient vectors $\{\alpha_m^*\}_{m=1}^M$ to collectively span the r -dimensional shared subspace and to do so in a well-conditioned manner. This does not require every pair of parties to have substantially different preferences. Rather, it requires sufficient variation across the collection of parties to identify all directions of the shared representation. Assumption 3.2 is a well-conditioned design condition on the feature-difference covariance matrix Σ_* . Since the BTL model depends on the reward parameters through the feature differences $\phi(s, a_1) - \phi(s, a_0)$, this assumption ensures that these feature differences contain enough information to distinguish different reward parameters. Our results can also be modified to accommodate the case where Σ_* is not invertible by replacing it with $\Sigma_* + \lambda I$ for any regularization parameter $\lambda \in \mathbb{R}^+$; see Remark 3.4 of [Zhu et al. \[2023\]](#).

Before moving on to sub-optimality, we first quantify the estimation error of the jointly learned party-specific reward parameters.

Theorem 3.3. *Suppose Assumptions 3.1, 3.2 hold. If the number of samples satisfies $n \gg d + \log(M/\delta)$*

for $\delta \in (0, 1)$, then with probability at least $1 - \delta$, for any $m \in [M]$,

$$\|\widehat{\theta}_m - \theta_m^*\|_{\Sigma_m} \lesssim \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)}.$$

Proof Sketch. We provide a proof outline here, and defer the details to Appendix B.1. First, we establish an overall bound on $\sum_{m=1}^M \|X_m \widehat{\theta}_m - X_m \theta_m^*\|^2$ by leveraging the strong convexity of $l_{\mathcal{D}}(U, \alpha)$, based on second-order Taylor expansion and a careful use of concentration inequalities. Next, we utilize Assumption 3.1 to derive a bound on $\|\widehat{\theta}_m - \theta_m^*\|_{\Sigma_m}$ in terms of U and α_m using the Davis-Kahan sin θ theorem and Bernstein-type inequality. ■

Theorem 3.3 implies that with probability at least $1 - \delta$, $\theta_m^* \in \Theta_B(\widehat{\theta}_m)$ for $\forall m \in [M]$. Then we specify $\Gamma(\delta, n, M, d, r) = K \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)}$ in Algorithm 1 with K a constant, thus completing the pessimistic component. It is a generalization of the upper bound of Lemma 3.1 in Zhu et al. [2023], which improves the previous bound of $O(\frac{d}{n})$ to $O(\frac{dr^2}{Mn})$ up to logarithmic factors when $r^2 \ll \min\{M, d\}$. We will see that the advancement also improves subsequent sub-optimality bounds.

Theorem 3.3 controls reward-estimation error in the covariance norm induced by the offline comparisons. To derive a policy-suboptimality guarantee, we need to convert this parameter estimation error into an error in the expected reward of the optimal policy. We capture this relationship through the following concentratability coefficient:

$$C^* = \max_m \|\Sigma_m^{-1/2} \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))\|_2. \quad (5)$$

We have the following sub-optimality guarantee for the pessimistic policy.

Theorem 3.4. *Under the same condition as Theorem 3.3, with probability at least $1 - \delta$,*

$$\text{SubOpt}(\widehat{\pi}) \lesssim C^* \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)}.$$

Proof Sketch. We provide a proof outline here, and defer the details to Appendix B.2. The sub-optimality can be decomposed as, $J(\pi^*) - J(\widehat{\pi}) = (J(\pi^*) - \widehat{J}(\pi^*)) + (\widehat{J}(\pi^*) - \widehat{J}(\widehat{\pi})) + (\widehat{J}(\widehat{\pi}) - J(\widehat{\pi}))$ and what we truly need to focus on is the first term $J(\pi^*) - \widehat{J}(\pi^*)$ since the latter two terms are smaller than zero with high probability. We then address the geometric form of the Nash welfare function by decomposing it into the aggregation of individual errors. ■

As claimed before, the sample complexity reduction described in Theorem 3.3 extends to Theorem 3.4, highlighting the advantage of low rank shared-representation learning. In our multi-party setting, sub-optimality depends on a concentratability coefficient C^* across the aggregation of M parties, setting it apart from single-party RLHF. Notably, C^* is assumed to be bounded [Zhan et al., 2024; Zhu et al., 2023], which provides a coverage guarantee for the feature vectors $\phi(s, \pi^*(s))$ from the dataset $\mathcal{D}$, thus yielding an accurate estimator.

To complement the upper bound, we proceed to investigate the lower bound in the following. We show that, for the problems with bounded concentratability coefficients C^* as defined in (5), we obtain the following result, whose proof is deferred to Appendix B.3.

Theorem 3.5 (Lower Bound). *Consider the family of instances $CB(M, \mathcal{C}) = \{\rho, \{s_m^i, a_{m,1}^i, a_{m,0}^i\}_{i \in [n], m \in [M]}, \Theta^* = U^* \alpha^* | C^* \leq C\}$, where C^* is defined in (5). For any bandit instance $Q(M) \in CB(M, \mathcal{C})$, let $\text{SubOpt}_{Q(M)}$ be the sub-optimality under instance $Q(M)$. Suppose $r > 6, n \gtrsim rC^2, C \geq 2$, then there exists a feature mapping ϕ such that the following lower bound holds.*

$$\inf_{\widehat{\pi}} \sup_{Q(M) \in CB(M, \mathcal{C})} \text{SubOpt}_{Q(M)}(\widehat{\pi}) \gtrsim C \sqrt{\frac{r}{n}}.$$

The lower bound, together with the previous upper bound, highlights the complexity challenges in multi-party RLHF. In this setting, we encounter separate samples for each party and a potentially larger concentratability coefficient C^* defined in (5), in contrast to the concentratability coefficient presented in [Zhu et al. \[2023\]](#). Consequently, our shared-representation learning technique plays an important role in achieving an improved sub-optimality bound. Obtaining tight dependence of our sub-optimality bound on feature dimensions is challenging. The $\sqrt{r}$ gap between Theorem 3.4 and Theorem 3.5 relates to an open problem in classical estimations, see a more detailed discussion on page 8 in [Tripuraneni et al. \[2020\]](#).

3.3 Efficiency and Welfare Guarantees

Beyond the optimality result, in this section, we demonstrate the efficiency guarantees of our method, as our proposed approach addresses social choice considerations. Section 3.2 establishes that the deterministic policy $\widehat{\pi}$ achieves Nash welfare close to that of the optimal policy π^* . Because the Nash welfare objective is strictly increasing in every positive party value, an exact Nash-welfare maximizer is Pareto efficient. This observation motivates us to investigate the concept of τ -approximate Pareto efficiency. We now examine the approximate efficiency guarantee inherited by the learned policy when its Nash-welfare suboptimality is small.

Definition 3.6 (τ -approximate Pareto Efficiency). A deterministic policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is said to be τ -approximate Pareto efficient, if there does not exist another deterministic policy π' such that for all party $m \in [M]$, $\mathbb{E}_{s \sim \rho} R_{\theta_m^*}(s, \pi'(s)) \geq \mathbb{E}_{s \sim \rho} R_{\theta_m^*}(s, \pi(s))$, and for at least one party $m \in [M]$, $\mathbb{E}_{s \sim \rho} R_{\theta_m^*}(s, \pi'(s)) \geq (1 + \tau) \mathbb{E}_{s \sim \rho} R_{\theta_m^*}(s, \pi(s))$.

This notion of efficiency ensures an agreement and balance among multiple parties, contributing to overall societal wellness by optimizing the collective outcomes. It's an adaptation of the definitions by [Aumann and Dombb \[2010\]](#); [Barman et al. \[2018\]](#) to RLHF settings. While previous work considered an efficiency notion that allows a τ improvement for all parties, we admit a τ improvement for only one party when maintaining others unchanged. We now show that the Nash-welfare suboptimality bound from Theorem 3.4 implies approximate Pareto efficiency of the learned deterministic policy.

Theorem 3.7. *Under the same condition as Theorem 3.3, with probability at least $1 - \delta$, $\widehat{\pi}$ is τ -approximate Pareto efficient, where $\tau \lesssim C^* M \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)}$.*

This result confirms that we cannot improve one party significantly without making another party worse, thus leading to an approximately efficient policy. Besides, the pessimistic policy $\widehat{\pi}$ follows the approximate Pigou-Dalton principle, which emphasizes the transfer of welfare from better-off individuals to weaker ones, ensuring a more equitable distribution of welfare

outcomes [Moulin, 2004]. The definition proof of the efficiency and welfare guarantees are given in Appendix B.4

3.4 Comparison with Various Social Welfare Functions

The preceding subsections focus on Nash welfare as the primary aggregation rule, since it balances overall welfare with protection against severe imbalance across parties. As outlined in Section 2.2, we provide a brief introduction to other social welfare functions and compare their results. Once the confidence sets for party-specific rewards have been constructed, we can replace the Nash aggregation in Line 3 of Algorithm 1 with another social welfare objective and optimize the corresponding pessimistic empirical value. To illustrate this flexibility, we consider two canonical alternatives introduced in Section 2: the Utilitarian objective, which optimizes average welfare, and the Egalitarian objective, which optimizes the welfare of the worst-off party. As shown below, both variants inherit the same statistical rate, so the choice of welfare function mainly determines the social criterion being optimized rather than the reward-learning difficulty.

Utilitarian ($\alpha = 1$) Under the Utilitarian paradigm, the goal is to maximize the arithmetic mean of the expected returns across all parties. Similarly, the (pessimistic) Utilitarian welfare function is defined as follows,

$$J(\pi) := \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi(s)),$$

$$\widehat{J}(\pi) = \min_{\Theta \in \Theta_B(\widehat{\Theta})} \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi(s)).$$

Egalitarian ($\alpha \rightarrow -\infty$) The Egalitarian approach prioritizes the worst-off party by maximizing the minimum expected return among all parties, denoted as $\min_m \mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi(s))$. Similarly, the (pessimistic) Egalitarian welfare function is defined as follows,

$$J(\pi) := \min_m \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi(s)),$$

$$\widehat{J}(\pi) = \min_{\Theta \in \Theta_B(\widehat{\Theta})} \min_m \mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi(s)).$$

For both cases, define $\widehat{\pi} = \arg \max_{\pi} \widehat{J}(\pi)$, $\pi^* = \arg \max_{\pi} J(\pi)$. We have the following sub-optimality bound.

Theorem 3.8. *Under the same condition as Theorem 3.3, with probability at least $1 - \delta$,*

$$\text{SubOpt}(\widehat{\pi}) \lesssim C^* \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)}.$$

Here $C^* = \max_m \|(\Sigma_m^{-1/2} \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s)))\|_2$.

Theorem 3.8 implies that $\widehat{\pi}$ is approximately optimal for the average/worst-off party, thus characterizing a relatively fair outcome. Each welfare function measures different aspects of social welfare, making them suitable for different practical settings.

Generalization to MDP As claimed earlier, the contextual bandit (CB) can be viewed as a special case of a Markov decision process (MDP) with finite horizons and transitions over states. For multi-party RLHF under MDP settings, preferences are based on the whole trajectories rather than single actions [Zhan et al., 2024]. The generalized algorithm for MDPs is similar to Algorithm 1, and sub-optimality upper bound is also guaranteed. However, in this context, we adopt a trajectory-based preference model and optimize a trajectory-wise value function. The key difference between MDPs and CBs is the inclusion of a trajectory-based occupancy measure $d^\pi(s) = \sum_{h=1}^H \mathbb{P}_h(s_h = s | \pi)$, which reduces to ρ for CBs. To proceed with the result, we adopt the observation from [Zhu et al., 2023] that:

$$\mathbb{E}_{s \sim \rho}[V_m^\pi(s)] = \mathbb{E}_{s \sim d^\pi} R_{m, \theta_m^*}(s, \pi(s)),$$

where $V_m^\pi(s) = \mathbb{E}[\sum_{h=1}^H R_m(s_h, a_h) | s_1 = s, \pi]$, and this expectation $\mathbb{E}$ is taken over the trajectory generated according to the transition kernel. The details are provided in Appendix A.3.

4 Extension to General Pairwise-Preference Model

The reward-representable regime in Section 3 represents each party’s preferences through a scalar reward function, which is natural for RLHF but also imposes structural restrictions on the preference relation. In particular, scalar reward models are most suitable when preferences can be summarized by an underlying ranking of actions or trajectories. Human preferences, however, may exhibit more complex patterns, including intransitivities observed in psychological experiments [Tversky, 1969]. To capture such cases, we now extend our framework to a more general regime, where preferences are represented directly through pairwise comparison preference probabilities rather than through latent reward functions. We retain the central structure of the preceding sections, where we learn a single collective policy from offline dataset collected from all parties. Because scalar expected preference value for each party is no longer available for social-welfare aggregation, we instead average the party-specific pairwise preference margins for each state-action pair tuple (s, a_0, a_1) and select a pessimistic von Neumann winner, which maximizes the worst-case aggregate preference margin against competing strategies. Throughout this section, we focus on the tabular setting with finite state and action spaces, where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. This allows us to avoid additional function-approximation assumptions and to work directly with state-wise preference matrices. It also leads naturally to finite zero-sum matrix games, which provide the algorithmic basis for computing von Neumann winners.

General Pairwise-Preference Model In the general pairwise-preference setting, we no longer assume that pairwise preferences are generated by reward differences. Instead, for each party m and state s , we represent preferences directly through the skew-symmetric matrix. Specifically, To ensure symmetry, we define the following preference matrix [Dudík et al., 2015]:

$$N_{s,m}^*(a_1, a_0) = 2\mathbb{P}_m(y = 1 | a_1, a_0, s) - 1, \quad \forall m \in [M], s \in \mathcal{S}.$$

Thus, $N_{s,m}^*$ is a skew-symmetric matrix. In this general model, we refrain from directly specifying reward functions to model human preferences. Consequently, it’s unfeasible to obtain an optimal policy based on an aggregated reward function. We consider an alternative solution concept, the *von Neumann winner* [Dudík et al., 2015], which corresponds to a probabilistic strategy that

outperforms every other strategy with at least a 50% probability of being preferred. Formally, a von Neumann winner of an $A \times A$ skew-symmetric preference matrix N is the max-min probabilistic strategy p^* , defined as follows, with matrix N describing a matrix game:

$$p^*, q^* = \arg \max_p \min_q p^\top N q = \arg \max_p \min_q \sum_{i,j} p_i q_j N_{ij}.$$

The value of the matrix game is defined as $p^{*\top} N q^*$, which equals 0 when N is skew-symmetric [Owen, 2013]. To connect with the previous model, if (2) holds, the von Neumann winner described in the following section corresponds to the Utilitarian welfare solution when $M = 1$ or $M = 2$. See Appendix A.2 for details.

4.1 Pessimistic Von Neumann Winner Algorithm

We now construct a pessimistic von Neumann winner from the offline party-indexed comparisons dataset. The procedure has two stages. First, we estimate the party-specific preference matrices, aggregate them using equal party weights, and construct entrywise confidence bounds. Second, for each state, we solve a zero-sum matrix game using the pessimistic payoff matrix and use its max-min strategy to define a randomized collective policy. Algorithm 2 summarizes the procedure.

Algorithm 2 Pessimistic Von Neumann Winner

Input: The datasets $\mathcal{D} = \bigcup_{m=1}^M \mathcal{D}_m$, a failure probability δ , and the initial distribution ρ .

Output: The pessimistic policy $\hat{\pi}$.

- 1: Calculate the preference matrix N_s in (6) for $\forall s \in \mathcal{S}$.
 - 2: Construct the pessimistic lower confidence bound $N_s - B_s$ through (6) for $\forall s \in \mathcal{S}$.
 - 3: Calculate the state-wise strategy $\hat{p}_s$ by solving the matrix game of $N_s - B_s$ for $\forall s \in \mathcal{S}$.
 - 4: Obtain the final randomized policy $\hat{\pi}$ (induced by $\hat{p}_s$) through (7).
-

We construct a pessimistic von Neumann winner in a Utilitarian manner to learn a randomized policy. The algorithm is described in Algorithm 2. It mainly consists of two steps: Step 1 is to calculate the aggregated preference matrices and construct the confidence bounds; Step 2 is to solve a zero-sum matrix game of each state and finally construct a corresponding randomized policy in the policy space.

Step 1: Preference matrix with confidence bonus. We construct preference matrices by aggregating each party’s preferences in a Utilitarian manner. Given $\mathcal{D}_m = \{(s_m^i, a_{m,1}^i, a_{m,0}^i, y_m^i)\}_{i=1}^n$, we define the *visiting number* of an action pair at one state as $n_0(m, a, a', s) := \#_m(a > a'; s) + \#_m(a' > a; s)$, where $\#_m(a > a'; s)$ denotes the number of observations in $\mathcal{D}_m$ of preferring a over a' at state s . We then Define the preference matrix and its confidence bonus as follows:

$$N_s(a, a') = \frac{1}{M} \sum_{m=1}^M N_{s,m}(a, a'), \quad B_s(a, a') = \sqrt{\frac{2 \log(4MSA^2/\delta)}{\min_m n_0(m, a, a', s) \vee 1}}, \quad \forall a \neq a' \in \mathcal{A}, \quad (6)$$

where δ is a given failure probability, and $N_{s,m}(a, a') = \frac{\#_m(a > a'; s) - \#_m(a' > a; s)}{\#_m(a > a'; s) + \#_m(a' > a; s)}$ if the denominator is non-zero; otherwise, it is defined as 0. The diagonal entries of these matrices are set to 0. Here, the matrix N_s is an aggregated preference matrix, while B_s measures the uncertainty of N_s .

Step 2: Max-min optimization. After constructing the confidence bounds, we seek a max-min probabilistic strategy under a pessimistic model. This is achieved by solving the following zero-sum matrix game using the lower confidence bound $N_s - B_s$ of the preference matrix, and then obtaining the final probabilistic strategy:

$$\widehat{p}_s, \widehat{q}_s = \arg \max_p \min_q p^\top (N_s - B_s) q, \forall s \in \mathcal{S}, \quad \widehat{p}(\pi) = \prod_{s \in \mathcal{S}} \widehat{p}_s(\pi(s)). \quad (7)$$

To be clear, $\widehat{p}_s$ is a probabilistic strategy for each state $s \in \mathcal{S}$, and $\widehat{p}(\pi)$ is a probabilistic strategy over the policy space Π , where Π encompasses all deterministic policy permutations (any $\pi : \mathcal{S} \rightarrow \mathcal{A}$). Thus, $\widehat{p}$ induces a randomized policy $\widehat{\pi} : \mathcal{S} \rightarrow \Delta(\mathcal{A})$.

