Provably Efficient Personalized Multi-Objective Bandits with Proactive Conversational Queries

Linfeng Cao[1] **Ming Shi**[2] **Ness B. Shroff**[1,3]

[1]Department of Computer Science and Engineering, The Ohio State University
[2]Department of Electrical Engineering, University at Buffalo
[3]Department of Electrical and Computer Engineering, The Ohio State University

Abstract

Personalized decision-making in multi-objective bandits requires learning user-specific trade-offs among competing objectives. Since arm utility depends on both unknown rewards and unknown preferences, existing methods infer preferences only from utility feedback, entangling preference learning with reward exploration. In practice, however, users often reveal their priorities through proactive conversational queries (e.g., "cheap and clean hotel"), yet this structured signal is not leveraged. We formalize a proactive query-based framework in which user queries provide structured preference signals. Modeling these signals via a Plackett–Luce subset choice model, we show that query-only learning is insufficient due to a fundamental shift-invariance barrier. To resolve this, we introduce MO-PQUCB, a hybrid algorithm that integrates query-based preference anchoring with bandit feedback through shift-invariant regularization and dual-exploration UCB. We prove that proactive queries accelerate preference estimation and yield improved regret scaling over prior preference-aware MO-MAB methods. Under corrupted queries, we further characterize statistical limits and design a robust estimator achieving near-optimal performance when the corruption is sparse. Experiments validate both theoretical and practical gains.

1 INTRODUCTION

Multi-objective decision-making arises in many applications such as personalized recommendation, dialogue planning and so on, where competing objectives (e.g., cost, quality, latency) must be balanced. The multi-objective multi-armed bandit (MO-MAB) framework [Drugan and Nowe, 2013] provides a principled model for sequential optimization under uncertainty. However, most existing MO-MAB methods focus on identifying Pareto-optimal arms [Drugan and Nowe, 2013, Turgay et al., 2018, Lu et al., 2019, Drugan, 2018, Balef and Maghsudi, 2023] without accounting for user-specific trade-offs, leading to one-size-fits-all solutions.

Preference-aware MO-MAB formulations address this limitation by modeling user preferences explicitly [Cao et al., 2025]. In these models, each user is associated with a latent preference vector over objectives, and utility is defined as the inner product between preferences and objective rewards. However, existing methods estimate preferences solely from bandit feedback, coupling reward exploration with preference learning and limiting statistical efficiency.

In practical interactive systems, users often reveal high-level priorities through proactive conversational queries (e.g., "cheap and clean hotel", Fig.1), which provide structured signals about objective trade-offs [Rui et al., 2025, Dao et al., 2024, Zhang, 2023]. Nevertheless, such signals are not explicitly leveraged in existing MO-MAB frameworks. In this work, we formalize a proactive query-based preference elicitation framework for MO-MAB. We model user queries as top-m objective rankings under a Plackett–Luce subset choice model parameterized by the user's latent preference. We first show that query-only preference learning suffers from a fundamental shift-invariance identifiability barrier, making additional feedback necessary for optimal decision-making. To overcome this limitation, we propose **MO-PQUCB**, a hybrid algorithm that integrates query-based preference anchoring with bandit feedback via shift-invariant regularization and a dual-exploration UCB strategy. This design leverages structured query feedback to accelerate preference estimation while retaining principled exploration of objective rewards.

Our contributions are summarized as follows:

- **Framework.** We introduce a proactive preference-aware MO-MAB formulation that incorporates structured query feedback. Our algorithm integrates query-based prefer-