4.2 Efficiency Guarantees

In this section, we demonstrate that we obtain an approximate von Neumann winner of the ground truth preference matrix. To see this, we define the population-wise preference matrices at state s as $N_s^*(a, a') = \frac{1}{M} \sum_{m=1}^M N_{s,m}^*(a, a')$, and the one in the policy space Π as $T^*(\pi, \pi') = \mathbb{E}_{s \sim \rho} N_s^*(\pi(s), \pi'(s))$, respectively. Here $N_s^*(a, a')$ denotes the aggregated preference between action pairs at state s , and $T^*(\pi, \pi')$ denotes the aggregated preference between deterministic policy pairs.

Recall that $\mathcal{D}_m$ is generated by distribution g as described in Section 2.1. We define the occupancy measure $d_s^g(a, a')$ as the probability of pair (a, a') appearing in the data generated by g for a given state s : $d_s^g(a, a') = g_a(a, a'|s) + g_a(a', a|s)$. Similarly, define $d_s^{\mu, \nu}(a, a')$ as the probability of (a, a') appearing in the data generated by policies (μ, ν) respectively for a given state s . These definitions are adopted from those in [Cui and Du, 2022]. Define $p_s^*, q_s^* = \arg \max_p \min_q p^\top N_s^* q$ as the ground truth max-min and min-max probabilistic strategy at state s .

We further assume that $\rho(s) > 0$ for each $s \in \mathcal{S}$. Since $|\mathcal{S}| < \infty$, this implies a uniform lower bound of $\rho(s)$ for $s \in \mathcal{S}$. Now we define the concentratability coefficient as follows

$$C^* = \max_{\nu, s, a \neq a'} \frac{d_s^{p_s^*, \nu}(a, a')}{\rho(s) d_s^g(a, a')},$$

where C^* is assumed to be bounded [Cui and Du, 2022]. This definition is similar to (5). It provides a coverage guarantee of the dataset: the ratio of the state-action occupancy induced by the optimal policies, to the data distribution, is bounded.

Till now, we can establish our main result, whose proof is deferred to Appendix C.1.

Theorem 4.1. *With probability at least $1 - \delta$, $\widehat{p}_s$ and $\widehat{p}$ are ε -approximate von Neumann winner, i.e.,*

$$\min_q \widehat{p}_s^\top N_s^* q \geq -\varepsilon, \forall s \in \mathcal{S} \quad \text{and} \quad \min_q \widehat{p}^\top T^* q \geq -\varepsilon,$$

where $\varepsilon = 8A \log(4MSA^2/\delta) \sqrt{\frac{C^*}{n}}$.

Since the value of the skew-symmetric matrix game is zero [Owen, 2013], Theorem 4.1 demonstrates that the learned strategy achieves a worst-case population preference margin of at least $-\varepsilon$, both statewise and over the induced policy space. Thus, no competing strategy can defeat the

learned strategy by more than ε in aggregate preference margin. The error bound ε decreases at rate $n^{-1/2}$, grows with the action-space size, and depends only logarithmically on the numbers of parties and states apart from their influence through the concentratability coefficient C^* .

As we focus on social choice aspects, we again present a (weak) Pareto efficiency result to ensure consensus among multiple individuals, adapted from the concept of *ex-post efficiency* [Fishburn, 1984].

Definition 4.2 (τ -Approximate Ex-post Efficiency). A probabilistic strategy p_s is τ -approximate ex-post efficient at state s if any Pareto-dominated alternative b receives probability $p_s(b) \leq \tau$. Here an alternative b is Pareto-dominated if there exists an alternative a s.t. $N_{s,m}^*(a, b) \geq 0$ for $\forall m \in [M]$.

Our definition differs from Zeng and Psomas [2020], where they considered a strategy that always outputs a τ -approximate Pareto efficient alternative. We consider an ignorable probability ($\leq \tau$) of outputting a Pareto-dominated (universally disliked) alternative instead. Thus, a smaller τ means that the learned strategy is less likely to select an action that admits a common replacement weakly preferred by every party. We now present the following efficiency result.

Theorem 4.3. *If the preference of each party is transitive, i.e., $N_{s,m}^*(x, y) \geq 0$ implies $N_{s,m}^*(x, z) \geq N_{s,m}^*(y, z)$ for $\forall z \in \mathcal{A}$, then with probability at least $1 - \delta$, $\widehat{p}_s$ is τ -approximate ex-post efficient at state s , where $\tau \lesssim \left(A \log(2MSA^2/\delta) \sqrt{\frac{C^*}{n}} \right)^{1/2}$.*

Theorem 4.3 guarantees that Pareto-dominated alternatives are unlikely to be selected. Here the transitivity is necessary, implying the challenge of the general model. The proof and further details are provided in Appendix C.2. This result, in line with Theorem 3.7, underscores the efficiency and welfare guarantees of multi-party RLHF and offers a solid foundation from a social choice perspective.

5 Numerical Illustrations

We provide two stylized contextual-bandit simulations to illustrate the behavior of the proposed methods under heterogeneous party preferences and unequal party sample sizes. The first considers the reward-representable model and compares pessimistic Nash-welfare optimization with a pooled reward-model baseline. The second applies the general pairwise-preference method without using the underlying reward representation. These simulations isolate the effect of retaining party identities and applying an explicit equal-weight aggregation rule.

5.1 Reward-Representable Model

We consider a reward-representable instance with an unknown low-rank representation, where $r \ll d$. Specifically, we construct $|\mathcal{S}| = 3$ states, $|\mathcal{A}| = 9$ actions, and $M = 24$ parties, and set $d = 24$ and $r = 3$. The state distribution is uniform over $\mathcal{S}$. We defer the complete feature construction, data-generating process, and assumption verification to the Appendix D.1. For every party, the offline dataset compares each of the eight non-reference actions $a_1, \dots, a_8$ with the reference action a_9 in every state. The minimum per-party sample size ranges over $n = 150, 300, \dots, 3000$, and every sample size is repeated 100 times. In the experiment, the parties are divided into three equally sized clusters of eight; four parties in each cluster provide n comparisons and the

other four provide $2n$ comparisons (Note that our results remain valid for varying sample sizes n across different parties by replacing n with $\min_m n_m$).

We compare our joint low-rank Pessimistic Nash Welfare algorithm with three plug-in baselines. 1) The independent baseline estimates a separate unrestricted d -dimensional reward parameter for every party and therefore does not exploit the shared representation. 2) The pooling-obs baseline discards party identities and fits one reward model by weighting every comparison equally, so parties with $2n$ observations contribute twice as much as parties with n observations. 3) The pooling-party baseline also fits a single reward model but normalizes each party’s contribution, giving every party equal total weight. The two pooling baselines therefore distinguish the effect of unequal observation counts from the more fundamental loss of party-specific preference information. All methods select from the complete class of $9^3 = 729$ deterministic policies, and their performance is evaluated using the true party-specific rewards.

Figure 2a shows that our method achieves lower Nash welfare suboptimality than the independent and pooling baselines throughout the reported sample-size range. The independent estimator improves as the number of comparisons increases but remains less accurate than the joint method. Together with the estimation-error results in the appendix, this pattern supports the benefit of learning the shared low-dimensional reward representation. By contrast, both pooling baselines maintain a substantial suboptimality gap because a single pooled reward model cannot represent the three clusters’ heterogeneous preferences in this construction.

We further evaluate the same Nash-targeted policies under the Utilitarian and Egalitarian objectives and report their suboptimality relative to pooling-obs in the right panel. We also report a party-level dispersion measure that examines the difference between parties’ average regrets (see Appendix D). All ratios remain below one, indicating that our Nash-targeted policy also has lower Utilitarian and Egalitarian suboptimality, and a more balanced party-level regret profile than the pooled policy. These results indicate that the empirical advantage of our method persists when the selected policy is evaluated under welfare criteria other than the one it was optimized for. This broader improvement across welfare criteria is consistent with the empirical findings of [Bakker et al., 2022]. Additional details on evaluation metrics and results on confidence-set coverage and estimation accuracy are reported in Appendix D.

5.2 General Pairwise Preference model

Additionally, we present the results for the general pairwise-preference model. We consider a separate one-state instance with three actions and two parties. Although we use a BTL model to generate a controlled population preference matrix, the proposed method and the pooling baseline observe only pairwise comparison outcomes; neither method receives the underlying rewards, features, or reward parameters. We query action pairs (a_1, a_2) ; (a_1, a_3) and (a_2, a_3) an equal number of times within each party, while the sample sizes of Party 2 is twice that of Party 1. Let n denote the minimum per-party sample size. Party 1 provides n comparisons, whereas Party 2 provides $2n$ comparisons, we range the n from 400 to 4000. Each sample size is repeated 500 times. The results are shown in Figure 2b. Again, as the sample size increases, the game-value suboptimality of the strategy learned by our method decreases toward zero as the sample size increases, whereas the observation-pooling baseline retains a non-vanishing suboptimality gap. This demonstrates the robustness of our approach in both structured and unstructured preference settings. We defer further experimental setup and analysis details in

Appendix D.2.

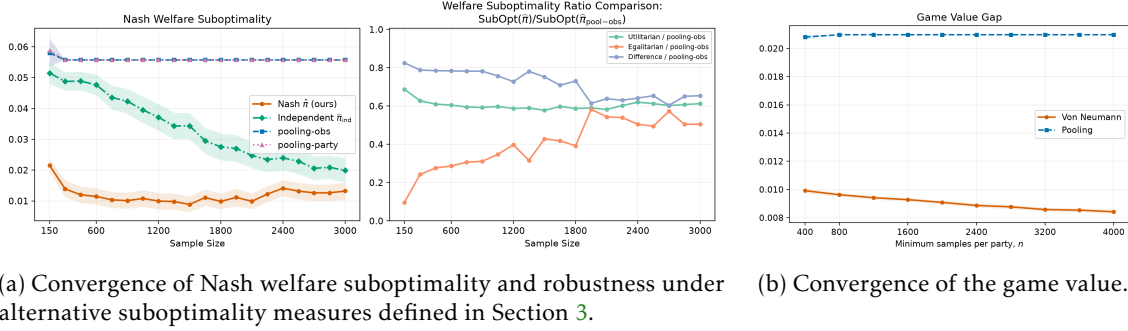


Figure 2: Simulation results.

6 Conclusion

Pluralistic alignment is not only a problem of learning from more people; it is also a problem of preserving meaningful disagreement and deciding explicitly how that disagreement should shape a single collective policy. This paper develops an offline finite-sample framework for doing so from party-indexed human comparisons. Rather than pooling heterogeneous feedback into one implicit reward model, our approach retains party-specific preference information throughout learning and separates the statistical task of estimating preferences from the normative choice of how those preferences should be aggregated. In the reward-representable regime, we exploit a shared low-rank structure to jointly estimate party-specific reward functions while preserving their differences. We then propagate the resulting statistical uncertainty through pessimistic Nash, Utilitarian, and Egalitarian welfare optimization. Under the stated realizability, diversity, normalization, and offline-coverage conditions, our guarantees characterize how shared structure, per-party estimation error, and data coverage jointly determine the quality and approximate efficiency of the learned collective policy. We further move beyond scalar reward representations to settings in which preferences may be stochastic, nontransitive, or cyclic. By estimating party-specific pairwise preferences and constructing a pessimistic von Neumann winner, we obtain a randomized collective policy with an approximate max-min preference guarantee. Under additional structural conditions, the policy also assigns limited probability to Pareto-dominated alternatives. Together, these results make preference aggregation an explicit and statistically analyzable component of RLHF, rather than an unintended consequence of data pooling. More broadly, they provide a principled foundation for collective learning systems in which both the statistical assumptions used to learn from heterogeneous feedback and the normative choices used to resolve disagreement are made transparent.

7 Discussion

KL penalty. In practice, people maximize a KL-regularized value function, which includes an additional term $\gamma D_{KL}(\hat{\pi}(\cdot|s) || \pi_{\text{ref}}(\cdot|s))$, where γ is a penalty parameter and π_{ref} is a reference policy. While our primary focus is the basic version, our results can be extended to the KL version,

where the sub-optimality gap becomes accompanied by an additional $\gamma \mathbb{E}_{s \sim \rho} D_{\text{KL}}(\widehat{\pi}(\cdot|s) \parallel \pi^*(\cdot|s))$ term [Xiong et al., 2023].

Computation. Algorithm 1 is an extension of algorithms in Zhan et al. [2024]; Zhu et al. [2023], all of which are typically computationally heavy. Here, we illustrate how to practically implement the algorithm. Step 2 of Algorithm 1 involves solving a constrained max-min optimization problem. By introducing the Lagrangian multiplier, which is akin to the approach employed in Appendix A of Zhan et al. [2024], it is feasible to apply gradient ascent-descent using policy gradient algorithms. In the general pairwise-preference model, finding a von Neumann winner, i.e., Step 2 of Algorithm 2, involves solving the matrix game. This can be achieved by linear programming [Barron, 2013], making it computationally efficient. Nevertheless, this efficiency for the general pairwise-preference model is achieved in tabular cases.

Future work. Several directions remain open. First, it would be useful to extend shared multi-party reward learning beyond linear reward models to general function classes while retaining finite-sample confidence guarantees and propagating the resulting uncertainty through social-welfare optimization. Second, an important open problem is to develop finite-sample methods that jointly quantify uncertainty in latent or overlapping party membership and in potentially evolving party-specific preferences, and propagate these sources of uncertainty through social-welfare optimization of a single collective policy. Third, bridging the gap between the present stylized experiments and practical LLM alignment requires scalable methods for party-specific reward estimation, uncertainty-aware welfare optimization, and collective policy fine-tuning over large response and policy spaces.

References

- Parand A. Alamdari, Soroush Ebadian, and Ariel D. Procaccia. Policy aggregation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 68308–68329. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7e670825a578392891ad40e93931b1e3-Paper-Conference.pdf.
- Yonatan Aumann and Yair Dombb. Pareto efficiency and approximate pareto efficiency in routing and load balancing games. In *International Symposium on Algorithmic Game Theory*, pages 66–77. Springer, 2010.
- Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.
- Siddharth Barman, Sanath Kumar Krishnamurthy, and Rohit Vaish. Finding fair and efficient allocations. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 557–574, 2018.
- Emmanuel N Barron. *Game theory: an introduction*. John Wiley & Sons, 2013.

- Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.
- Avinandan Bose, Simon Shaolei Du, and Maryam Fazel. Offline multi-task transfer RL with representational penalization. *arXiv preprint arXiv:2402.12570*, 2024.
- Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Florian Brandl, Felix Brandt, and Hans Georg Seedig. Consistent probabilistic social choice. *Econometrica*, 84(5):1839–1880, 2016.
- Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.
- Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Singh Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6116–6135. PMLR, 2024. URL <https://proceedings.mlr.press/v235/chakraborty24b.html>.
- Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, volume 162, pages 3773–3793. PMLR, 2022.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, page 4302–4310, 2017.
- Qiwen Cui and Simon S Du. When are offline two-player zero-sum Markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791, 2022.
- Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.
- Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.
- Sara Fish, Paul Gözl, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC ’24, page 985, New York, NY, USA, 2024. Association for Computing Machinery.
- Peter C Fishburn. Condorcet social choice functions. *SIAM Journal on applied Mathematics*, 33(3):469–489, 1977.
- Peter C Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

- Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for ai alignment from human feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80439–80465. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf.
- Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1 – 6, 2012.
- Peizhong Ju, Arnob Ghosh, and Ness Shroff. Achieving fairness in multi-agent MDP using reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- Frank P Kelly, Aman K Maulloo, and David Kim Hong Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, 49:237–252, 1998.
- Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. Strategyproof reinforcement learning from human feedback. In *Advances in Neural Information Processing Systems*, volume 38, pages 101431–101464. Curran Associates, Inc., 2025. doi: 10.52202/085713-3396. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/92f43b1d33fae4aa1958f75317f0cec1-Paper-Conference.pdf.
- Seong Jin Lee, Will Wei Sun, and Yufeng Liu. Low-rank contextual reinforcement learning from heterogeneous human feedback. *Journal of the American Statistical Association*, 121(554): 1037–1050, 2026. doi: 10.1080/01621459.2026.2674413. URL <https://doi.org/10.1080/01621459.2026.2674413>.
- Gene Li, Cong Ma, and Nati Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.
- Sai Li and Linjun Zhang. Multi-dimensional domain generalization with low-rank structures. *arXiv preprint arXiv:2309.09555*, 2023.
- Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.
- Renpu Liu, Peng Wang, Donghao Li, Cong Shen, and Jing Yang. A shared low-rank adaptation approach to personalized rlhf. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1405–1413. PMLR, 2025. URL <https://proceedings.mlr.press/v258/liu25d.html>.
- Debmalya Mandal and Jiarui Gan. Socially fair reinforcement learning. *arXiv preprint arXiv:2208.12584*, 2022.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.
- Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.
- Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegl, Andrea

- Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36743–36768. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/munos24a.html>.
- Włodzimierz Ogryczak, Patrice Perny, and Paul Weng. A compromise programming approach to multiobjective markov decision processes. *International Journal of Information Technology & Decision Making*, 12(05):1021–1053, 2013.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Guillermo Owen. *Game theory*. Emerald Group Publishing, 2013.
- Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E. Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024.
- Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.
- Ronald L Rivest and Emily Shen. An optimal single-winner preferential voting system based on game theory. In *International Workshop on Computational Social Choice*, pages 399–410. Citeseer, 2010.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.
- Amartya Sen. *Collective choice and social welfare*. Harvard University Press, 2018.
- Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8905–8915. PMLR, 13–18 Jul 2020.
- Umer Siddique, Abhinav Sinha, and Yongcan Cao. Fairness in preference-based reinforcement learning. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.
- Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*, 2024.
- Gokul Swamy, Christoph Dann, Rahul Kidambi, Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st*

- International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47345–47377. PMLR, 2024. URL <https://proceedings.mlr.press/v235/swamy24a.html>.
- Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, 2020.
- Amos Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024.
- Tianze Wang, Dongnan Gui, Yifan Hu, Shuhang Lin, and Linjun Zhang. MPO: An efficient post-processing framework for mixing diverse preference alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 62958–62979. PMLR, 2025. URL <https://proceedings.mlr.press/v267/wang25af.html>.
- Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? a theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, 2025. doi: 10.1080/01621459.2025.2555067. URL <https://doi.org/10.1080/01621459.2025.2555067>.
- Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 34:27395–27407, 2021.
- Nuoya Xiong and Aarti Singh. Projection optimization: A general framework for multi-objective and multi-group rlhf. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 69017–69058. PMLR, 2025. URL <https://proceedings.mlr.press/v267/xiong25c.html>.
- Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.
- Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102:315–323, 2014.
- David Zeng and Alexandros Psomas. Fairness-efficiency tradeoffs in dynamic fair division. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 911–912, 2020.
- Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline

- preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
- Guojun Zhang, Saber Malekmohammadi, Xi Chen, and Yaoliang Yu. Proportional fairness in federated learning. *arXiv preprint arXiv:2202.01666*, 2022.
- Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2023.
- Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Further Details on Methods

A.1 Social Welfare Function Details

Here, we provide additional details regarding social welfare functions to complement the discussion in Section 2.2. In welfare economics, a social welfare function aggregates individual utilities towards collective welfare and thus guides social decisions to align with the group’s preferences. A reasonable social welfare function that satisfies six desirable axioms has been shown to lie in a one-parameter family of isoelastic social welfare functions W_α defined in (1) (or its monotonically increasing transformation).