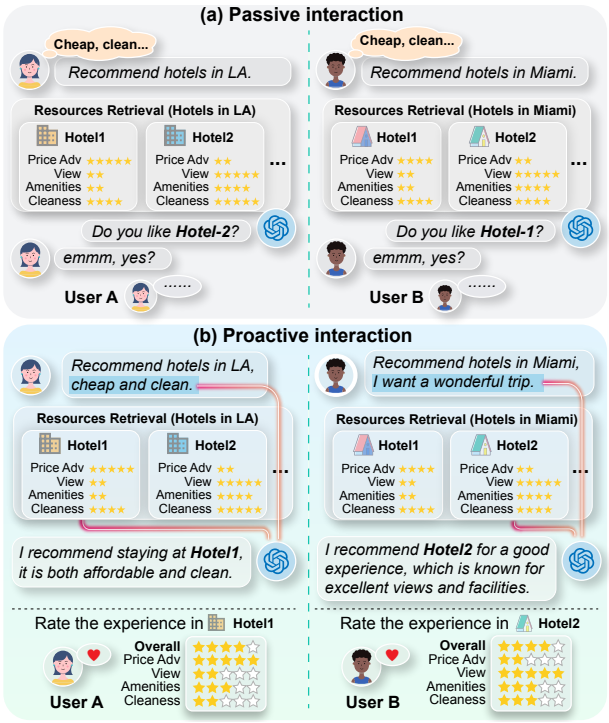


Figure 1: Illustration of interaction styles in conversational recommendation. (a) Passive interaction: the system proposes items or keywords and receives binary feedback, providing limited information about user trade-offs. (b) Proactive interaction: the user expresses high-level priorities in the initial query, revealing structured preference signals that enable more informative objective trade-off inference.

ence estimation with bandit learning to shrink confidence regions over personalized utilities.

- **Theory.** We characterize the statistical benefit of proactive queries and prove that MO-PQUCB achieves near-optimal regret, improving regret scaling of $\sqrt{\log T}$ over prior preference-aware MO-MAB methods.
- **Robustness.** We analyze learning under corrupted queries, establish a fundamental lower bound identifying the learnability regime, and design a robust estimator achieving near-optimal performance under sparse corruption.
- **Empirical validation.** Experiments on synthetic and real-world datasets demonstrate improved personalization and robustness, and integration with large language models highlights practical applicability.

2 RELATED WORK

The multi-objective multi-armed bandit (MO-MAB) framework extends classical bandits to vector-valued rewards. Early work primarily focused on Pareto-optimal arm identification and Pareto regret minimization [Drugan and Nowe, 2013, Turgay et al., 2018, Lu et al., 2019, Drugan, 2018, Balef and Maghsudi, 2023]. While these methods provide principled global optimization guarantees, they do not account for user-specific trade-offs across objectives.

Preference-aware MO-MAB formulations incorporate user-specific trade-offs across objectives. Some works assume known structured orders (e.g., lexicographic hierarchies) and optimize without learning preferences [Hüyük and Tekin, 2021, Cheng et al., 2024]. Another line of work models preferences as latent vectors and defines utility as the inner product between preference and objective reward vectors, learning them jointly [Cao et al., 2025]. In this regime, preferences are inferred solely from bandit feedback, coupling preference learning with reward exploration.

Conversational and interactive bandit methods leverage user interaction to guide exploration [Christakopoulou et al., 2016, Zhang et al., 2020, Xie et al., 2021, Zhao et al., 2022]. However, most existing frameworks are system-driven, relying on keyword prompts or binary feedback, which provide limited structural information about objective trade-offs. In contrast, our method supports proactive user-initiated queries beyond system-driven binary interaction and achieves improved regret scaling. We leave the detailed discussion and systematic comparison including regret bound with these methods in Appendix A.

3 PROBLEM FORMULATION

3.1 PREFERENCE-AWARE MO-MAB

We adopt the preference-aware MO-MAB framework proposed in Cao et al. [2025]. Specifically, we consider a multi-objective multi-armed bandit (MO-MAB) setting with K arms, N users, and D objective dimensions. At each round $t \in [T]$, user $n \in [N]$ is presented with an available arm set $\mathcal{A}_t^n \subseteq [K]$. The learner selects $a_{n,t} \in \mathcal{A}_t^n$ and observes both the objective reward vector $\boldsymbol{r}_{a_{n,t},t} \in [0,1]^D$ and the resulting scalar utility $g_{a_{n,t},t}$.

Objective Rewards. For each arm i, the objective reward vector $\boldsymbol{r}_{i,t}$ is drawn i.i.d. from an unknown distribution supported on $[0,1]^D$, with mean $\boldsymbol{\mu}_i$.