Among these functions, the Nash welfare function ($\alpha = 0$) stands out as the unique solution that additionally satisfies *independence of individual scales (IIS)*; and the Utilitarian welfare function ($\alpha = 1$) is the unique one that additionally satisfies *independence of individual zeros (IIZ)*. The definitions of these properties are provided below.

Definition A.1 (Independence of Individual Scales of Utilities (IIS)). A solution is independent of individual scales of utilities if, for $\forall m \in [M]$, transforming the m -th utility function u_m to $c_m u_m$ with $c_m \in \mathbb{R}^+$ does not change the solution.

Definition A.2 (Independence of Individual Zeros of Utilities (IIZ)). A solution is independent of individual zeros of utilities if, for $\forall m \in [M]$, transforming the m -th utility function u_m to $c_m + u_m$ with $c_m \in \mathbb{R}$ does not change the solution.

Additionally, taking the limit $\alpha \rightarrow -\infty$ yields Egalitarian welfare function, which can be reduced to consider $\min_m u_m$ for simplicity.

In the main article, we introduce the Nash welfare function as the primary illustrative example. The Nash welfare function is advantageous due to its IIS property, which eliminates the potential issues of dominance arising from differences in utility scales. Moreover, one can also consider the Nash bargaining version which introduces an objective notion of individual “zero utility”. This involves the normalization of individual zeros $\{u_m^0\}_{m=1}^M$, and is formulated as follows:

$$\arg \max_u \prod_{m=1}^M (u_m - u_m^0)^{1/M}, \text{ s.t. } u_m > u_m^0 \text{ for } \forall m \in [M].$$

Thus, Nash bargaining solution satisfies the IIZ property.

A.2 Connections Between Two Models

In this section, we outline the connection between the reward-based and general reward-free models. The average-wise definition of the von Neumann winner in Section 4.1 is grounded in the Utilitarian welfare function. Here, we provide the rationale behind this choice by considering a voting scenario. In the context of a voting election, each voter has an absolute preference ranking over the set of alternatives. For any pair of alternatives $a \neq a'$, we have $N_{s,m}(a, a') \in \{1, -1\}$ for $a \neq a'$. Aggregating over all voters, $N_s(a, a')$ thus represents the disparity between the count of individuals preferring a over a' and the count of those preferring a' over a . Consequently, a von Neumann winner represents the outcome favored by a majority of voters. Here the normalization for $N_{s,m}$ is crucial to avoid dominance due to unequal sample sizes or preference weights.

We have claimed that, under the additional assumption of a reward-based model, the von Neumann winner in the general model consistently aligns with the solution of the Utilitarian welfare function when $M = 1$ or 2 . We give a brief illustration as follows.

- Case $M = 1$: In this scenario, there exists a Condorcet winner [Fishburn, 1977] determined by the optimal reward. Consequently, the Condorcet winner that maximizes the reward function is trivially the von Neumann winner.
- Case $M = 2$: We first establish the following inequality for any real numbers x, y ,

$$\frac{1}{2} \left(\frac{1}{1+e^{-x}} + \frac{1}{1+e^{-y}} \right) \geq \frac{1}{2} \Leftrightarrow x + y \geq 0.$$

Next, we substitute $x = R_1(s, a_1) - R_1(s, a_0)$, $y = R_2(s, a_1) - R_2(s, a_0)$, where $R_m(s, a)$ denotes the reward function for the m -th party. Using the result above, we obtain:

$$\frac{1}{2} [\mathbb{P}_1(a_1 > a_0 | s) + \mathbb{P}_2(a_1 > a_0 | s)] \geq \frac{1}{2} \Leftrightarrow R_1(s, a_1) + R_2(s, a_1) \geq R_1(s, a_0) + R_2(s, a_0).$$

This equivalence demonstrates that the aggregate preference between any action pair, based on the average of the rewards across both parties, determines the von Neumann winner.

A.3 Generalization to MDPs

We now generalize our results for the reward-based model in Section 3 to MDPs where the preference is based on a whole trajectory instead of a single action. Examples of such settings include Atari games and robot locomotion [Christiano et al., 2017], where the preference depends on historical behaviors. Consider a MDP with M individuals (parties), each of whom has its own reward function. Formally, we consider the environment $E = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^{H-1}, \{R_m\}_{m=1}^M, \rho)$, where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, H is the horizon length, $P_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ for $h \in [H-1]$ are known transition kernels, $R_m : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function for the m -th individual (or party), and $\rho : \mathcal{S} \rightarrow \Delta(\mathcal{S})$ is an initial state distribution. When $H = 1$ and there are no transitions, it is reduced to the contextual bandit (CB) setting. Define the occupancy measure $d^\pi : \mathcal{S} \rightarrow \mathbb{R}$ as $d^\pi(s) = \sum_{h=1}^H \mathbb{P}_h(s_h = s | \pi)$, which becomes $\rho(s)$ when $H = 1$. We are provided with a dataset $\mathcal{D} = \bigcup_{m=1}^M \mathcal{D}_m$, where $\mathcal{D}_m = \{(s_m^i, \tau_{m,1}^i, \tau_{m,0}^i, y_m^i)\}_{i=1}^n$ is the pairwise comparison dataset for the m -th individual. For the i -th sample in $\mathcal{D}_m$, an initial state s_m^i is sampled from ρ and two trajectories $\tau_{m,1}^i = (a_{m,1}^i, \dots, s_{m,H}^i, a_{m,H}^i)$, $\tau_{m,0}^i = (a_{m,1}^{i'}, \dots, s_{m,H}^{i'}, a_{m,H}^{i'})$ are sampled from a distribution $g_\tau(\tau_{m,1}^i, \tau_{m,0}^i | s_m^i)$. Thus $(s_m^i, \tau_{m,1}^i, \tau_{m,0}^i)$ is generated by a generation distribution with density $g = g_\tau(\cdot | s) \rho(s)$. We then observe a binary outcome y_m^i from a Bernoulli distribution $\mathbb{P}_m(y_m^i = 1 | \tau_{m,1}^i, \tau_{m,0}^i)$ as follows,

$$\mathbb{P}_m(y = 1 | \tau_1, \tau_0, s) = \Phi \left(\sum_{h=1}^H R_{m, \theta_m^*}(s_h, a_h) - \sum_{h=1}^H R_{m, \theta_m^*}(s_h', a_h') \right),$$

where Assumption 3.1, 3.2 hold with the feature difference in Σ_* changed to the cumulative difference over the trajectory, i.e.,

$$\Sigma_* = \mathbb{E}_g \left(\sum_{h=1}^H (\phi(s_h, a_h) - \phi(s_h', a_h')) \right) \left(\sum_{h=1}^H (\phi(s_h, a_h) - \phi(s_h', a_h')) \right)^\top. \quad (8)$$

The generalized algorithm for MDPs is similar to Algorithm 1. However, here we adopt a trajectory-based preference model and thus optimize a trajectory-wise value function. We are going to discuss the differences in detail below.

Denote $X_{m,i} = \sum_{h=1}^H (\phi(s_{m,h}^i, a_{m,h}^i) - \phi(s_{m,h}^{i'}, a_{m,h}^{i'}))$ now as the difference between the cumulative feature in two trajectories. Define the data covariance matrices as $\Sigma_m = \frac{1}{n} \sum X_{m,i} X_{m,i}^\top$ for $\forall m \in [M]$. We again minimize the negative log-likelihood:

$$\widehat{U}, \widehat{\alpha} \in \arg \min_{U \in \mathcal{O}_{d \times r}, \|\alpha_m\|_2 \leq B} l_{\mathcal{D}}(U, \alpha),$$

$$l_{\mathcal{D}}(U, \alpha) = -\frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \left\{ \mathbf{1}(y_m^i = 1) \log(\Phi(\langle U \alpha_m, X_{m,i} \rangle)) + \mathbf{1}(y_m^i = 0) \log(\Phi(\langle U \alpha_m, -X_{m,i} \rangle)) \right\}.$$

Subsequently, we use Nash welfare function to aggregate individual preferences. Here we point out that the key difference between MDPs and CBs is the inclusion of the trajectory-based measure d^π , which stems from the observation [Zhu et al., 2023] that:

$$\mathbb{E}_{s \sim \rho}[V_m^\pi(s)] = \mathbb{E}_{s \sim d^\pi} R_{m, \theta_m^*}(s, \pi(s)),$$

where $V_m^\pi(s) = \mathbb{E}[\sum_{h=1}^H R_m(s_h, a_h) | s_1 = s, \pi]$, and this expectation $\mathbb{E}$ is taken over the trajectory generated according to the transition kernel. Since the transition distribution P is known, we can directly calculate d^π for any given π . Define the trajectory-wise true and pessimistic Nash welfare function J and $\widehat{J}$ as follows:

$$J(\pi) = \prod_{m=1}^M [\mathbb{E}_{s \sim d^\pi} R_{m, \theta_m^*}(s, \pi(s))]^{1/M},$$

$$\widehat{J}(\pi) = \min_{\Theta \in \Theta_B(\widehat{\Theta})} \prod_{m=1}^M [\mathbb{E}_{s \sim d^\pi} R_{m, \theta_m}(s, \pi(s))]^{1/M},$$

where $\Theta_B(\widehat{\Theta}_m) = \{\theta \in \Theta_B : \|\theta - \widehat{\theta}_m\|_{\Sigma_m} \leq K' \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}}\}$, K' is a fixed parameter.

Define $C^* = \max_m \mathbb{E}_{s \sim d^{\pi^*}} \|(\Sigma_m^{-1/2} \phi(s, \pi^*(s)))\|_2$. Similarly, we assume that the optimal policy value for every party is bounded below by a positive constant, i.e., $\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \theta_m^*}(s, \pi^*(s)) \geq r_0$ for $\forall m \in [M]$. We have the following result, whose proof is deferred to Appendix B.6.

Theorem A.3. *Under the same condition as Theorem 3.3 (with Σ_* replaced by (8)), we have, then with probability at least $1 - \delta$, for any $m \in [M]$,*

$$J(\pi^*) - J(\widehat{\pi}) \lesssim C^* \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}}.$$

The sub-optimality scales with the concentratability coefficient C^* . The primary distinction between Theorem B.8 and Theorem 3.4 lies in the trajectory-wise form of the concentratability coefficient C^* . We can also attain efficiency and fairness guarantees similar to Section 3.3 accordingly.

B Technical Details for Reward-Based Model

We denote the low-dimensional parameters and the original parameters as follows:

$$\begin{aligned}\alpha &= (\alpha_1, \dots, \alpha_M), \quad \alpha^* = (\alpha_1^*, \dots, \alpha_M^*), \quad \widehat{\alpha} = (\widehat{\alpha}_1, \dots, \widehat{\alpha}_M) \in \mathbb{R}^{r \times M}, \\ \Theta &= (\theta_1, \dots, \theta_M), \quad \Theta^* = (\theta_1^*, \dots, \theta_M^*), \quad \widehat{\Theta} = (\widehat{\theta}_1, \dots, \widehat{\theta}_M) = (\widehat{U}\widehat{\alpha}_1, \dots, \widehat{U}\widehat{\alpha}_M) \in \mathbb{R}^{d \times M}.\end{aligned}$$

We denote the feature difference vectors matrices as follows:

$$X_{m,i} = \phi(s_m^i, a_{m,1}^i) - \phi(s_m^i, a_{m,0}^i) \in \mathbb{R}^d, \quad X_m = (X_{m,1}, \dots, X_{m,n})^\top \in \mathbb{R}^{n \times d}, \quad \forall m, i.$$

Through out this section, we define $\theta, \widehat{\theta}, \theta^* \in \mathbb{R}^{dM}$ as the vectorization of $\Theta, \widehat{\Theta}, \Theta^*$. We write $\Sigma \succeq (\leq) \Sigma'$ if $\Sigma - \Sigma'$ is positive (negative) semidefinite.

B.1 Proof of Theorem 3.3

Lemma B.1 (Covariance concentration). *Suppose Assumption 3.2 holds. If the number of samples satisfies $n \gg d + \log(M/\delta)$ for $\delta \in (0, 1)$, then with probability at least $1 - \delta/10$, we have: for any $m \in [M]$,*

$$0.9\Sigma_* \leq \Sigma_m \leq 1.1\Sigma_*. \quad (9)$$

Proof. First, we fix $m \in [M]$ and recall the definition $X_{m,i} := \phi(s_m^i, a_{m,1}^i) - \phi(s_m^i, a_{m,0}^i)$. It follows that $\Sigma_m = \frac{1}{n} \sum_{i=1}^n X_{m,i} X_{m,i}^\top = \frac{1}{n} X_m^\top X_m$, and its expectation satisfies $\mathbb{E}_g \Sigma_m = \Sigma_*$. Since ϕ is bounded from the assumption, $X_{m,i}$ $i \in [n]$ are independent sub-Gaussian random vectors. Consequently, with probability at least $1 - 2\exp(-t^2)$,

$$\begin{aligned}\|\Sigma_m - \Sigma_*\| &= \left\| \frac{1}{n} \sum_{i=1}^n X_{m,i} X_{m,i}^\top - \Sigma_* \right\| \\ &\lesssim \|\Sigma_*\| \max(\eta, \eta^2), \quad \eta = \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}.\end{aligned}$$

See Theorem 4.6.1 in [Vershynin \[2018\]](#) or Lemma 7 in [Tripuraneni et al. \[2020\]](#) for reference. Thus, since $\|\Sigma_*\| \leq C_{\max}$, it follows that with probability at least $1 - \delta$,

$$\|\Sigma_m - \Sigma_*\| \lesssim \sqrt{\frac{d + \log(1/\delta)}{n}}.$$

Next, we apply a union bound over all $m \in [M]$. We substitute δ with δ/M , and obtain that, with probability at least $1 - \delta$, for any $m \in [M]$

$$\|\Sigma_m - \Sigma_*\| \lesssim \sqrt{\frac{d + \log(M/\delta)}{n}}$$

Thus, the bound $\|\Sigma_m - \Sigma_*\| \leq 0.1\|\Sigma_*\|$ holds for any $m \in [M]$ given $n \gg d + \log(M/\delta)$. ■

Theorem B.2 (Guarantee on source data). *Suppose Assumption 3.1, 3.2 hold. If the number of samples satisfies $n \gg d + \log(M/\delta)$ for $\delta \in (0, 1)$, then with probability at least $1 - \delta/5$,*

$$\sum_{m=1}^M \|X_m \widehat{U} \widehat{\alpha}_m - X_m U^* \alpha_m^*\|^2 \lesssim Mr + rd \log d + \log(1/\delta). \quad (10)$$

Proof. We assume that (9) is true, which happens with probability at least $1 - \delta/10$ according to Lemma B.1.

Step 1: Optimality of $\widehat{\theta}$. Consider the following loss function, along with its first and second derivatives,

$$\begin{aligned} l_{\mathcal{D}}(\theta) &= -\frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \log[1(y_m^i = 1) \frac{1}{\exp(-\langle \theta_m, X_{m,i} \rangle) + 1} + 1(y_m^i = 0) \frac{\exp(-\langle \theta_m, X_{m,i} \rangle)}{\exp(-\langle \theta_m, X_{m,i} \rangle) + 1}], \\ \nabla_{\theta_m} l_{\mathcal{D}}(\theta) &= -\frac{1}{Mn} \sum_{i=1}^n \left[1(y_m^i = 1) \frac{\exp(-\langle \theta_m, X_{m,i} \rangle)}{\exp(-\langle \theta_m, X_{m,i} \rangle) + 1} - 1(y_m^i = 0) \frac{1}{\exp(-\langle \theta_m, X_{m,i} \rangle) + 1} \right] X_{m,i}, \\ \nabla_{\theta_m}^2 l_{\mathcal{D}}(\theta) &= \frac{1}{Mn} \sum_{i=1}^n \frac{\exp(-\langle \theta_m, X_{m,i} \rangle)}{(\exp(-\langle \theta_m, X_{m,i} \rangle) + 1)^2} \cdot X_{m,i} X_{m,i}^\top. \end{aligned}$$

Thus we have,

$$\begin{aligned} \nabla_{\theta} l_{\mathcal{D}}(\theta) &= (\nabla_{\theta_1} l_{\mathcal{D}}(\theta)^\top, \dots, \nabla_{\theta_M} l_{\mathcal{D}}(\theta)^\top)^\top, \\ \nabla_{\theta}^2 l_{\mathcal{D}}(\theta) &= \text{diag}(\nabla_{\theta_1}^2 l_{\mathcal{D}}(\theta), \dots, \nabla_{\theta_M}^2 l_{\mathcal{D}}(\theta)). \end{aligned}$$

Regarding the second derivative, since $|\langle \theta_m, X_{m,i} \rangle| \leq 2BL$, we have $\frac{\exp(-\langle \theta_m, X_{m,i} \rangle)}{(\exp(-\langle \theta_m, X_{m,i} \rangle) + 1)^2} \geq \gamma$ where $\gamma = \frac{1}{2 + \exp(-2BL) + \exp(2BL)}$. Therefore, we obtain the lower bound

$$\nabla_{\theta_m}^2 l_{\mathcal{D}}(\theta) \geq \frac{\gamma}{Mn} \sum_{i=1}^n X_{m,i} X_{m,i}^\top = \frac{\gamma}{M} \Sigma_m.$$

Define $\Sigma_{\text{sum}} = \text{diag}(\Sigma_1, \dots, \Sigma_M) \in \mathbb{R}^{Mn \times Mn}$, we conclude from strong convexity that

$$l_{\mathcal{D}}(\widehat{\theta}) - l_{\mathcal{D}}(\theta^*) - \langle \nabla l_{\mathcal{D}}(\theta^*), \widehat{\theta} - \theta^* \rangle \geq \frac{\gamma}{2M} (\widehat{\theta} - \theta^*)^\top \Sigma_{\text{sum}} (\widehat{\theta} - \theta^*) = \frac{\gamma}{2M} \|\widehat{\theta} - \theta^*\|_{\Sigma_{\text{sum}}}^2.$$