User Preferences. Each user n is associated with an unknown mean preference vector $\overline{\boldsymbol{c}}_n \in \mathbb{R}^D$, indicating preference on D objectives. At each round t, the instantaneous preference vector $\boldsymbol{c}_{n,t} \in [0,B]^D$ is drawn i.i.d. from an unknown distribution supported on $[0,B]^D$ with mean $\overline{\boldsymbol{c}}_n$.

Independence. For all t, n, i, we assume the reward vector $\boldsymbol{r}_{i,t}$ and preference vector $\boldsymbol{c}_{n,t}$ are independent. This is realistic in many applications since $\boldsymbol{c}_t$ and $\boldsymbol{r}_t$ are *inherently driven* by independent factors of user characteristics and arm attributes [Cao et al., 2025]. For instance, an individual user's personal preferences do not affect a hotel's location, environment, or pricing, and vice versa.

Utility. At round t, when arm $a_{n,t}$ is selected for user n, the observed utility is $g^n_{a_{n,t},t} = \langle \boldsymbol{c}_{n,t}, \boldsymbol{r}_{a_{n,t},t} \rangle$. Under the independence assumption, we have $\mathbb{E}[g^n_{i,t}] = \langle \overline{\boldsymbol{c}}_n, \boldsymbol{\mu}_i \rangle$.

Regret. Let $a^*_{n,t} = \arg\max_{i \in \mathcal{A}^n_t} \langle \overline{\boldsymbol{c}}_n, \boldsymbol{\mu}_i \rangle$ denote the optimal arm for user n at round t. We define the cumulative regret over horizon T as

$$R(T) = \sum_{t=1}^{T} \sum_{n=1}^{N} \left\langle \overline{\boldsymbol{c}}_n, \boldsymbol{\mu}_{a^*_{n,t}} - \boldsymbol{\mu}_{a_{n,t}} \right\rangle. \tag{1}$$

The goal is to minimize $R(T)$ by efficiently matching all users with arms that align with their individual preferences.

3.2 PROACTIVE PREFERENCE QUERY ELICITATION (QE)

At each round t, prior to arm selection, each user $n \in [N]$ proactively provides high-level preference feedback via query interactions. We represent this feedback as a top-$m_{n,t}$ ordered subset $S^n_t = [\sigma_{n,t}(1), \ldots, \sigma_{n,t}(m_{n,t})]$, where $m_{n,t} \leq D$ may vary across rounds. The ranking S^n_t indicates the user's most preferred $m_{n,t}$ objectives, ordered from most to least preferred [1]. In practice, such rankings can be extracted from natural-language queries using a language interface (e.g., BERT or GPT). For theoretical analysis, we assume the ranked subset S^n_t is directly observed.

We model the latent full ranking σ_t using the Plackett–Luce (PL) model [Marden, 1996, Sec. 5.6.1] parameterized by the user's latent mean preference vector $\overline{\boldsymbol{c}} \in \mathbb{R}^D$. Under the induced PL subset choice model [Saha and Gopalan, 2019], the probability of observing the top-$m_{n,t}$ ranking S^n_t is

$$\mathbb{P}(S^n_t \mid \overline{\boldsymbol{c}}) = \prod_{j=1}^{m_{n,t}} \frac{\exp(\overline{c}_{S^n_t(j)})}{\sum_{k \in [D] \setminus \{S^n_t(1), \ldots, S^n_t(j-1)\}} \exp(\overline{c}_k)}. \tag{2}$$

This corresponds to sequentially selecting $m_{n,t}$ objectives without replacement under the PL distribution. The model generalizes several classical cases: when $m_{n,t} = 1$, $D = 2$, it reduces to the Bradley–Terry–Luce (BTL) model [Bradley and Terry, 1952]; when $m_{n,t} = D$, it recovers the full ranking PL model [Zhu et al., 2023]. *The top-m formulation reflects practical settings where users disclose only their most prioritized objectives rather than a complete ranking.*

4 WARM UP: STATISTICAL BENEFIT OF PROACTIVE QE

We analyze the statistical benefit of proactive query elicitation for preference estimation under the PL subset model. The analysis reveals a fundamental shift-invariance barrier and motivates our hybrid algorithm design.