Here the matrix-induced norm is defined as $\|x\|_{\Sigma} = \sqrt{x^\top \Sigma x}$. In the following, we also use $\|Z\|_{\Sigma_{\text{sum}}}$ to denote the Σ_{sum} -norm of Z 's vectorization when Z is a matrix. From the optimality of $\widehat{\theta}$, we have $l_{\mathcal{D}}(\widehat{\theta}) - l_{\mathcal{D}}(\theta^*) \leq 0$, it follows that

$$-\langle \nabla l_{\mathcal{D}}(\theta^*), \widehat{\theta} - \theta^* \rangle \geq \frac{\gamma}{2M} \|\widehat{\theta} - \theta^*\|_{\Sigma_{\text{sum}}}^2. \quad (11)$$

Using Cauchy-Schwartz inequality immediately implies that

$$\|l_{\mathcal{D}}(\theta^*)\|_{\Sigma_{\text{sum}}^{-1}} \|\widehat{\theta} - \theta^*\|_{\Sigma_{\text{sum}}} \geq \frac{\gamma}{2M} \|\widehat{\theta} - \theta^*\|_{\Sigma_{\text{sum}}}^2. \quad (12)$$

Regarding the first derivative, notice that $y_m^i | X_{m,i} \sim \text{Ber}\left(\frac{1}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1}\right)$, we define a random vector $V_m \in \mathbb{R}^n$ with components $V_{m,i}$ as follows,

$$\begin{aligned} V_{m,i} &= \mathbf{1}(y_m^i = 1) \frac{\exp(-\langle \theta_m^*, X_{m,i} \rangle)}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1} - \mathbf{1}(y_m^i = 0) \frac{1}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1} \\ &= \begin{cases} \frac{\exp(-\langle \theta_m^*, X_{m,i} \rangle)}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1} & \text{w.p. } \frac{1}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1}, \\ \frac{-1}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1} & \text{w.p. } \frac{\exp(-\langle \theta_m^*, X_{m,i} \rangle)}{\exp(-\langle \theta_m^*, X_{m,i} \rangle) + 1}. \end{cases} \end{aligned} \quad (13)$$

Here V_m solely depends on y_m when conditioned on X_m . From the definition of V_m , we have $\nabla_{\theta_m} l_{\mathcal{D}}(\theta^*) = -\frac{1}{Mn} X_m^\top V_m$. So conditioned on X_m , V_m is a random vector with independent components that satisfy $\mathbb{E}V_i = 0, \|V_{m,i}\| \leq 1$, and is therefore sub-Gaussian. Denote $A_m = \frac{1}{n^2} X_m \Sigma_m^{-1} X_m^\top$, we can then establish the following bound [see, Appendix B.1, [Zhu et al., 2023](#)]:

$$\text{tr}(A_m) \leq \frac{d}{n}, \quad \text{tr}(A_m^2) \leq \frac{d}{n^2}, \quad \|A_m\| \leq \frac{1}{n}.$$

By applying Bernstein's inequality for sub-Gaussian random variables in a quadratic form with $x = (V_1^\top, \dots, V_M^\top)^\top$ and $A = \text{diag}(A_1, \dots, A_M)$ [see, Theorem 2.1, [Hsu et al., 2012](#)], we directly obtain that, with probability at least $1 - \delta/20$,

$$M^2 \|l_{\mathcal{D}}(\theta^*)\|_{\Sigma_{\text{sum}}^{-1}}^2 = \sum_{m=1}^M \frac{1}{n^2} V_m^\top X_m \Sigma_m^{-1} X_m^\top V_m \lesssim \frac{Md + \log(1/\delta)}{n}. \quad (14)$$

Combining (14) with (12), we obtain that, with probability at least $1 - \delta/20$,

$$\|\widehat{\theta} - \theta^*\|_{\Sigma_{\text{sum}}} \lesssim \sqrt{\frac{Md + \log(1/\delta)}{n}}. \quad (15)$$

This implies that

$$\begin{aligned} Md + \log(1/\delta) &\gtrsim \sum_{m=1}^M (\widehat{\theta}_m - \theta_m^*)^\top X_m^\top X_m (\widehat{\theta}_m - \theta_m^*) \\ &\geq n \sum_{m=1}^M (\widehat{\theta}_m - \theta_m^*)^\top 0.9 \Sigma_* (\widehat{\theta}_m - \theta_m^*) && \text{(using (9))} \\ &\geq 0.9n C_{\min} \|\widehat{\theta} - \theta^*\|^2 && \text{(using Assumption 3.2)} \\ &\gtrsim n \|\widehat{\theta} - \theta^*\|^2. \end{aligned}$$

Thus, with probability at least $1 - \delta/20$,

$$\|\widehat{\theta} - \theta^*\|_F^2 = \|\widehat{\theta} - \theta^*\|^2 \lesssim \sqrt{\frac{Md + \log(1/\delta)}{n}}. \quad (16)$$

Here we use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix.

Step 2: Combining with low-dimensional assumption. Notice that $\text{rank}(\widehat{\Theta} - \Theta^*) \leq 2r$, so we can rewrite it as $\widehat{\Theta} - \Theta^* = QR = (Q\mathbf{r}_1, \dots, Q\mathbf{r}_M)$ where $Q \in \mathcal{O}_{d \times 2r}$, $R = (\mathbf{r}_1, \dots, \mathbf{r}_M) \in \mathbb{R}^{2r \times M}$. Thus the vectorization can be rewritten as $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = [(Q\mathbf{r}_1)^\top, \dots, (Q\mathbf{r}_M)^\top]^\top$. For each $m \in [M]$ we further rewrite $X_m Q = Z_m W_m$ where $Z_m \in \mathcal{O}_{n \times 2r}$, $W_m \in \mathbb{R}^{2r \times 2r}$. Then we have

$$\begin{aligned}
-\langle \nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*), \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \rangle &= \sum_{m=1}^M \frac{1}{Mn} V_m^\top X_m Q \mathbf{r}_m \\
&= \frac{1}{Mn} \sum_{m=1}^M V_m^\top Z_m W_m \mathbf{r}_m \\
&\leq \frac{1}{Mn} \sum_{m=1}^M \|Z_m^\top V_m\| \|W_m \mathbf{r}_m\| \\
&\leq \frac{1}{Mn} \sqrt{\sum_{m=1}^M \|Z_m^\top V_m\|^2} \sqrt{\sum_{m=1}^M \|W_m \mathbf{r}_m\|^2} \\
&= \frac{1}{Mn} \sqrt{\sum_{m=1}^M \|Z_m^\top V_m\|^2} \sqrt{\sum_{m=1}^M \|Z_m W_m \mathbf{r}_m\|^2} \\
&= \frac{1}{Mn} \sqrt{\sum_{m=1}^M \|Z_m^\top V_m\|^2} \sqrt{\sum_{m=1}^M \|X_m Q \mathbf{r}_m\|^2} \tag{17}
\end{aligned}$$

where the second term can be calculated as $\sum_{m=1}^M \|X_m Q \mathbf{r}_m\|^2 = \sum_{m=1}^M \|X_m(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_m^*)\|^2 = n\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\Sigma_{\text{sum}}}^2$.

Step 3: Applying ε -net argument. Next, we are going to give a high-probability upper bound on $\sum_{m=1}^M \|Z_m^\top V_m\|^2$ using the randomness of V_m . Since Z_m depends on Q which depends on V , we use an ε -net argument to cover all possible $Q \in \mathcal{O}_{d \times 2r}$. First, for any fixed $\bar{Q} \in \mathcal{O}_{d \times 2r}$, we let $X_m \bar{Q} = \bar{Z}_m \bar{W}_m$ where $\bar{Z}_m \in \mathcal{O}_{n \times 2r}$. The $\bar{Z}_m$'s defined in this way are independent of V .

From the orthonormality of $\bar{Z}_m$, we can establish the following equation:

$$\text{tr}(\bar{Z}_m \bar{Z}_m^\top) = r, \quad \text{tr}(\bar{Z}_m \bar{Z}_m^\top)^2 = r, \quad \|\bar{Z}_m \bar{Z}_m^\top\| = 1$$

Again, use Bernstein's inequality for sub-Gaussian random variables in the quadratic form as we did in (14), we obtain that, with probability at least $1 - \delta'$,

$$\sum_{m=1}^M \|\bar{Z}_m^\top V_m\|^2 = \sum_{m=1}^M V_m \bar{Z}_m \bar{Z}_m^\top V_m \lesssim Mr + \log(1/\delta').$$

Combining this with (17), we obtain that

$$-\langle \nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*), \bar{Q}R \rangle \leq \sqrt{\frac{Mr + \log(1/\delta')}{M^2 n}} \|\bar{Q}R\|_{\Sigma_{\text{sum}}}.$$

Lemma A.5 in Du et al. [2021] implies that: there exists an ε -net $\mathcal{N}$ of $\mathcal{O}_{d \times 2r}$ in Frobenius norm such that $|\mathcal{N}| \leq \left(\frac{6\sqrt{2r}}{\varepsilon}\right)^{2rd}$. Applying a union bound over $\mathcal{N}$, we obtain that, with probability at

least $1 - |\mathcal{N}|\delta'$,

$$-\langle \nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*), \bar{Q}R \rangle \leq \sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} \|\bar{Q}R\|_{\Sigma_{\text{sum}}}, \quad \forall \bar{Q} \in \mathcal{N}. \quad (18)$$

Choosing $\delta' = \frac{\delta}{20(\frac{6\sqrt{2}r}{\varepsilon})^{2rd}}$, we obtain that (18) holds with probability at least $1 - \delta/20$.

Next, we proceed to apply ε -net. Let $\bar{Q} \in \mathcal{N}$ such that $\|\bar{Q} - Q\|_F \leq \varepsilon$, then we have

$$\begin{aligned} \|(\bar{Q} - Q)R\|_{\Sigma_{\text{sum}}}^2 &= \frac{1}{n} \sum_{m=1}^M \|X_m(\bar{Q} - Q)\mathbf{r}_m\|^2 \\ &\leq \frac{1}{n} \sum_{m=1}^M \|X_m\|^2 \|\bar{Q} - Q\|^2 \|\mathbf{r}_m\|^2 \\ &\leq \|\Sigma_*\| \varepsilon^2 \sum_{m=1}^M \|\mathbf{r}_m\|^2 \\ &\leq C_{\max} \varepsilon^2 \|R\|_F^2 \\ &= C_{\max} \varepsilon^2 \|QR\|_F^2 \\ &\lesssim \frac{Md + \log(1/\delta)}{n} \varepsilon^2 \end{aligned} \quad (\text{using (16)}) \quad (19)$$

Step 4: Finishing the proof. We have the following inequalities,

$$\begin{aligned} &\frac{\gamma}{2M} \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_{\Sigma_{\text{sum}}}^2 \\ &\leq -\langle \nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*), QR \rangle \quad (\text{using (11)}) \\ &= -\langle \nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*), \bar{Q}R \rangle - \langle \nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*), (\bar{Q} - Q)R \rangle \\ &\lesssim \sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} \|\bar{Q}R\|_{\Sigma_{\text{sum}}} + \|\nabla l_{\mathcal{D}}(\boldsymbol{\theta}^*)\|_{\Sigma_{\text{sum}}}^{-1} \|(\bar{Q} - Q)R\|_{\Sigma_{\text{sum}}} \quad (\text{using (18)}) \\ &\lesssim \sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} (\|QR\|_{\Sigma_{\text{sum}}} + \|(\bar{Q} - Q)R\|_{\Sigma_{\text{sum}}}) \\ &\quad + \sqrt{\frac{Md + \log(1/\delta)}{M^2n}} \|(\bar{Q} - Q)R\|_{\Sigma_{\text{sum}}} \quad (\text{using (14)}) \\ &\leq \sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} \|QR\|_{\Sigma_{\text{sum}}} + \sqrt{\frac{Md + \log(1/\delta')}{M^2n}} \|(\bar{Q} - Q)R\|_{\Sigma_{\text{sum}}} \quad (\text{using } r < d, \delta' \leq \delta) \\ &\lesssim \sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} \|QR\|_{\Sigma_{\text{sum}}} + \sqrt{\frac{Md + \log(1/\delta')}{M^2n}} \sqrt{\frac{Md + \log(1/\delta)}{n}} \varepsilon \quad (\text{using (19)}) \\ &\leq \sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} \|QR\|_{\Sigma_{\text{sum}}} + \frac{Md + \log(1/\delta')}{Mn} \varepsilon. \quad (\text{using } \delta' \leq \delta) \end{aligned}$$

Thus, we have

$$\|QR\|_{\Sigma_{\text{sum}}}^2 \lesssim M \left(\sqrt{\frac{Mr + \log(1/\delta')}{M^2n}} \|QR\|_{\Sigma_{\text{sum}}} + \frac{Md + \log(1/\delta')}{Mn} \varepsilon \right).$$

Therefore, we let $\varepsilon = r/d$ and recall that $\delta' = \frac{\delta}{20(\frac{6\sqrt{2}r}{\varepsilon})^{2rd}}$, it follows that

$$\begin{aligned}\|QR\|_{\Sigma_{\text{sum}}} &\lesssim \max\left(\sqrt{\frac{Mr + \log(1/\delta')}{n}}, \sqrt{\frac{Md + \log(1/\delta')}{n}}\varepsilon\right) \\ &\leq \sqrt{\frac{Mr + \log(1/\delta')}{n}} \\ &= \sqrt{\frac{Mr + rd \log(r/\varepsilon) + \log(1/\delta)}{n}} \\ &\leq \sqrt{\frac{Mr + rd \log d + \log(1/\delta)}{n}}.\end{aligned}$$

Finally, notice that

$$\|QR\|_{\Sigma_{\text{sum}}}^2 = \sum_{m=1}^M (\widehat{\theta}_m - \theta_m^*)^\top \Sigma_m (\widehat{\theta}_m - \theta_m^*) = \frac{1}{n} \sum_{m=1}^M \|X_m \widehat{U} \widehat{\alpha}_m - X_m U^* \alpha_m^*\|^2,$$

and the high-probability results used in (9), (16), (18), we obtain the needed inequality with a failure probability at most $\frac{\delta}{10} + \frac{\delta}{20} + \frac{\delta}{20} = \frac{\delta}{5}$. This concludes our proof. $\blacksquare$

Theorem B.3 (Guarantee on target data). *Suppose Assumption 3.1, 3.2 hold. If the number of samples satisfies $n \gg d + \log(M/\delta)$ for $\delta \in (0, 1)$, then with probability at least $1 - \delta$, for any $m \in [M]$,*

$$\|\widehat{\theta}_m - \theta_m^*\|_{\Sigma_m} = \|\widehat{U} \widehat{\alpha}_m - U^* \alpha_m^*\|_{\Sigma_m} \lesssim \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}}.$$

Proof. We assume Lemma B.1 and Theorem B.2 are true, which happens with probability at least $1 - 2\delta/5$.

Step 1: Finding a rotation of $\widehat{U}$ near U^* . Theorem B.2 implies that

$$\sum_{m=1}^M (\widehat{\theta}_m - \theta_m^*)^\top \Sigma_m (\widehat{\theta}_m - \theta_m^*) \lesssim \frac{Mr + rd \log d + \log(1/\delta)}{n}.$$

Thus, Lemma B.1 further shows that

$$\sum_{m=1}^M (\widehat{\theta}_m - \theta_m^*)^\top \Sigma_* (\widehat{\theta}_m - \theta_m^*) \lesssim \frac{Mr + rd \log d + \log(1/\delta)}{n}.$$

Therefore, combining this with Assumption 3.2 implies that

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq \frac{Mr + rd \log d + \log(1/\delta)}{n}.$$

Then, we bound the singular values of $\Theta^* = U^* \alpha^*$. Note that α^* is a $r \times M$ matrix, by definition, we have the smallest singular value $\nu = \sigma_r(\frac{\Theta^{*\top} \Theta^*}{M}) = \sigma_r(\frac{\alpha^{*\top} \alpha^*}{M})$, and the worst-case condition

number as $\kappa = \sigma_1(\frac{\Theta^{*\top}\Theta^*}{M})/\nu = \sigma_1(\frac{\alpha^{*\top}\alpha^*}{M})/\nu$. Note that when Θ is well-conditioned in the sense that $\nu > 0$ and $\kappa \leq O(1)$, the assumption $\|\alpha_m^*\| = \Theta(1)$ implies that $\nu \geq \Omega(\frac{1}{r})$ [Tripuraneni et al., 2020]. Therefore, we have the following inequality:

$$\sqrt{\frac{\sigma_1(\Theta^{*\top}\Theta^*)}{\sigma_r^2(\Theta^{*\top}\Theta^*)}} = \frac{\sqrt{M\nu\kappa}}{M\nu} \leq \sqrt{\frac{r}{M}}.$$

Notice that $\widehat{U}$ and U^* are orthogonal. By applying the general Davis-Kahan $\sin \theta$ theorem [see, Theorem 4, Yu et al., 2014] for matrices $\widehat{\Theta}$ and Θ^* , it follows that, there exists an orthogonal matrix $\widehat{O} \in \mathbb{R}^{r \times r}$ such that

$$\begin{aligned} \|\widehat{U}\widehat{O} - U^*\|_F &\lesssim \sqrt{\frac{\sigma_1(\Theta^{*\top}\Theta^*)}{\sigma_r^2(\Theta^{*\top}\Theta^*)}} \min(r^{1/2}\|\widehat{\Theta} - \Theta^*\|, \|\widehat{\Theta} - \Theta^*\|_F) \\ &\lesssim \sqrt{\frac{r}{M}} \sqrt{\frac{Mr + rd \log d + \log(1/\delta)}{n}} \\ &= \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(1/\delta)}{Mn}}. \end{aligned} \quad (20)$$

Step 2: Bounding $\widehat{\alpha}_m$. For simplicity, suppose $\|\widehat{U} - U^*\|_F \lesssim \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(1/\delta)}{Mn}}$. We can rotate $\widehat{U}$ to achieve this. Then $\widehat{\alpha}_m$ is obtained by minimizing the following loss function:

$$l_{\mathcal{D}_m}(\alpha_m) = -\frac{1}{n} \sum_{i=1}^n \log[\mathbf{1}(y_m^i = 1) \frac{1}{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle) + 1} + \mathbf{1}(y_m^i = 0) \frac{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle)}{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle) + 1}].$$