[1] Our study does not fundamentally require every round query. The key quantity is the effective query information. If informative queries occur with linear frequency (i.e., $\sum_{t=1}^{T} m_t = \Theta(T)$), the same bound holds up to constants

4.1 QE-BASED PREFERENCE ESTIMATOR

For each user $n \in [N]$, given the observed ranking sequence $\mathcal{S}^n_t = \{S^n_i\}_{i=1}^t$, we estimate the latent preference vector under the PL subset model via classic maximum likelihood estimation (MLE), the classic algorithm in learning to rank and RLHF [Christiano et al., 2017, Ouyang et al., 2022, Wang et al., 2024, Xiong et al., 2024].

Let $\ell_{\mathrm{PL}}(S; \boldsymbol{c})$ denote the negative log-likelihood of a top-m ranking S under the PL model:

$$\ell_{\mathrm{PL}}(S; \boldsymbol{c}) = -\sum_{j=1}^{m} \log \frac{\exp(\boldsymbol{c}_{S(j)})}{\sum_{k \in [D] \setminus \{S(1), \ldots, S(j-1)\}} \exp(\boldsymbol{c}_k)}, \tag{3}$$

and the MLE estimator aims at minimizing

$$\mathcal{L}_{\mathcal{S}^n_t}(\boldsymbol{c}) = \sum_{i=1}^{t} \ell_{\mathrm{PL}}(S^n_i; \boldsymbol{c}).$$

However, the PL model is invariant to additive shifts of the parameter vector [Shah et al., 2016]: for any $a \in \mathbb{R}$, $\boldsymbol{c}$ and $\boldsymbol{c} + a\mathbf{1}$ induce the same distribution. To ensure identifiability, we work with the mean-centered preference vector $\overline{\boldsymbol{c}}^0_n = \overline{\boldsymbol{c}}_n - \frac{\|\overline{\boldsymbol{c}}_n\|_1}{D} \boldsymbol{e}$, and restrict estimation to the zero-mean subspace

$$\Theta^0_C = \left\{ \boldsymbol{c} \in \mathbb{R}^D : \|\boldsymbol{c}\|_\infty \leq B,\ \boldsymbol{c}^\top \mathbf{1} = 0 \right\}.$$

The QE-based estimator of user n is therefore defined as

$$\hat{\boldsymbol{c}}^{(\mathrm{QE})}_{n,t} = \arg\min_{\boldsymbol{c} \in \Theta^0_C} \mathcal{L}_{\mathcal{S}^n_t}(\boldsymbol{c}). \tag{4}$$

The following lemma establishes a convergence guarantee for the QE-based estimator to the mean-centered preference vector $\overline{\boldsymbol{c}}^0_n$. The proof is provided in Appendix D.

Lemma 1. *Let $\overline{m}_{n,t} = \frac{1}{t} \sum_{i=1}^{t} m_{n,i}$ denote the average number of elicited objectives up to round t. For all $t \geq C_0 \log(T)$ with $C_0 = 2D(16e^{2B} + 8)^2$, there exists a constant $C_1 > 0$ such that the QE-based estimator uniformly holds with high probability that:*

$$\|\hat{\boldsymbol{c}}^{(QE)}_{n,t} - \overline{\boldsymbol{c}}^0_n\|_2 \leq C_1 D \sqrt{\frac{\log(t)}{\overline{m}_{n,t} t}}.$$

Insight 1: Efficiency of Proactive QE. Proactive queries yield an estimator with error $\widetilde{O}\big((\overline{m}_{n,t} t)^{-\frac{1}{2}}\big)$ on mean-centered preference, showing that the effective sample size scales with $\overline{m}_{n,t} t$, i.e., richer elicitation improves learning.

4.2 IS QE ALONE SUFFICIENT?

While proactive QE enables statistically efficient estimation of the mean-centered preference vector, an important question remains: does query feedback alone suffice for optimal decision-making in MO-MAB?

We show that the answer is negative. Query-only learning is fundamentally insufficient, as formalized below.

Theorem 1 (Lower Bound for QE-Only Learning)**.** *Consider a preference-aware MO-MAB environment with $D \geq 2$ objectives and at least two Pareto-optimal arms whose objective rewards differ. For any bandit algorithm that relies solely on QE-based preference estimates, there exist latent preference vectors $\{\overline{\boldsymbol{c}}^n\}_{n\in[N]}$ such that $R(T) = \Omega(T)$.*

The proof is deferred to Appendix E. The lower bound arises from the shift-invariance of the PL model. QE-based estimators can only recover preferences up to an additive constant. However, MO-MAB utilities depend on the absolute preference vector: $\langle \boldsymbol{c}, \boldsymbol{\mu}_i \rangle$ and additive shifts modify arm rankings whenever $\sum_d \boldsymbol{\mu}_i(d)$ differs across arms. Thus, indistinguishable preference vectors under QE can induce different optimal arms, leading to linear regret.

Insight 2: Necessity of Additional Feedback. Query-only learning is inherently non-identifiable under shift-invariance, rendering optimal decision-making impossible.

5 MO-PQUCB: LEARNING WITH QUERIES AND BANDIT FEEDBACK

The preceding analysis and insights reveal a structural tension: while QE enables efficient relative preference estimation, it is insufficient for optimal decision-making. We now introduce *MO-PQUCB* (Multi-Objective Proactive Query UCB), a unified framework that integrates proactive query elicitation with bandit feedback, as shown in Algorithm 1.

The central idea is to use QE to anchor the relative preference structure, while leveraging bandit utility observations to calibrate the absolute preference scale and guide exploration. Algorithm 1 consists of two tightly coupled components: (1) a collaborative preference estimator that regularizes bandit-based preference learning using QE anchors, and (2) a preference-aware UCB policy that balances uncertainty in both reward and preference estimation.

5.1 QE-ANCHORED PREFERENCE ESTIMATOR

The QE estimate $\hat{\boldsymbol{c}}_t^{(\mathrm{QE})}$ identifies preferences only up to an additive constant. To recover the full preference vector, we incorporate bandit feedback through a regularized least-squares formulation, using the QE estimate as a structural regularization anchor.

Algorithm 1 MO-PQUCB

Input: $\alpha \in (0,1)$, $\lambda > 0$, $\{\beta_t\}$, ρ
▷ **Initialization:**
for each user $n \in [N]$ **do**
  $\boldsymbol{U}_\lambda \leftarrow (\lambda+1)\boldsymbol{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^\top$;
  $\boldsymbol{V}_{n,0} \leftarrow (1-\alpha)\boldsymbol{U}_\lambda$;
  $\mathcal{S}_0^n \leftarrow \emptyset$
end for
for each arm $i \in [K]$ **do**
  $N_{i,1} \leftarrow 0, \quad \hat{\boldsymbol{r}}_{i,1} \leftarrow \mathbf{0}, \quad \gamma_{i,1} \leftarrow 0$
end for
for $t = 1$ to T **do**
  for each interacting user $n \in [N]$ **do**
    ▷ **Query Phase:**
    Receive user query and obtain ranked subset S_t^n
    $\mathcal{S}_t^n \leftarrow \mathcal{S}_{t-1}^n \cup \{S_t^n\}$
    ▷ **Preference Estimation:**
    Estimate preference anchor $\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{QE})}$ by Eq. (4)
    Update collaborative preference estimate $\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}$ by (5)
    ▷ **Arm Selection:**
    Select $a_{n,t}$ with dual-exploration UCB policy by (8)
    Observe $\boldsymbol{r}_{a_{n,t},t}$ and $g^n_{a_{n,t},t}$
    ▷ **Bandit Sufficient Statistics Update:**
    $\boldsymbol{V}_{n,t+1} \leftarrow \boldsymbol{V}_{n,t} + \alpha\, \boldsymbol{r}_{a_{n,t},t}\boldsymbol{r}_{a_{n,t},t}^\top$
  end for
  ▷ **Arm Statistics Update:**
  for each arm $i \in [K]$ **do**
    Update $N_{i,t+1}$ and $\hat{\boldsymbol{r}}_{i,t+1}$ by (7) and (6)
    $\gamma_{i,t+1} \leftarrow \sqrt{\frac{\log(t/\rho)}{\max\{1, N_{i,t+1}\}}}$
  end for
end for