Its first and second derivatives are given by,

$$\begin{aligned} \nabla l_{\mathcal{D}_m}(\alpha_m) &= -\frac{1}{n} \sum_{i=1}^n \left[\mathbf{1}(y_m^i = 1) \frac{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle)}{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle) + 1} - \mathbf{1}(y_m^i = 0) \frac{1}{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle) + 1} \right] \widehat{U}^\top X_{m,i}, \\ \nabla^2 l_{\mathcal{D}_m}(\alpha_m) &= \frac{1}{n} \sum_{i=1}^n \frac{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle)}{(\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle) + 1)^2} \cdot \widehat{U}^\top X_{m,i} X_{m,i}^\top \widehat{U}. \end{aligned}$$

Similar to (12) in the proof of Theorem B.2, we have the following bound

$$\|\widehat{\alpha}_m - \alpha_m^*\|_{\widehat{U}^\top \Sigma_m \widehat{U}} \lesssim \|\nabla l_{\mathcal{D}_m}(\alpha_m^*)\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}}. \quad (21)$$

Here, the gradient is given by $\nabla l_{\mathcal{D}_m}(\alpha_m^*) = -\frac{1}{n} \widehat{U}^\top X_m^\top \widehat{V}_m$, where $\widehat{V}_{m,i} = V_{m,i} + \Delta_i$, $V_{m,i}$ is defined in (13), and Δ_i is defined as follows:

$$\Delta_i = f_i(\widehat{U}) - f_i(U^*), \quad f_i(U) = \frac{\exp(-\langle U\alpha_m^*, X_{m,i} \rangle)}{\exp(-\langle U\alpha_m^*, X_{m,i} \rangle) + 1}.$$

We aim to bound Δ_i by considering the gradient of $f_i(U)$:

$$\nabla f_i(U) = \frac{\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle)}{(\exp(-\langle \widehat{U}\alpha_m, X_{m,i} \rangle) + 1)^2} \cdot X_{m,i} \alpha_m^\top$$

Notice that $\|X_{m,i}\|_2 \leq B$ and $\|\alpha_m\|_2 = \Theta(1)$, therefore $\|X_{m,i}\alpha_m^\top\|_F^2 = \|X_{m,i}\|_2^2 \|\alpha_m\|_2^2$ is bounded, implying that the gradient $\nabla f_i(U)$ is also bounded. Thus,

$$|\Delta_i| = |f_i(\widehat{U}) - f_i(U^*)| \leq \|\widehat{U} - U^*\|_F \|X_{m,i}\alpha_m^\top\|_F \lesssim \|\widehat{U} - U^*\|_F. \quad (22)$$

Therefore, it follows that

$$\begin{aligned} \|\nabla l_{\mathcal{D}_m}(\alpha_m^*)\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} &= \left\| -\frac{1}{n} \widehat{U}^\top X_m^\top \widehat{V}_m \right\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} \\ &= \left\| -\frac{1}{n} \widehat{U}^\top X_m^\top V_m \right\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} + \left\| -\frac{1}{n} \widehat{U}^\top X_m^\top \Delta \right\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}}. \end{aligned}$$

Regarding the first term, note that $c_{\min} I_d \leq \Sigma_m \leq c_{\max} I_d$, we have

$$\left\| -\frac{1}{n} \widehat{U}^\top X_m^\top V_m \right\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} \lesssim \left\| \frac{1}{n} \widehat{U}^\top X_m^\top V_m \right\|_2,$$

and thus

$$\begin{aligned} &\left\| -\frac{1}{n} \widehat{U}^\top X_m^\top V_m \right\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} \\ &\lesssim \left\| \frac{1}{n} \widehat{U}^\top X_m^\top V_m \right\|_2 \\ &\leq \frac{1}{n} \|U^{*\top} X_m^\top V_m\|_2 + \frac{1}{n} \|(\widehat{U} - U^*)^\top X_m^\top V_m\|_2, \\ &\leq \frac{1}{n} \|U^{*\top} X_m^\top V_m\|_2 + \frac{1}{n} \|\widehat{U} - U^*\|_F \|X_m^\top V_m\|_2 \\ &\lesssim \sqrt{\frac{r + \log(1/\delta)}{n}} + \|\widehat{U} - U^*\|_F \sqrt{\frac{d + \log(1/\delta)}{n}} \quad (\text{By Theorem 2.1 in [Hsu et al., 2012]}) \\ &\lesssim \sqrt{\frac{r + \log(1/\delta)}{n}} + \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(1/\delta)}{Mn}} \sqrt{\frac{d + \log(1/\delta)}{n}} \end{aligned}$$

Under the sample-size condition $n \gg d + \log(M/\delta)$, it is dominated by the first term.

Regarding the second term, denote $\tilde{X}_m = X_m \widehat{U}$, it follows that

$$\begin{aligned} \left\| -\frac{1}{n} \widehat{U}^\top X_m^\top \Delta \right\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} &= \sqrt{\frac{1}{n} \Delta^\top \tilde{X}_m^\top (\tilde{X}_m^\top \tilde{X}_m)^{-1} \tilde{X}_m \Delta} \\ &\leq \frac{1}{\sqrt{n}} \|\Delta\|_2 \\ &\leq \max_i |\Delta_i| \\ &\lesssim \|\widehat{U} - U^*\|_F \quad (\text{using (22)}) \end{aligned}$$

Thus, (21) implies that

$$\|\widehat{\alpha}_m - \alpha_m^*\|_{\widehat{U}^\top \Sigma_m \widehat{U}} \lesssim \|\nabla l_{\mathcal{D}_m}(\alpha_m^*)\|_{(\widehat{U}^\top \Sigma_m \widehat{U})^{-1}} \lesssim \sqrt{\frac{r + \log(1/\delta)}{n}} + \|\widehat{U} - U^*\|_F. \quad (23)$$

Step 3: Finishing the proof. Therefore, we have the following inequalities,

$$\|\widehat{U} \widehat{\alpha}_m - U^* \alpha_m^*\|_{\Sigma_m} \leq \|\widehat{U} \widehat{\alpha}_m - \widehat{U} \alpha_m^*\|_{\Sigma_m} + \|\widehat{U} \alpha_m^* - U^* \alpha_m^*\|_{\Sigma_m}$$

$$\begin{aligned}
&= \|\widehat{\alpha}_m - \alpha_m^*\|_{\widehat{U}^\top \Sigma_m \widehat{U}} + \sqrt{\alpha_m^{*\top} (\widehat{U} - U^*)^\top \Sigma_m (\widehat{U} - U^*) \alpha_m^*} \\
&\lesssim \|\widehat{\alpha}_m - \alpha_m^*\|_{\widehat{U}^\top \Sigma_m \widehat{U}} + \|\alpha_m^*\|_2 \|(\widehat{U} - U^*)^\top \Sigma_m (\widehat{U} - U^*)\|^{1/2} && \text{(using (9))} \\
&\lesssim \|\widehat{\alpha}_m - \alpha_m^*\|_{\widehat{U}^\top \Sigma_m \widehat{U}} + \|\widehat{U} - U^*\|_F && \text{(using Assumption 3.2)} \\
&\lesssim \sqrt{\frac{r + \log(1/\delta)}{n}} + \|\widehat{U} - U^*\|_F && \text{(using (23))} \\
&\lesssim \sqrt{\frac{r + \log(1/\delta)}{n}} + \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(1/\delta)}{Mn}} && \text{(using (20))} \\
&\lesssim \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(1/\delta) \right)}
\end{aligned}$$

Notice that the high-probability results are used in Lemma B.1, Theorem B.2, and (23), thus we obtain the needed inequality with a failure probability at most $\frac{2\delta}{5} + \frac{\delta}{20} \leq \frac{3\delta}{5}$ for a fixed m . Finally, applying a union bound for $\forall m \in [M]$ and substituting δ with δ/M finish the proof. ■

B.2 Proof of Theorem 3.4

Lemma B.4. *If $B \geq x_i, y_i \geq \varepsilon > 0$ for $\forall i \in [M]$, then*

$$|(x_1 \cdots x_M)^{1/M} - (y_1 \cdots y_M)^{1/M}| \leq \frac{B}{M\varepsilon} \sum_{m=1}^M |x_m - y_m|.$$

Proof. Without loss of generality, we assume that $y_1 \cdots y_M \geq x_1 \cdots x_M$.

$$\begin{aligned}
&(y_1 \cdots y_M)^{1/M} - (x_1 \cdots x_M)^{1/M} \\
&= (x_1 \cdots x_M)^{1/M} \left\{ \left(\left(1 + \frac{y_1 - x_1}{x_1}\right) \cdots \left(1 + \frac{y_M - x_M}{x_M}\right) \right)^{1/M} - 1 \right\} \\
&\leq (x_1 \cdots x_M)^{1/M} \left\{ \left(\left(1 + \frac{y_1 - x_1}{x_1}\right) + \cdots + \left(1 + \frac{y_M - x_M}{x_M}\right) \right) / M - 1 \right\} && \text{(using mean inequality)} \\
&\leq \frac{B}{M} \left(\frac{y_1 - x_1}{x_1} + \cdots + \frac{y_M - x_M}{x_M} \right) && \text{(using } x_i \leq B \text{)} \\
&\leq \frac{B}{M} \left(\frac{|y_1 - x_1|}{x_1} + \cdots + \frac{|y_M - x_M|}{x_M} \right) \\
&\leq \frac{B}{M} \left(\frac{|y_1 - x_1|}{\varepsilon} + \cdots + \frac{|y_M - x_M|}{\varepsilon} \right) && \text{(using } x_i \geq \varepsilon \text{)} \\
&= \frac{B}{M\varepsilon} \sum_{m=1}^M |x_m - y_m|.
\end{aligned}$$

We conclude the proof. ■

We then provided the detailed proof of Theorem 3.4.

Proof. Decompose the sub-optimality into three terms:

$$J(\pi^*) - J(\widehat{\pi}) = (J(\pi^*) - \widehat{J}(\pi^*)) + (\widehat{J}(\pi^*) - \widehat{J}(\widehat{\pi})) + (\widehat{J}(\widehat{\pi}) - J(\widehat{\pi})).$$

The second term, $\widehat{J}(\pi^*) - \widehat{J}(\widehat{\pi}) \leq 0$, holds directly by the definition of $\widehat{\pi}$, as $\widehat{\pi}$ maximizes $\widehat{J}$. From Theorem 3.3, with probability at least $1 - \delta$, $\theta_m^* \in \Theta_B(\widehat{\theta}_m)$ for $\forall m$. This allows us to bound the third term as follows:

$$\widehat{J}(\widehat{\pi}) - J(\widehat{\pi}) = \min_{\Theta \in \Theta_B(\widehat{\Theta})} \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \widehat{\pi}(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi}(s))]^{1/M} \leq 0.$$

Thus we can focus on the first term. Let $(\tilde{\theta}_1, \dots, (\tilde{\theta}_M) = \arg \min_{\Theta \in \Theta_B(\widehat{\Theta})} \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi^*(s))]^{1/M}$, which represents the pessimistic parameter chosen from lower confidence bound for the optimal policy π^* .

From the assumption, we have $\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s)) \geq r_0$. Besides, with probability at least $1 - \delta$, we have $\tilde{\theta}_m, \theta_m^* \in \Theta_B(\widehat{\theta}_m)$ for $\forall m$. Using this, we can compare the reward functions evaluated at $\tilde{\theta}_m$ and θ_m^* :

$$\begin{aligned} |R_{m, \tilde{\theta}_m}(s, \pi^*(s)) - R_{m, \theta_m^*}(s, \pi^*(s))| &= |(\theta_m^* - \tilde{\theta}_m)^\top \phi(s, \pi^*(s))| \\ &\leq \|\theta_m^* - \tilde{\theta}_m\|_2 \|\phi(s, \pi^*(s))\|_2 \\ &\lesssim \|\theta_m^* - \tilde{\theta}_m\|_{\Sigma_m} \|\phi(s, \pi^*(s))\|_2 \\ &\lesssim \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}}. \end{aligned}$$

The last inequality holds based on the fact that $\tilde{\theta}_m \in \Theta_B(\widehat{\theta}_m)$ and $\theta_m^* \in \Theta_B(\widehat{\theta}_m)$ with probability at least $1 - \delta$ for $\forall m$, as well as the boundedness of ϕ . Thus, for sufficiently large n , we also have $\mathbb{E}_{s \sim \rho} R_{m, \tilde{\theta}_m}(s, \pi^*(s)) \geq r_0/2$. Therefore, with probability at least $1 - \delta$, we can conclude that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &\leq J(\pi^*) - \widehat{J}(\pi^*) \\ &= \max_{\Theta \in \Theta_B(\widehat{\Theta})} \left(\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi^*(s))]^{1/M} \right) \\ &= \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \tilde{\theta}_m}(s, \pi^*(s))]^{1/M} \\ &\lesssim \frac{1}{M} \sum_{m=1}^M |\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s)) - \mathbb{E}_{s \sim \rho} R_{m, \tilde{\theta}_m}(s, \pi^*(s))| \quad (\text{using Lemma B.4}) \\ &= \frac{1}{M} \sum_{m=1}^M |(\theta_m^* - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))|. \end{aligned}$$

Separately, for each m we have

$$|(\theta_m^* - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))|$$

$$\begin{aligned}
&= |(\theta_m^* - \widehat{\theta}_m + \widehat{\theta}_m - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))| \\
&\leq |(\theta_m^* - \widehat{\theta}_m)^\top \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))| + |(\widehat{\theta}_m - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))| \\
&\leq \|\theta_m^* - \widehat{\theta}_m\|_{\Sigma_m} \|\mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))\|_{\Sigma_m^{-1}} + \|\widehat{\theta}_m - \tilde{\theta}_m\|_{\Sigma_m} \|\mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))\|_{\Sigma_m^{-1}} \quad (\text{Cauchy-Schwartz inequality}) \\
&\lesssim \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)} \|(\Sigma_m^{-1/2} \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s)))\|_2
\end{aligned}$$

The last inequality holds because $\tilde{\theta}_m \in \Theta_B(\widehat{\theta}_m)$ and $\theta_m^* \in \Theta_B(\widehat{\theta}_m)$ with probability at least $1 - \delta$ for $\forall m$. Therefore, we conclude that

$$J(\pi^*) - J(\widehat{\pi}) \lesssim \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)} C^*.$$

where $C^* = \max_m \|\Sigma_m^{-1/2} \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))\|_2$. ■

Alternatively, one can consider another logarithm form of the Nash welfare function as in [Ju et al. \[2024\]](#).

B.3 Proof of Theorem 3.5

Proof. The proof is adapted from Theorem 3.10 of [Zhu et al. \[2023\]](#). In this context, we consider $\tilde{\phi} = U^{*\top} \phi \in \mathbb{R}^r$ for $U^* = (\mathbf{I}_r \quad \mathbf{0})^\top \in \mathcal{O}_{d \times r}$, and the original feature ϕ can be directly reconstructed from $\tilde{\phi}$. Without loss of generality, assume that $r/3$ is some integer. We set $\mathcal{S} = \{1, \dots, r/3\}$, $\mathcal{A} = \{a_1, a_2, a_3\}$, and $\rho = \text{Unif}([1, 2, \dots, |\mathcal{S}|])$. The feature functions and the number of observations for each state-action pair are defined as follows:

$$\begin{aligned}
\tilde{\phi}(s, a_1) &= e_{3s-1}, \quad \tilde{\phi}(s, a_2) = e_{3s-2}, \quad \tilde{\phi}(s, a_3) = 0, \\
n(s, a_1, a_3) &= (n/|\mathcal{S}|)(1 - 2/C^2), \quad n(s, a_2, a_3) = (n/|\mathcal{S}|)(2/C^2).
\end{aligned}$$

Let $v_{-1} = (1/r, 1/r + \Delta, -2/r - \Delta)$, $v_1 = (1/r + 2\Delta, 1/r + \Delta, -2/r - 3\Delta)$ where $\Delta = \mathcal{C}\sqrt{|\mathcal{S}|/n}$. We construct $2^{|\mathcal{S}|}$ instances, indexed by $\tau \in \{-1, 1\}^{|\mathcal{S}|}$, where each $\alpha_\tau^* = (v_{\tau_1}, \dots, v_{\tau_{|\mathcal{S}|}})^\top$. Under each α_τ , the optimal policy $\pi^*(s)$ is either a_1 or a_2 . It can be verified that $\|\Sigma^{-\frac{1}{2}} \mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))\|_2 \leq \mathcal{C}$ and $\|\theta_\tau^*\| = \|\alpha_\tau^*\|$ is bounded, given that $r > 6, n \gtrsim r\mathcal{C}^2, \mathcal{C} \geq 2$. Here $\Sigma^{-\frac{1}{2}}$ is regarded as the Moore-Penrose pseudoinverse, which is well-defined here because $\mathbb{E}_{s \sim \rho} \phi(s, \pi^*(s))$ lies in the column space of Σ . By applying Assouad's lemma as outlined in [Zhu et al. \[2023\]](#), we can conclude that there exists some $\alpha_\tau, \tau \in \{-1, 1\}^{|\mathcal{S}|}$ that achieves the lower bound for the case of $M = 1$.

Now we construct M copies of the above contextual bandit instance, denoted as $\mathcal{Q}_m, m \in [M]$. For this lower-bound construction, the M party-specific datasets are perfectly coupled. Specifically, we first generate a single dataset $\mathcal{D}$ containing n comparison observations, including their realized preference outcomes, and then set $\mathcal{D}_1 = \dots = \mathcal{D}_M = \mathcal{D}$. Consequently, the learner observes only n distinct random comparisons rather than Mn . This coupled data-generation paradigm is used exclusively for the lower-bound proof; throughout the rest of the paper, the party-specific datasets are generated independently. We have shown that $C^* \leq \mathcal{C}$ and $\|\theta_m^*\| = \|\alpha_m^*\|$ are bounded. Thus, we can define an instance $\mathcal{Q}(M) \in \text{CB}(M, \mathcal{C})$ where the multi-party instance

$\mathcal{Q}(M) = \mathcal{Q}_1 \times \cdots \times \mathcal{Q}_M$ represents the combination of each individual instance for a single party. For any $\widehat{\pi}$, we have

$$\begin{aligned} \text{SubOpt}_{\mathcal{Q}(M)}(\widehat{\pi}) &= \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi}(s))]^{1/M} \\ &= \mathbb{E}_{s \sim \rho} R_{\theta^*}(s, \pi^*(s)) - \mathbb{E}_{s \sim \rho} R_{\theta^*}(s, \widehat{\pi}(s)) \gtrsim C \sqrt{\frac{r}{n}}. \end{aligned}$$

which is exactly the result we aim to achieve. To ensure that the reward exceeds r_0 , we can add one additional dimension to both the feature $\tilde{\phi}$ and the parameter α . ■

B.4 Proof of Theorem 3.7

Proof. For simplicity, we will use $R(s, \pi)$ to denote $R(s, \pi(s))$. Consider another policy $\tilde{\pi}$ which satisfies the condition $\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \tilde{\pi}) \geq \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})$ for $\forall m < M$, then

$$\begin{aligned} & [\mathbb{E}_{s \sim \rho} R_{M, \theta_M^*}(s, \tilde{\pi})]^{1/M} \\ &= \frac{\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \tilde{\pi})]^{1/M}}{\prod_{m=1}^{M-1} [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \tilde{\pi})]^{1/M}} \\ &\leq \frac{\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*)]^{1/M}}{\prod_{m=1}^{M-1} [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})]^{1/M}} \quad (\text{using optimality of } \pi^* \text{ and relationship of } \tilde{\pi}, \widehat{\pi}) \\ &\leq \frac{\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*)]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})]^{1/M} + \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})]^{1/M}}{\prod_{m=1}^{M-1} [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})]^{1/M}} \\ &= [\mathbb{E}_{s \sim \rho} R_{M, \theta_M^*}(s, \widehat{\pi})]^{1/M} \left(1 + \frac{\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*)]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})]^{1/M}}{\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \widehat{\pi})]^{1/M}} \right) \\ &\leq [\mathbb{E}_{s \sim \rho} R_{M, \theta_M^*}(s, \widehat{\pi})]^{1/M} \left(1 + \frac{J(\pi^*) - J(\widehat{\pi})}{J(\widehat{\pi})} \right) \\ &\leq [\mathbb{E}_{s \sim \rho} R_{M, \theta_M^*}(s, \widehat{\pi})]^{1/M} \left(1 + \frac{J(\pi^*) - J(\widehat{\pi})}{J(\pi^*) - (J(\pi^*) - J(\widehat{\pi}))} \right) \\ &\leq [\mathbb{E}_{s \sim \rho} R_{M, \theta_M^*}(s, \widehat{\pi})]^{1/M} \left(1 + \frac{K \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right) C^*}}{r_0/2} \right). \end{aligned}$$

The last inequality holds due to Theorem 3.4 and the fact that $J(\pi^*) \geq r_0$, for sufficiently large n . By raising both sides to the power of M , we obtained the desired conclusion. ■

We further show that $\widehat{\pi}$ satisfies the approximate Pigou-Dalton principle, which is an approximation of the traditional Pigou-Dalton principle [Moulin \[2004\]](#).

Definition B.5 (τ -approximate Pigou-Dalton Principle). A solution π satisfies the τ -approximate Pigou-Dalton principle, if for any policy π' such that

$$- \mathbb{E}_{s \sim \rho} R_{i, \theta_i^*}(s, \pi'(s)) + \mathbb{E}_{s \sim \rho} R_{j, \theta_j^*}(s, \pi'(s)) = \mathbb{E}_{s \sim \rho} R_{i, \theta_i^*}(s, \pi(s)) + \mathbb{E}_{s \sim \rho} R_{j, \theta_j^*}(s, \pi(s)),$$

$$\begin{aligned}
& - |\mathbb{E}_{s \sim \rho} R_{i, \theta_i^*}(s, \pi'(s)) - \mathbb{E}_{s \sim \rho} R_{j, \theta_j^*}(s, \pi'(s))| = |\mathbb{E}_{s \sim \rho} R_{i, \theta_i^*}(s, \pi(s)) - \mathbb{E}_{s \sim \rho} R_{j, \theta_j^*}(s, \pi(s))|, \\
& - \mathbb{E}_{s \sim \rho} R_{k, \theta_k^*}(s, \pi'(s)) = \mathbb{E}_{s \sim \rho} R_{k, \theta_k^*}(s, \pi(s)) \text{ for any } k \neq i, j,
\end{aligned}$$

it holds that $J(\pi') - J(\pi) \leq \tau$. Equivalently, the welfare function increases by at most τ in a move reducing the gap (difference) between any two individual values $\mathbb{E}_{s \sim \rho} R_{i, \theta_i^*}, \mathbb{E}_{s \sim \rho} R_{j, \theta_j^*}$ and retaining their total value.

Using the bound on $\prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \widehat{\pi}(s))]^{1/M}$, it directly follows that:

Corollary B.6. *Suppose Assumption 3.1, 3.2 hold. If the number of samples satisfies $n \gg d + \log(M/\delta)$ for $\delta \in (0, 1)$, then with probability at least $1 - \delta$, $\widehat{\pi}$ is τ -approximate Pigou-Dalton principle, where*

$$\tau \lesssim C^* \sqrt{\frac{1}{n} \left(r^2 + \frac{dr^2 \log d}{M} + \log(M/\delta) \right)}.$$

B.5 Proofs in Section 3.4

We begin with the **Utilitarian** welfare function. The proof itself is straightforward to add up individual errors for each party. Here we mainly demonstrate the most significant difference of the proof.

Proof. Again, we consider only the first term in the expression: $J(\pi^*) - J(\widehat{\pi}) = (J(\pi^*) - \widehat{J}(\pi^*)) + (\widehat{J}(\pi^*) - \widehat{J}(\widehat{\pi})) + (\widehat{J}(\widehat{\pi}) - J(\widehat{\pi}))$ as in the proof of Theorem 3.4. Define $\tilde{\theta}_m$ in a similar manner. We proceed as follows:

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}) & \leq J(\pi^*) - \widehat{J}(\pi^*) \\
& = \max_{\Theta \in \Theta_B(\Theta)} \left\{ \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s)) - \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi^*(s)) \right\} \\
& = \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi^*(s)) - \frac{1}{M} \sum_{m=1}^M \mathbb{E}_{s \sim \rho} R_{m, \tilde{\theta}_m}(s, \pi^*(s)) \\
& \leq \frac{1}{M} \sum_{m=1}^M \left| \mathbb{E}_{s \sim \rho} (R_{m, \theta_m^*}(s, \pi^*(s)) - R_{m, \tilde{\theta}_m}(s, \pi^*(s))) \right| \\
& = \frac{1}{M} \sum_{m=1}^M \left| \mathbb{E}_{s \sim \rho} (\theta_m^* - \tilde{\theta}_m)^\top \phi(s, \pi^*(s)) \right|.
\end{aligned}$$

Thus, the remaining steps follow similarly to the proof of Theorem 3.4. ■

Then, we move to the **Egalitarian** welfare function. We introduce a lemma first.

Lemma B.7. *Suppose $a_1, \dots, a_n$ and $b_1, \dots, b_n$ are two groups of real values, sorted as $a_{(1)} \leq \dots \leq a_{(n)}$ and $b_{(1)} \leq \dots \leq b_{(n)}$, respectively. Then, the following inequality holds:*

$$a_{(i)} - b_{(i)} \leq \max_{1 \leq k \leq n} (a_k - b_k).$$

Proof. Let $a_{(i)} = a_{s_i}, b_{(j)} = b_{t_j}$, where $s_i, t_j \in [n]$. For $i = 1$, we have

$$a_{(1)} - b_{(1)} = a_{s_1} - b_{t_1} = (a_{s_1} - a_{t_1}) + (a_{t_1} - b_{t_1}) \leq 0 + \max_k (a_k - b_k) = \max_k (a_k - b_k).$$

We complete the case for $i = 1$. In general, for any $i \in [n]$, we consider two cases:

Case 1: If $a_{s_i} \leq a_{t_i}$, we have

$$a_{(i)} - b_{(i)} = a_{s_i} - b_{t_i} = (a_{s_i} - a_{t_i}) + (a_{t_i} - b_{t_i}) \leq 0 + \max_k (a_k - b_k) = \max_k (a_k - b_k).$$

Case 2: Otherwise $a_{s_i} > a_{t_i}$. We know that there are only $i - 1$ values smaller than a_{s_i} , i.e., $a_{(1)}, \dots, a_{(i-1)} \leq a_{(i)} = a_{s_i}$. Thus there exists $l \leq i$ such that $a_{t_l} \geq a_{s_i}$. Thus, we have

$$a_{(i)} - b_{(i)} = a_{s_i} - b_{(i)} \leq a_{s_i} - b_{(l)} \leq a_{t_l} - b_{t_l} \leq \max_k (a_k - b_k).$$

We conclude the proof. ■

With such a lemma, the proof of the Egalitarian case is similar to Theorem 3.4. Here we mainly demonstrate the most significant difference of the proof.

Proof. Again, we consider only the first term in the expression: $J(\pi^*) - J(\widehat{\pi}) = (J(\pi^*) - \widehat{J}(\pi^*)) + (\widehat{J}(\pi^*) - \widehat{J}(\widehat{\pi})) + (\widehat{J}(\widehat{\pi}) - J(\widehat{\pi}))$ as in the proof of Theorem 3.4. Define $\tilde{\theta}_m$ in a similar manner. We proceed as follows:

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &\leq J(\pi^*) - \widehat{J}(\pi^*) \\ &= \max_{\Theta \in \Theta_B(\widehat{\Theta})} \left\{ \min_m \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi(s)) - \min_m \mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi(s)) \right\} \\ &= \min_m \mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi(s)) - \min_m \mathbb{E}_{s \sim \rho} R_{m, \tilde{\theta}_m}(s, \pi(s)) \\ &\leq \max_m \left(\mathbb{E}_{s \sim \rho} R_{m, \theta_m^*}(s, \pi(s)) - \mathbb{E}_{s \sim \rho} R_{m, \tilde{\theta}_m}(s, \pi(s)) \right) \quad (\text{using Lemma B.7}) \\ &\leq \max_m \left| \mathbb{E}_{s \sim \rho} (\theta_m^* - \tilde{\theta}_m)^\top \phi(s, \pi^*(s)) \right|. \end{aligned}$$

Thus, the remaining steps follow similarly to the proof of Theorem 3.4. ■

B.6 Proof of Theorem A.3

We begin with an estimation error bound similar to Theorem 3.3.

Theorem B.8. Suppose Assumption 3.1, 3.2 hold, with the feature difference in Σ_* of Assumption 3.2 changed to cumulative difference over the trajectory. If the number of samples satisfies $n \gg d + \log(M/\delta)$ for $\delta \in (0, 1)$, then with probability at least $1 - \delta$, for any $m \in [M]$,

$$\|\theta_m^* - \widehat{\theta}_m\|_{\Sigma_m} \lesssim \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}}.$$

Proof. This proof is the same as Theorem 3.3. In the MDP setting, we just consider $X_{m,i} = \sum_{h=1}^H (\phi(s_{m,h}^i, a_{m,h}^i) - \phi(s_{m,h}^{i'}, a_{m,h}^{i'}))$ with $\|X_{m,i}\|_2 \leq 2BLH$ instead. ■

Then, we give the proof of Theorem A.3.

Proof. Decompose the sub-optimality into three terms:

$$J(\pi^*) - J(\widehat{\pi}) = (J(\pi^*) - \widehat{J}(\pi^*)) + (\widehat{J}(\pi^*) - \widehat{J}(\widehat{\pi})) + (\widehat{J}(\widehat{\pi}) - J(\widehat{\pi})).$$

We can again focus on the first term. Let $(\tilde{\theta}_1, \dots, \tilde{\theta}_M) = \arg \min_{\Theta \in \Theta_B(\widehat{\Theta})} \prod_{m=1}^M [\mathbb{E}_{s \sim \rho} R_{m, \theta_m}(s, \pi^*(s))]^{1/M}$, which represents the pessimistic parameter chosen from lower confidence bound for the optimal policy π^* .

From the assumption, we have $\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \theta_m^*}(s, \pi^*(s)) \geq r_0$. Besides, with probability at least $1 - \delta$, we have $\tilde{\theta}_m, \theta_m^* \in \Theta_B(\widehat{\theta}_m)$ for $\forall m$. Using this, we can compare the reward functions evaluated at $\tilde{\theta}_m$ and θ_m^* :

$$\begin{aligned} |R_{m, \tilde{\theta}_m}(s, \pi^*(s)) - R_{m, \theta_m^*}(s, \pi^*(s))| &= |(\theta_m^* - \tilde{\theta}_m)^\top \phi(s, \pi^*(s))| \\ &\leq \|\theta_m^* - \tilde{\theta}_m\|_2 \|\phi(s, \pi^*(s))\|_2 \\ &\lesssim 1/H \|\theta_m^* - \tilde{\theta}_m\|_{\Sigma_m} \|\phi(s, \pi^*(s))\|_2 \\ &\lesssim 1/H \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}} \end{aligned}$$

The last inequality holds based on the fact that $\tilde{\theta}_m \in \Theta_B(\widehat{\theta}_m)$ and $\theta_m^* \in \Theta_B(\widehat{\theta}_m)$ with probability at least $1 - \delta$ for $\forall m$, as well as the boundedness of ϕ . Thus, for sufficiently large n , we also have that $\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \tilde{\theta}_m}(s, \pi^*(s)) \geq r_0/2$. Therefore, with probability at least $1 - \delta$, we can conclude that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &\leq J(\pi^*) - \widehat{J}(\pi^*) \\ &= \max_{\Theta \in \Theta_B(\widehat{\Theta})} \left(\prod_{m=1}^M [\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \theta_m^*}(s, \pi^*(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \theta_m}(s, \pi^*(s))]^{1/M} \right) \\ &= \prod_{m=1}^M [\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \theta_m^*}(s, \pi^*(s))]^{1/M} - \prod_{m=1}^M [\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \tilde{\theta}_m}(s, \pi^*(s))]^{1/M} \\ &\lesssim \frac{1}{M} \sum_{m=1}^M |\mathbb{E}_{s \sim d^{\pi^*}} R_{m, \theta_m^*}(s, \pi^*(s)) - \mathbb{E}_{s \sim d^{\pi^*}} R_{m, \tilde{\theta}_m}(s, \pi^*(s))| \quad (\text{using Lemma B.4}) \\ &= \frac{1}{M} \sum_{m=1}^M |(\theta_m^* - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))| \end{aligned}$$

Separately, for each m we have

$$\begin{aligned} &|(\theta_m^* - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))| \\ &= |(\theta_m^* - \widehat{\theta}_m + \widehat{\theta}_m - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))| \\ &\leq |(\theta_m^* - \widehat{\theta}_m)^\top \mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))| + |(\widehat{\theta}_m - \tilde{\theta}_m)^\top \mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))| \\ &\leq \|\theta_m^* - \widehat{\theta}_m\|_{\Sigma_m} \|\mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))\|_{\Sigma_m^{-1}} + \|\widehat{\theta}_m - \tilde{\theta}_m\|_{\Sigma_m} \|\mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s))\|_{\Sigma_m^{-1}} \quad (\text{Cauchy-Schwartz inequality}) \\ &\lesssim \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}} \|(\Sigma_m^{-1/2} \mathbb{E}_{s \sim d^{\pi^*}} \phi(s, \pi^*(s)))\|_2 \end{aligned}$$

The last inequality holds because $\tilde{\theta}_m \in \Theta_B(\widehat{\theta}_m)$ and $\theta_m^* \in \Theta_B(\widehat{\theta}_m)$ with probability at least $1 - \delta$ for $\forall m$. Therefore, we conclude that

$$J(\pi^*) - J(\widehat{\pi}) \lesssim \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}} C^*.$$

where $C^* = \max_m \mathbb{E}_{s \sim d^{\pi^*}} \|(\Sigma_m^{-1/2} \phi(s, \pi^*(s)))\|_2$. ■

C Technical Details for Reward-Free Model

C.1 Proof of Theorem 4.1

We begin by introducing two fundamental concentration inequalities that will be useful in our analysis.

Proposition C.1 (McDiarmid inequality, [Vershynin \[2018\]](#)). *Let $X_1, \dots, X_n$ be independent random variables in $\mathcal{X}$, let $g : \mathcal{X} \rightarrow \mathbb{R}$ be a function of $X_1, \dots, X_n$ s.t. $\forall x_1, \dots, x_n, x'_i \in \mathcal{X}, |h(x_1, \dots, x_i, \dots, x_n) - h(x_1, \dots, x'_i, \dots, x_n)| \leq c$. Then for all $\delta > 0$, with probability at least $1 - \delta$, we have*

$$|h(X_1, \dots, X_n) - E(h(X_1, \dots, X_n))| \leq c \sqrt{\frac{n \log(2/\delta)}{2}}.$$

Proposition C.2 (Binomial concentration inequality, [Xie et al. \[2021\]](#)). *Suppose $N \sim \text{Ber}(n, p)$, which is a Bernoulli distribution with parameter $n \geq 1$ and $p \in [0, 1]$. Then with probability at least $1 - \delta$, we have*

$$\frac{p}{N \vee 1} \leq \frac{8 \log(1/\delta)}{n}.$$

Lemma C.3. *With probability at least $1 - \delta$,*

$$(1) |N_{s,m}(a, a') - N_{s,m}^*(a, a')| \leq \sqrt{\frac{2 \log(2MSA^2/\delta)}{n_0(m, a, a', s) \vee 1}}, \text{ for } \forall m, s, a, a';$$

$$(2) |N_s(a, a') - N_s^*(a, a')| \leq \sqrt{\frac{2 \log(2MSA^2/\delta)}{\min_m n_0(m, a, a', s) \vee 1}}, \text{ for } \forall s, a, a'.$$

Proof. (1) For fixed m, s, a, a' , define $n' = n_0(m, a, a', s)$. The cases where $n' = 0$ is trivial. For $n' \geq 1$, let $X_{m,i}$ for $i \in [n']$ represent all the binary responses corresponding to the pair (s, a, a') in $\mathcal{D}_m$. Let $h(X) = \sum_{i=1}^{n'} X_i / n'$. Since $|X_i - X'_i| / n' \leq 1/n'$, we apply Proposition C.1 and substitute c with $1/n'$, then with probability at least $1 - \delta$, we have

$$\begin{aligned} & \left| \frac{\#_m(a > a'; s)}{\#_m(a > a'; s) + \#_m(a' > a; s)} - \frac{N_{s,m}^*(a, a') + 1}{2} \right| \\ &= \left| \frac{\#_m(a > a'; s)}{n'} - \frac{N_{s,m}^*(a, a') + 1}{2} \right| \\ &= \left| \frac{\#_m(a > a'; s)}{n'} - \mathbb{P}_m(y = 1 | s, a, a') \right| \end{aligned}$$

$$\leq \frac{1}{n'} \sqrt{\frac{n' \log(2/\delta)}{2}}$$

So for fixed m, s, a, a' , with probability at least $1 - \delta$,

$$|N_{s,m}(a, a') - N_{s,m}^*(a, a')| \leq 2 \left| \frac{N_{s,m}(a, a') + 1}{2} - \frac{N_{s,m}^*(a, a') + 1}{2} \right| \leq \sqrt{\frac{2 \log(2/\delta)}{n_0(m, a, a', s)}}.$$

Then with probability at least $1 - MSA^2\delta$,

$$|N_{s,m}(a, a') - N_{s,m}^*(a, a')| \leq \sqrt{\frac{2 \log(2/\delta)}{n_0(m, a, a', s)}} \quad \forall s, m, a, a'.$$

Substitute δ with $\delta/(MSA^2)$, we complete the first part.