For each user n, we define the hybrid estimator

$$\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})} = \arg\min_{\boldsymbol{c}\in\mathbb{R}^D} \Big(\alpha \sum_{\ell=1}^{t-1} \big(\langle \boldsymbol{c}, \boldsymbol{r}_{a_\ell^n,\ell}\rangle - g^n_{a_\ell^n,\ell}\big)^2 + (1-\alpha)\big\|\boldsymbol{c} - \hat{\boldsymbol{c}}_{n,t}^{(\mathrm{QE})}\big\|^2_{\boldsymbol{U}_\lambda} \Big).$$

where $\alpha \in (0,1)$ balances the alignment trade-off between bandit driven feedback and QE anchoring, and

$$\boldsymbol{U}_\lambda = \boldsymbol{U} + \lambda \boldsymbol{I}, \quad \boldsymbol{U} = \boldsymbol{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^\top.$$

Here $\boldsymbol{U}$ is the projector onto the subspace orthogonal to $\mathbf{1}$.

The key structural design lies in aligning the regularization geometry with the identifiability structure induced by QE. Since QE identifies preferences only up to an additive shift along $\mathbf{1}$, the projector $\boldsymbol{U}$ restricts regularization to the orthogonal subspace, while leaving the shift direction unconstrained. Consequently, bandit feedback uniquely

determines the previously unidentifiable shift component, leading to consistent recovery of the full preference vector. The optimization admits the closed-form solution:

$$\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})} = \boldsymbol{V}_{n,t}^{-1}\big(\alpha \sum_{\ell=1}^{t-1} g_{a_{n,\ell},\ell}^{n} \boldsymbol{r}_{a_{n,\ell},\ell} + (1-\alpha)\boldsymbol{U}_{\lambda}\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{QE})}\big), \quad (5)$$

where $\boldsymbol{V}_{n,t} = \alpha \sum_{\ell=1}^{t-1} \boldsymbol{r}_{a_{n,\ell},\ell}\boldsymbol{r}_{a_{n,\ell},\ell}^{\top} + (1-\alpha)\boldsymbol{U}_{\lambda}$.

We characterize the estimation error of the QE-anchored hybrid preference estimator as follows.

Lemma 2. *For any user $n \in [N]$, for all $t \geq C_0 \log\left(\frac{T}{\rho}\right)$, the estimation error of the proposed QE-anchored preference estimator (Eq.5) admits a high-probability upper confidence bound with some $C_2 > 0$:*

$$\|\hat{\boldsymbol{c}}_{n,t}^{(HB)} - \overline{\boldsymbol{c}}_n\|_{\boldsymbol{V}_{n,t}} \leq C_2\Bigg(\underbrace{\alpha DB\sqrt{D\log\left(\alpha Dt\right)}}_{\textit{bandit noise term}} + \underbrace{D\sqrt{\frac{(1-\alpha)\log(t)}{\overline{m}_{n,t}t}}}_{\textit{QE anchoring noise term}}\Bigg).$$

The proof is deferred to Appendix G. This result shows that the hybrid estimator improves over prior methods by leveraging proactive QE. The first term, scaled by α, grows as $O(\sqrt{\log t})$, matching the confidence bounds in standard conversational contextual bandits [Zhang et al., 2020, Zhao et al., 2022] and preference-aware MO-MAB [Cao et al., 2025]. In contrast, the QE-based term shrinks at rate $\tilde{O}(1/(\sqrt{\overline{m}_{n,t}t}))$, enabling faster shrinkage of confidence region. The parameter α thus governs the trade-off between proactive QE anchoring and bandit-driven feedback, directly controlling how the algorithm balances exploration and preference regularization. Such trade-off can be optimized by choosing α (and λ) to achieve near-optimal regret (see Corollary 2.1).

5.2 DUAL-EXPLORATION UCB POLICY

Given the hybrid preference estimate $\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}$ for each user n, our goal is to construct an upper confidence bound (UCB) policy for the expected utility $\langle \overline{\boldsymbol{c}}_n, \boldsymbol{\mu}_i \rangle$ of each arm $i \in [K]$.

Dual exploration challenge. Preference-aware MO-MAB exhibits a fundamental *global–local exploration* dilemma [Cao et al., 2025]. For each user n, the learner must simultaneously: **(1)** *Global exploration:* sample diverse arms to ensure that the Gram matrix $\boldsymbol{V}_{n,t}$ is well-conditioned, enabling accurate preference recovery; **(2)** *Local exploration:* repeatedly sample a specific arm to reduce uncertainty in its reward estimate.