(2) With probability at least $1 - \delta$, for $\forall a, a', s$,

$$\begin{aligned} |N_s(a, a') - N_s^*(a, a')| &= \left| \frac{1}{M} \sum_{m=1}^M N_{s,m}(a, a') - \frac{1}{M} \sum_{m=1}^M N_{s,m}^*(a, a') \right| \\ &\leq \frac{1}{M} \sum_{m=1}^M |N_{s,m}(a, a') - N_{s,m}^*(a, a')| \\ &\leq \sqrt{\frac{2 \log(2MSA^2/\delta)}{\min_m n_0(m, a, a', s)}}. \end{aligned}$$

Thus we conclude the proof. ■

Now we present the detailed proof of Theorem 4.1.

Proof. First, from Lemma C.3, with probability at least $1 - \delta/2$,

$$|N_s(a, a') - N_s^*(a, a')| \leq B_s(a, a') = \sqrt{\frac{2 \log(4MSA^2/\delta)}{\min_m n_0(m, a, a', s) \vee 1}}, \quad \forall s, a, a'.$$

We note that $n_0(m, a, a', s) \sim B(n, \rho(s)d_s^g(a, a'))$, then from Proposition C.2, with probability at least $1 - \delta/2$,

$$\frac{\rho(s)d_s^g(a, a')}{n_0(m, a, a', s) \vee 1} \leq \frac{8 \log(2MSA^2/\delta)}{n}, \quad \forall m, s, a, a'. \quad (24)$$

Thus the above inequalities hold simultaneously with probability at least $1 - \delta$. Next, recall that $\widehat{p}_s, \widehat{q}_s = \arg \max_p \min_q p^\top (N_s - B_s) q$ and $p_s^*, q_s^* = \arg \max_p \min_q p^\top N_s^* q$, we have:

$$\begin{aligned} \min_q \widehat{p}_s^\top N_s^* q &\geq \min_q \widehat{p}_s^\top (N_s - B_s) q && \text{(using } N_s^* \geq N_s - B_s) \\ &= \widehat{p}_s^\top (N_s - B_s) \widehat{q}_s \\ &\geq p_s^{*\top} (N_s - B_s) \widehat{q}_s && \text{(using the definition of } \widehat{p}_s) \end{aligned}$$

$$\begin{aligned}
&= p_s^{*\top} N_s^* \widehat{q}_s + p_s^{*\top} (N_s - N_s^*) \widehat{q}_s - p_s^{*\top} B_s \widehat{q}_s \\
&\geq \min_q p_s^{*\top} N_s^* q - 2p_s^{*\top} B_s \widehat{q}_s && \text{(using } N_s^* \geq N_s - B_s) \\
&= -2p_s^{*\top} B_s \widehat{q}_s && \text{(the value of the matrix game } N_s^* = 0) \\
&= -2\mathbb{E}_{\substack{a \sim p_s^* \\ a' \sim \widehat{q}_s}} B_s(a, a') \\
&= -2\mathbb{E}_{\substack{a \sim p_s^* \\ a' \sim \widehat{q}_s}} \sqrt{\frac{2\log(4MSA^2/\delta)}{\min_m n_0(m, a, a', s) \vee 1}} \\
&\geq -2\mathbb{E}_{\substack{a \sim p_s^* \\ a' \sim \widehat{q}_s}} 4\sqrt{\frac{\log^2(4MSA^2/\delta)}{n\rho(s)d_s^g(a, a')}} && \text{(using (24)).}
\end{aligned}$$

To simplify the notation, let $\eta = \log(4MSA^2/\delta)$, and denote $d_s^\pi = d_s^{p_s^*, \widehat{q}_s}$. We have

$$\begin{aligned}
\min_q \widehat{p}_s^\top N_s^* q &\geq -8 \sum_{a, a'} \eta d_s^\pi(a, a') \sqrt{\frac{1}{n\rho(s)d_s^g(a, a')}} \\
&= -8 \sum_{a, a'} \sqrt{d_s^\pi(a, a')} \eta \sqrt{\frac{d_s^\pi(a, a')}{n\rho(s)d_s^g(a, a')}} \\
&\geq -8 \sqrt{\sum_{a, a'} d_s^\pi(a, a') A^2} \eta \sqrt{\frac{d_s^\pi(a, a')}{n\rho(s)d_s^g(a, a')}} && \text{(Cauchy-Schwartz inequality)} \\
&\geq -8A\eta \sqrt{\frac{d_s^\pi(a, a')}{n\rho(s)d_s^g(a, a')}} \\
&\geq -8A\eta \sqrt{\frac{C^*}{n}} := -\varepsilon.
\end{aligned}$$

Thus we conclude that $\widehat{p}_s$ satisfies $\min_q \widehat{p}_s^\top N_s^* q \geq -\varepsilon$, which implies the following equivalent conditions:

$$\begin{aligned}
\sum_{a \in \mathcal{A}} \widehat{p}_s(a) N_s^*(a, a') &\geq -\varepsilon, \quad \forall a', s, \\
\sum_{a \in \mathcal{A}} \widehat{p}_s(a) &= 1, \quad \forall s.
\end{aligned} \tag{25}$$

Our objective is to establish that the probability distribution $\widehat{p}$ over policy space Π satisfies the following conditions:

$$\begin{aligned}
\sum_{\pi \in \Pi} \widehat{p}(\pi) T^*(\pi, \pi') &\geq -\varepsilon, \quad \forall \pi', \\
\sum_{\pi \in \Pi} \widehat{p}(\pi) &= 1.
\end{aligned}$$

Since $\widehat{p}(\pi) = \prod_{s \in \mathcal{S}} \widehat{p}_s(\pi(s))$, it follows straightforwardly that: $\sum_{\pi \in \Pi} \widehat{p}(\pi) = 1$. Besides, notice that

$$\sum_{\pi \in \Pi} \widehat{p}(\pi) T^*(\pi, \pi') = \sum_{\pi \in \Pi} \widehat{p}(\pi) \sum_{s \in \mathcal{S}} \rho_s N_s^*(\pi(s), \pi'(s))$$

$$= \sum_{s \in \mathcal{S}} \rho_s \sum_{\pi \in \Pi} \widehat{p}(\pi) N_s^*(\pi(s), \pi'(s)). \quad (26)$$

We now separate the terms for different states. For simplicity, let $\mathcal{S} = \{s_i\}_{i=1}^S$ and $\widehat{p}_{s_i} = \widehat{p}_i$ for $i \in [S]$. Thus, for each state s_i , we have

$$\begin{aligned} & \sum_{\pi \in \Pi} \widehat{p}(\pi) N_{s_i}^*(\pi(s_i), \pi'(s_i)) \\ &= \sum_{\pi \in \Pi} \prod_{i=1}^S \widehat{p}_i(\pi(s_i)) N_{s_i}^*(\pi(s_i), \pi'(s_i)) \\ &= \sum_{\pi \in \Pi} \left[\prod_{j \neq i} \widehat{p}_j(\pi(s_j)) \right] \widehat{p}_i(\pi(s_i)) N_{s_i}^*(\pi(s_i), \pi'(s_i)) \\ &= \left(\sum_{\substack{\pi(s_j) \in \mathcal{A} \\ \forall j \neq i}} \prod_{j \neq i} \widehat{p}_j(\pi(s_j)) \right) \left(\sum_{\pi(s_i) \in \mathcal{A}} \widehat{p}_i(\pi(s_i)) N_{s_i}^*(\pi(s_i), \pi'(s_i)) \right) \geq -\varepsilon. \end{aligned}$$

The last inequality holds because the first term equals 1, and the second term is greater than or equal to $-\varepsilon$ for all $a' = \pi'(s_i)$ using (25). Therefore, combining with (26) implies that

$$\sum_{\pi \in \Pi} \widehat{p}(\pi) T^*(\pi, \pi') \geq - \sum_s \rho_s \varepsilon = -\varepsilon.$$

Now we conclude the proof. ■

C.2 Proof of Theorem 4.3

This proof follows a result similar to the one presented in Fishburn [1984].

Proof. Suppose $N_{s,m}^*(a, b) \geq 0$ for $\forall m$ and the inequality holds for some specific m , then from transitivity we have $N_{s,m}^*(a, c) \geq N_{s,m}^*(b, c)$ for $\forall c$. Therefore, $N_s^*(a, b) > 0$ and $N_s^*(a, c) \geq N_s^*(b, c)$.

Now, define $\widehat{p}'_s$ as the probability distribution $\widehat{p}_s$ with the following modifications:

$$\widehat{p}'_s(a) = \widehat{p}_s(a) + \widehat{p}_s(b), \quad \widehat{p}'_s(b) = 0, \quad \widehat{p}'_s(c) = \widehat{p}_s(c), c \notin \{a, b\}.$$

It then follows that

$$\begin{aligned} \widehat{p}'_s{}^\top N_s^* \widehat{p}_s &= N_s^*(a, b) [(\widehat{p}_s(a) + \widehat{p}_s(b)) \widehat{p}_s(b) - 0 \cdot \widehat{p}_s(a)] \\ &+ \sum_{c \in \mathcal{S} \setminus \{a, b\}} N_s^*(a, c) [(\widehat{p}_s(a) + \widehat{p}_s(b)) \widehat{p}_s(c) - \widehat{p}_s(c) \widehat{p}_s(a)] \\ &+ \sum_{c \in \mathcal{S} \setminus \{a, b\}} N_s^*(b, c) [0 \cdot \widehat{p}_s(c) - \widehat{p}_s(c) \widehat{p}_s(b)] \\ &= \widehat{p}_s(b) \left[N_s^*(a, b) (\widehat{p}_s(a) + \widehat{p}_s(b)) + \sum_{c \in \mathcal{S} \setminus \{a, b\}} \widehat{p}_s(c) (N_s^*(a, c) - N_s^*(b, c)) \right] \end{aligned}$$

$$\geq \widehat{p}_s(b)^2 N_s^*(a, b). \quad (27)$$

Notice that with probability at least $1 - \delta$, we have

$$\widehat{p}_s'^\top N_s^* \widehat{p}_s = -\widehat{p}_s^\top N_s^* \widehat{p}_s' \leq \varepsilon. \quad (\text{using } \widehat{p}_s^\top N_s^* \widehat{p}_s' \geq \min_q \widehat{p}_s^\top N_s^* q \geq -\varepsilon)$$

where ε is defined in Theorem 4.1. Therefore, (27) implies that $\varepsilon \geq \widehat{p}_s(b)^2 N_s^*(a, b)$, which leads to the following inequality for $\widehat{p}_s(b)$:

$$\begin{aligned} \widehat{p}_s(b) &\leq \sqrt{\varepsilon / N_s^*(a, b)} = \left(8A \log(2MSA^2/\delta) \sqrt{\frac{C^*}{n} / N_s^*(a, b)} \right)^{\frac{1}{2}} \\ &= K \left(A \log(2MSA^2/\delta) \sqrt{\frac{C^*}{n}} \right)^{1/2}, \end{aligned}$$

where K is a constant that depends on $N_s^*(a, b)$. ■

We now proceed to demonstrate the **necessity of transitivity**. Consider a simple example with three alternatives: Suppose that $N_{s,m}^*(a, b) = 1, N_{s,m}^*(b, c) = 1, N_{s,m}^*(c, a) = 1$ for $\forall m$. This means that for each individual, given comparison of (a, b) , $a > b$ with probability 1. The same holds for the other pairs. Therefore, all individuals have deterministic preferences with respect to the three alternatives, and the preference matrix takes the following form:

$$N_s = N_s^* = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}.$$

Since everyone has absolute preferences, there is no need to introduce a confidence bound for the rankings. As a result, the von Neumann winner directly yields the outcome $(1/3, 1/3, 1/3)$. However, as item b is Pareto dominated by item a , the result stated in the previous theorem does not hold anymore, where transitivity is assumed. Without transitivity, it may not have consistent preferences. The transitivity is indeed a necessary condition for the result to hold.

D Experimental Details

This appendix provides additional details for the experiments in Section 5.

D.1 Experimental Details for Section 5.1

D.1.1 Dataset Construction.

We construct a finite contextual-bandit instance designed to illustrate two features of the proposed method. First, the parties have heterogeneous preferences but share a low-dimensional reward representation. Second, the comparison features span a substantially larger ambient space, so the learner must identify the reward-relevant subspace from the observations. Specifically, we consider 3 states, 9 actions, and $M = 24$ parties, with ambient dimension $d = 24$ and latent dimension $r = 3$. The state distribution ρ is uniform over $\mathcal{S} = \{s_1, s_2, s_3\}$, and $\mathcal{A} = \{a_1, \dots, a_9\}$. We write $[q] = \{1, \dots, q\}$.

The true parameter matrix satisfies

$$\Theta^* = (\theta_1^*, \dots, \theta_M^*) = U^* \alpha^*, \quad \alpha^* = (\alpha_1^*, \dots, \alpha_M^*) \in \mathbb{R}^{3 \times 24},$$

where

$$U^* = \begin{pmatrix} I_3 \\ 0_{21 \times 3} \end{pmatrix} \in \mathcal{O}_{24 \times 3}.$$

Thus, the true party parameters lie in the three-dimensional subspace spanned by the first three ambient coordinates. The representation U^* is used only to generate the data and is unavailable to the learner.

We divide the parties into three equally sized clusters. For $j \in [3]$, let

$$G_j = \{8(j-1) + 1, \dots, 8j\}.$$

Each cluster is associated with one of the three latent reward directions. For every $j \in [3]$ and $k \in [4]$, we independently draw

$$\zeta_{j,k} \sim \text{Unif}(0, 0.05).$$

We assign to the eight parties in G_j the antithetic perturbations

$$\{\varepsilon_m : m \in G_j\} = \{\zeta_{j,1}, \dots, \zeta_{j,4}, -\zeta_{j,1}, \dots, -\zeta_{j,4}\},$$

and define

$$\alpha_m^* = (1 + \varepsilon_m)e_j, \quad \theta_m^* = U^* \alpha_m^*, \quad m \in G_j.$$

The continuous perturbations make the parties within each cluster distinct almost surely, while their antithetic structure ensures

$$\sum_{m \in G_j} \varepsilon_m = 0, \quad \frac{1}{|G_j|} \sum_{m \in G_j} \alpha_m^* = e_j.$$

Thus, the construction introduces party-level heterogeneity without shifting the three cluster centers. Because these centers span the three latent directions, the resulting party-parameter matrix has rank 3. The parameters are generated once and held fixed across all repetitions, so only the sampled comparison labels vary between repetitions.

We next construct the feature vectors. Write $\phi(s, a) = (\phi_{1:3}(s, a), \phi_{4:24}(s, a))$, where the first three coordinates determine the true rewards. They are given by

action	$\phi_{1:3}(s_1, a)$	$\phi_{1:3}(s_2, a) = \phi_{1:3}(s_3, a)$
a_1	$(7/2, -1/4, -3/4)^\top$	$(13/4, -1/2, -1/2)^\top$
a_2	$(-1/2, 3, -1/2)^\top$	$(-1/2, 3, -1/2)^\top$
a_3	$(-1/2, -1/2, 3)^\top$	$(-1/2, -1/2, 3)^\top$
a_4	$(3/4, 3/4, 1/2)^\top$	$(3/4, 3/4, 1/2)^\top$
$a_5, \dots, a_8$	$(0, 0, 0)^\top$	$(0, 0, 0)^\top$
a_9	$(-1, -1, -1)^\top$	$(-1, -1, -1)^\top$

This construction gives the actions distinct roles. Actions a_1, a_2 , and a_3 are specialist actions favored by clusters G_1, G_2 , and G_3 , respectively, whereas a_4 provides a compromise that assigns positive reward to all three clusters. Actions $a_5, \dots, a_8$ carry no true reward signal, and a_9

serves as the common comparison reference. The construction therefore creates a transparent collective-choice problem in which policies must trade off the interests of the three party clusters.

We use the remaining 21 coordinates to provide ambient feature coverage. Let $e_t^{(24)}$ denote the t -th standard basis vector in $\mathbb{R}^{24}$, and write

$$\bar{\phi}_{\ell,h} = \phi_{1:3}(s_\ell, a_h).$$

Define

$$\mathcal{N} = \{(1, 4)\} \cup \left(\{2, 3\} \times [4]\right) \cup \left([3] \times \{5, 6, 7, 8\}\right),$$

and, for $(\ell, h) \in \mathcal{N}$, let

$$c(\ell, h) = \begin{cases} 4, & (\ell, h) = (1, 4), \\ 4 + 4(\ell - 2) + h, & \ell \in \{2, 3\}, h \in [4], \\ 12 + 4(\ell - 1) + (h - 4), & \ell \in [3], h \in \{5, 6, 7, 8\}. \end{cases}$$

Set

$$\psi_{\ell,h} = \begin{cases} 3e_{c(\ell,h)}^{(24)}, & (\ell, h) \in \mathcal{N}, \\ 0_{24}, & \text{otherwise,} \end{cases}$$

and define the complete feature vector as

$$\phi(s_\ell, a_h) = \begin{pmatrix} \bar{\phi}_{\ell,h} \\ 0_{21} \end{pmatrix} + \psi_{\ell,h}.$$

For example, (s_1, a_4) receives nuisance coordinate 4; $(s_2, a_1), \dots, (s_2, a_4)$ receive coordinates 5, ..., 8; $(s_3, a_1), \dots, (s_3, a_4)$ receive coordinates 9, ..., 12; and the state-action pairs involving $a_5, \dots, a_8$ receive coordinates 13, ..., 24. All other state-action pairs have zero nuisance coordinates.

Because every θ_m^* is supported on the first three coordinates, these nuisance features do not affect any party's true reward. Their role is statistical: they provide 21 mutually orthogonal comparison directions and make the ambient comparison covariance full rank. An unrestricted estimator must therefore operate in all 24 dimensions, whereas the joint estimator can reduce estimation error by learning that the party rewards share a three-dimensional representation.

This construction satisfies the assumptions used in our analysis. In particular, the party-parameter matrix has rank 3, the comparison covariance is positive definite, and the three party clusters provide nondegenerate coverage of all latent directions. Moreover, we take the latent preference parameter norm bound $B = 1.06$ and normalization constant $C_0 = 8.5$, the parameter and feature bounds satisfy

$$C_0 - B \max_{s,a} \|\phi(s, a)\|_2 > 0,$$

so all values entering the Nash welfare objective and its pessimistic counterpart remain positive. The concentratability coefficient is also finite because the comparison distribution covers the feature directions used by the optimal policy.

Building on the above construction, we now describe data collection process. For every state, each action a_h , $h \in [8]$, is compared against the reference action a_9 . This produces the 24 comparison cells

$$\{(s_\ell, a_h, a_9) : \ell \in [3], h \in [8]\},$$

and each party's comparisons are allocated as evenly as possible across these cells. Let $X_{\ell,h} = \phi(s_\ell, a_h) - \phi(s_\ell, a_9)$. The preference label is generated according to the BTL model in Section 3, with winning probability $\Phi(\langle \theta_m^*, X_{\ell,h} \rangle)$.

The horizontal-axis quantity n in Figure 2a denotes the minimum per-party sample size. Within each cluster G_j , random half of the parties provide n comparisons each, while other parties provide $2n$ comparisons each. The total number of comparisons across all cells is $36n$. We use $n = 150, 300, \dots, 3000$ and conduct 100 independent repetitions at every sample size.

D.1.2 Learning methods and baselines.

Our method applies the joint low-rank MLE followed by the pessimistic Nash welfare optimization in Algorithm 1. It jointly estimates the shared representation $\widehat{U}$ and the party-specific coefficients $\{\widehat{\alpha}_m\}_{m=1}^M$, subject to

$$U^\top U = I_r, \quad \|U\alpha_m\|_2 \leq B, \quad B = 1.06.$$

The numerical procedure alternates between updating the shared representation and updating the party-specific coefficients, using three initializations.

We compare our method with three plug-in baselines. The independent baseline $\widehat{\pi}_{\text{ind}}$ estimates one unrestricted d -dimensional parameter for each party and therefore isolates the benefit of learning a shared low-dimensional representation. The remaining two baselines discard party identities, pool the observations, and fit a single reward parameter, but differ in how they weight the observations. Let $\ell_m(\theta)$ denote the unnormalized BTL negative log-likelihood for party m . The pooling-obs baseline $\widehat{\pi}_{\text{pool-obs}}$ solves

$$\min_{\|\theta\|_2 \leq B} \sum_{m=1}^M \ell_m(\theta),$$

so each comparison observation receives equal weight and parties with more comparisons contribute more strongly. In contrast, the pooling-party baseline $\widehat{\pi}_{\text{pool-party}}$ solves

$$\min_{\|\theta\|_2 \leq B} \frac{1}{M} \sum_{m=1}^M \frac{\ell_m(\theta)}{n_m},$$

so every party receives equal total weight regardless of its sample size. Comparing these two pooling baselines separates the effect of unequal observation counts from the more fundamental loss of information caused by pooling heterogeneous party preferences.

All policies are deterministic. We enumerate the full policy class $\Pi = \mathcal{A}^S$ with $|\Pi| = 9^3 = 729$, and define $\mu_\pi = \mathbb{E}_{s \sim \rho}[\phi(s, \pi(s))]$. For party m , let

$$\Sigma_m = \frac{1}{n_m} \sum_{i=1}^{n_m} X_{mi} X_{mi}^\top.$$

The confidence set used by our method is

$$\Theta_B(\widehat{\theta}_m) = \left\{ \theta : \|\theta\|_2 \leq B, \|\theta - \widehat{\theta}_m\|_{\Sigma_m} \leq \Gamma \right\},$$

where

$$\Gamma = K \sqrt{\frac{r^2}{n} + \frac{dr^2 \log d + r \log(M/\delta)}{Mn}},$$

and we set $K = 2.6, \delta = 0.05$. We shall validate empirically in D.1.4 that this choice of K provides consistently high marginal and simultaneous empirical coverage across the reported sample-size range. For every candidate policy, the implementation computes the pessimistic value of each party by solving

$$\min_{\theta \in \Theta_B(\hat{\theta}_m)} C_0 + \mu_\pi^\top \theta.$$

It then selects the policy that maximizes the geometric mean of these party-specific worst-case values. The optimization is performed over the exact intersection of the confidence ellipsoid and the parameter ball. The true parameters are used only to evaluate the resulting policies and are never provided to the learning methods.

D.1.3 Evaluation Metric.

Let $\hat{\pi}$, $\hat{\pi}_{\text{ind}}$, $\hat{\pi}_{\text{pool-obs}}$, and $\hat{\pi}_{\text{pool-party}}$ denote the policies produced by our method, the independent baseline, and the two pooling baselines, respectively. For a deterministic policy π , define party m 's true value as

$$V_m(\pi) = \mathbb{E}_{s \sim \rho}[R_{m, \theta_m^*}(s, \pi(s))].$$

We evaluate the Nash, Utilitarian, and Egalitarian welfare functions

$$J_N(\pi) = \left(\prod_{m=1}^M V_m(\pi) \right)^{1/M}, \quad J_U(\pi) = \frac{1}{M} \sum_{m=1}^M V_m(\pi), \quad J_E(\pi) = \min_{m \in [M]} V_m(\pi).$$

For $F \in \{N, U, E\}$, define the corresponding welfare suboptimality by

$$\text{SubOpt}_F(\hat{\pi}) = \max_{\pi \in \Pi} J_F(\pi) - J_F(\hat{\pi}).$$

The left panel of Figure 2a reports the mean Nash welfare suboptimality of our method and the three baselines, together with 95% Monte Carlo confidence intervals over the 100 repetitions. The right panel evaluates the policies produced by our method and pooling-obs under the Utilitarian and Egalitarian objectives. These policies remain Nash-targeted and are not re-optimized for the alternative welfare objectives. For each $F \in \{U, E\}$, the reported quantity is

$$\frac{\mathbb{E}_{\text{rep}}[\text{SubOpt}_F(\hat{\pi})]}{\mathbb{E}_{\text{rep}}[\text{SubOpt}_F(\hat{\pi}_{\text{pool-obs}})]}.$$

This is the ratio of the mean suboptimalities across repetitions.

We additionally measure the dispersion of party-level regret. For any learned policy $\tilde{\pi}$, define party m 's average classical regret as

$$\overline{\text{Regret}}_m(\tilde{\pi}) = \mathbb{E}_{\text{rep}} \left[\max_{\pi \in \Pi} V_m(\pi) - V_m(\tilde{\pi}) \right].$$

The mean pairwise difference in average party regret is

$$D(\tilde{\pi}) = \frac{1}{M(M-1)} \sum_{m \neq m'} |\overline{\text{Regret}}_m(\tilde{\pi}) - \overline{\text{Regret}}_{m'}(\tilde{\pi})|.$$

The Difference curve in the right panel reports $\frac{D(\hat{\pi})}{D(\hat{\pi}_{\text{pool-obs}})}$, where a ratio below one means that our method produces a smaller mean pairwise disparity in parties' average regrets than pooling-obs. This metric describes how evenly regret is distributed across parties in this experiment, serving as a partial evidence of fairness.

D.1.4 Confidence-set coverage.

For party m , confidence-set coverage is the empirical probability of the event $\|\hat{\theta}_m - \theta_m^*\|_{\Sigma_m} \leq \Gamma$. Marginal coverage averages this event over parties and repetitions, whereas simultaneous coverage requires the event to hold for all 24 parties in the same repetition. The latter is the more demanding criterion and corresponds more closely to the joint event required when optimizing a collective policy. Figure 3 reports both measures for $K \in \{2, 2.4, 2.6, 3\}$.

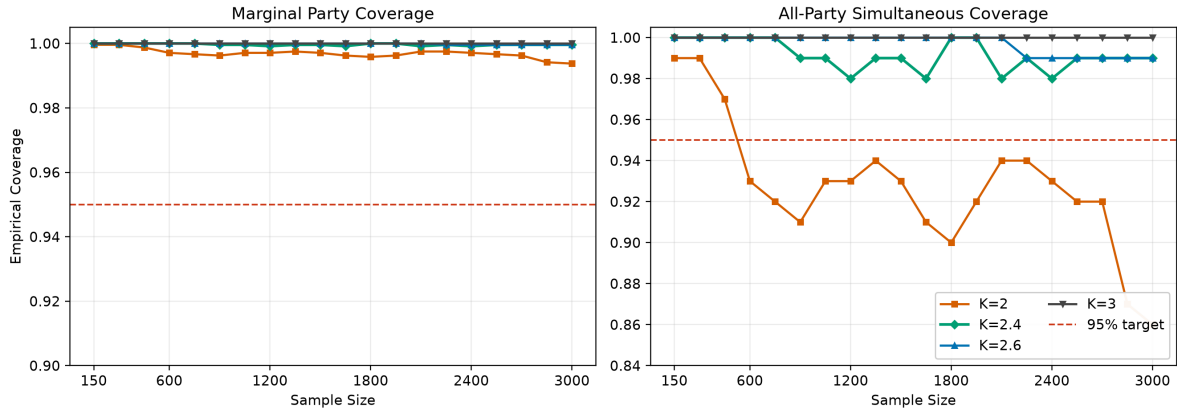


Figure 3: Empirical confidence-set coverage. The left panel reports marginal coverage over parties and repetitions, while the right panel reports simultaneous coverage over all 24 parties.

For the main choice $K = 2.4$, the aggregate marginal coverage is 99.96%, while simultaneous coverage averaged over the sample-size grid is 99.15%. The minimum simultaneous coverage at an individual sample size is 98%. Thus, $K = 2.4$ provides consistently high marginal and simultaneous empirical coverage across the reported sample-size range under the current experimental setup.

D.1.5 Estimation accuracy and efficiency.

To separate the benefit of shared-representation learning from the effect of pessimistic policy selection, we compare the joint and independent estimators using

$$\mathcal{E}_{\Sigma_*}(\hat{\Theta}) = \left\{ \frac{1}{M} \sum_{m=1}^M \|\hat{\theta}_m - \theta_m^*\|_{\Sigma_*}^2 \right\}^{1/2}.$$

The fixed population covariance Σ_* is used only for this diagnostic; the pessimistic policy optimization continues to use the party-specific empirical covariance matrices Σ_m .

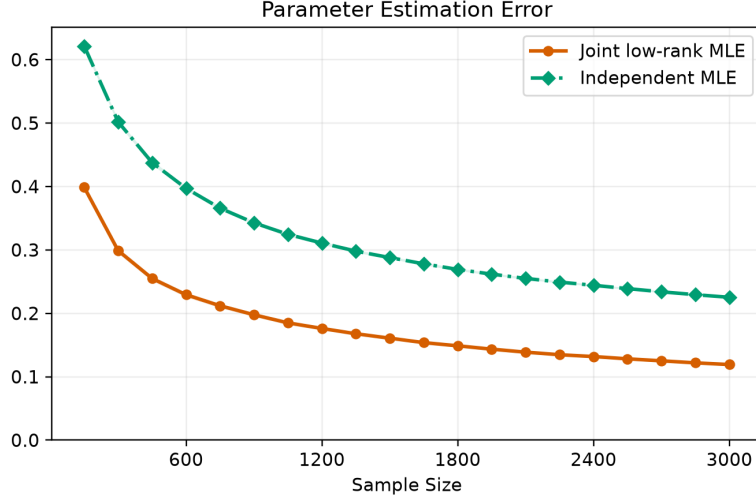


Figure 4: Population-covariance-weighted estimation error of the joint low-rank and independent MLEs.

The joint estimator has lower estimation error than the independent estimator throughout the reported sample-size range. At $n = 150$, its mean error is 0.399, compared with 0.622 for the independent estimator. At $n = 3000$, the corresponding errors are 0.119 and 0.225. Both estimators improve as the sample size increases, but the persistent gap shows the statistical benefit of estimating the shared three-dimensional representation rather than learning 24 coordinates separately for every party.

D.2 Further Details for the General Pairwise-Preference Experiment (Section 5.2)

This section provides implementation and diagnostic details for the general pairwise-preference experiment reported in Figure 2b. We use a separate one-state setting. Although we use a BTL model to generate a controlled population preference matrix, the proposed method and the pooling baseline observe only pairwise comparison outcomes. Denote the state by s , let $\mathcal{A} = \{a_1, a_2, a_3\}$, and set $M = 2$. The feature vectors and ground-truth parameters are

$$\phi(a_1) = (1, 0)^\top, \quad \phi(a_2) = (0, 1)^\top, \quad \phi(a_3) = (0, 0)^\top, \quad \theta_1^* = (1, 0.1), \quad \theta_2^* = (0.2, 1).$$

Thus, $R_{\theta_m^*}(s, a) = \langle \theta_m^*, \phi(a) \rangle$. We use the BTL model only to generate a controlled population preference matrix:

$$\mathbb{P}_m(y = 1 \mid a, a', s) = \Phi(R_{\theta_m^*}(s, a) - R_{\theta_m^*}(s, a')), \quad \Phi(x) = \frac{1}{1 + \exp(-x)}.$$

Neither the proposed method nor the pooling baseline observes the rewards, features, or parameters. Following Section 4, the party-specific population preference matrix is

$$N_{s,m}^*(a, a') := 2\mathbb{P}_m(y = 1 \mid a, a', s) - 1, \quad N_{s,m}^*(a, a) = 0,$$

and the equally weighted population target is $N_s^* := \frac{1}{M} \sum_{m=1}^M N_{s,m}^* = (N_{s,1}^* + N_{s,2}^*)/2$.

Sampling and estimation. The minimal party sample size is $n \in \{400, 800, \dots, 4000\}$. We allocate $|\mathcal{D}_1| = n$ comparisons to Party 1 and $|\mathcal{D}_2| = 2n$ to Party 2, and distribute each party's comparisons evenly among the three unordered action pairs. Each setting is repeated 500 times. Using the notation in Section 4.1, let

$$n_0(m, a, a', s) := \#_m(a > a'; s) + \#_m(a' > a; s)$$

and define

$$N_{s,m}(a, a') := \frac{\#_m(a > a'; s) - \#_m(a' > a; s)}{n_0(m, a, a', s)}.$$

We then form the party-balanced matrix and its confidence bonus as

$$N_s(a, a') = \frac{1}{M} \sum_{m=1}^M N_{s,m}(a, a'), \quad B_s(a, a') = \sqrt{\frac{2 \log(4MSA^2/\delta)}{\min_m n_0(m, a, a', s) \vee 1}},$$

for $a \neq a'$, with zero diagonal and $\delta = 0.05$. Here $S = 1$ and $A = 3$. We implement Algorithm 2 exactly by solving

$$\widehat{p}_s, \widehat{q}_s = \arg \max_{p \in \Delta(A)} \min_{q \in \Delta(A)} p^\top (N_s - B_s) q.$$

The observation-pooling baseline first pools the two parties' comparison counts to construct $N_{s,\text{pool}}$, and then solves the corresponding matrix game without a confidence bonus. Both methods are evaluated on the same simulated data in each repetition.

Evaluation and the pooling limit. Because N_s^* is skew-symmetric, its game value is zero. We therefore report the nonnegative game-value suboptimality

$$- \min_{q \in \Delta(A)} \widehat{p}_s^\top N_s^* q.$$

Figure 2b shows its mean and 95% Monte Carlo confidence interval over the 500 repetitions. For this instance,

$$N_s^* \approx \begin{pmatrix} 0 & 0.02098 & 0.28089 \\ -0.02098 & 0 & 0.25604 \\ -0.28089 & -0.25604 & 0 \end{pmatrix}.$$

Hence a_1 is the Condorcet winner of N_s^* : it has positive population preference margin against every other action. In contrast, because Party 2 provides twice as many comparisons, observation pooling converges to $(N_{s,1}^* + 2N_{s,2}^*)/3$, whose Condorcet winner is a_2 . Selecting a_2 has suboptimality 0.02098 under the equally weighted target N_s^* , which explains the non-vanishing pooling curve. Algorithm 2 preserves the intended equal party weights by normalizing within each party before aggregation.

Lower-bound coverage. We additionally examine whether $N_s - cB_s$ forms a valid simultaneous lower confidence bound for the population preference matrix. For each $c \in \{0.2, 0.3, \dots, 1.0\}$, we test whether

$$N_s(a, a') - cB_s(a, a') \leq N_s^*(a, a')$$

holds simultaneously for all $a \neq a'$. We remark that the multiplier c is not an algorithmic hyperparameter: Algorithm 2 and the main experiment use $N_s - B_s$, corresponding to $c = 1$.

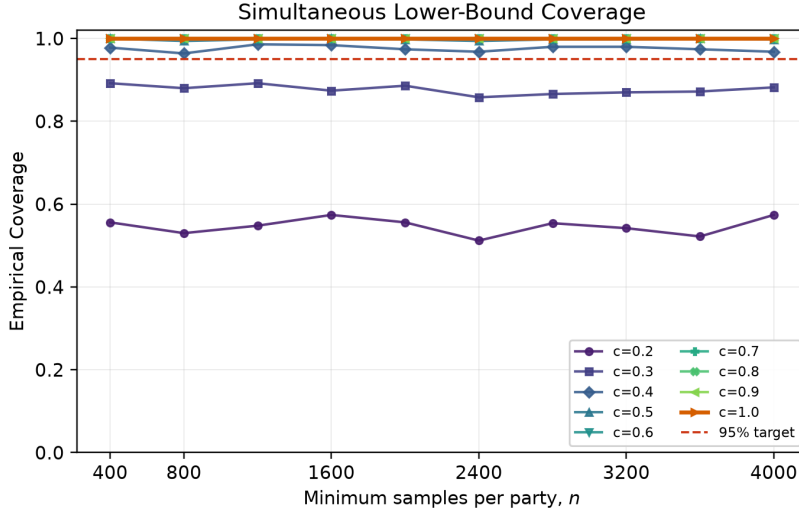


Figure 5: Empirical simultaneous lower-bound coverage for the general pairwise-preference experiment. Each curve corresponds to a diagnostic multiplier c on B_s , and the dashed line marks the nominal $1 - \delta = 95\%$ level. Algorithm 2 and the main experiment use the full confidence bonus ($c = 1$).

Because N_s and N_s^* are skew-symmetric and B_s is symmetric, this is equivalent to simultaneous two-sided coverage over the three unordered pairs. Figure 5 shows that the full bonus covers in every repetition at every sample size. Coverage is at least 99.4% for $c = 0.5$ and falls below the nominal 95% level for $c \leq 0.3$, indicating that the theoretical bonus is conservative on this instance.