Thus, the UCB must explicitly account for uncertainty in both reward estimation and preference estimation. Inspired by the analysis of PRUCB Cao et al. [2025], we adopt their dual-exploration framework as our bandit backbone (i.e., dual-exploration policy eq.(8) with the corresponding reward and preference uncertainties designs).

Objective reward estimation. We estimate the objective reward of each arm $i \in [K]$ via the collaborative empirical mean across all users:

$$\hat{\boldsymbol{r}}_{i,t} = \frac{\sum_{j=1}^{t-1}\sum_{n\in[N]} \boldsymbol{r}_{a_j^n,j}\mathbb{1}_{\{a_j^n=i\}}}{\max\{1, N_{i,t}\}}, \quad (6)$$

$$N_{i,t} = \sum_{j=1}^{t-1}\sum_{n\in[N]} \mathbb{1}_{\{a_j^n=i\}}, \quad (7)$$

where $N_{i,t}$ is the number of times arm i selected up to round $t-1$.

Utility confidence bound. By the concentration of the reward estimator and the confidence region of the hybrid preference estimator, we can then bound the utility estimate as follows. The proof is deferred to Appendix F.

Lemma 3. *For any user $n \in [N]$, any arm $i \in [K]$, with probability at least $1 - \frac{D\rho^2\pi^2}{3}$, it holds uniformly for all $t \geq C_0 \log\left(\frac{T}{\rho}\right)$ that*

$$\langle \overline{\boldsymbol{c}}_n, \boldsymbol{\mu}_i \rangle \leq \langle \hat{\boldsymbol{c}}_{n,t}^{(HB)}, \hat{\boldsymbol{r}}_{i,t} \rangle + B_{i,t}^{n(r)} + B_{i,t}^{n(c)}.$$

Here the two bonus terms are defined as follows.

Reward uncertainty.

$$B_{i,t}^{n(r)} = \gamma_{i,t}\|\hat{\boldsymbol{c}}_{n,t}^{(HB)}\|_1, \quad \gamma_{i,t} = \sqrt{\log(t/\rho)/\max\{1, N_{i,t}\}}.$$

This term captures uncertainty arising from estimation of the arm reward vector.

Preference uncertainty.

$$B_{i,t}^{n(c)} = \beta_t^n \, \|\hat{\boldsymbol{r}}_{i,t} + \gamma_{i,t}\mathbf{1}\|_{\boldsymbol{V}_{n,t}^{-1}},$$

where $\beta_t^n = C_2\left(\alpha\sqrt{D^3\log(\alpha Dt)} + D\sqrt{\frac{(1-\alpha)\log t}{\overline{m}_{n,t}\, t}}\right)$, denotes the radius of the confidence region for $\hat{\boldsymbol{c}}_{n,t}^{(HB)}$ (see Eq. (16) *for the explicit constant). This term captures uncertainty from hybrid preference estimation.*

Arm selection. By the principle of optimism in the face of uncertainty [Auer et al., 2002], the user-specific policy is

$$a_{n,t} = \arg\max_{i\in\mathcal{A}_t^n}\Big(\langle \hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}, \hat{\boldsymbol{r}}_{i,t}\rangle + B_{i,t}^{n(r)} + B_{i,t}^{n(c)}\Big). \quad (8)$$

The two bonus terms explicitly reflect the dual-exploration structure: $B_{i,t}^{n(r)}$ promotes local exploration by reducing arm-level reward uncertainty, while $B_{i,t}^{n(c)}$ encourages global exploration to improve conditioning of the preference estimator. Together, they enable efficient learning under heterogeneous and latent user preferences.

5.3 THEORETICAL GUARANTEES

We state the main regret results in a compact form, highlighting the dependence on key parameters; the full versions and proof are deferred to Appendix H.

Theorem 2 (User-Specific Regret)**.** *For any user $n \in [N]$, with high probability, the regret of MO-PQUCB satisfies*

$$R^n(T) = O\Big(\underbrace{\big(\sqrt{\tfrac{D^3}{\overline{m}_{n,T}}} + BD^{\frac{3}{2}}\sqrt{\alpha \log T}\big)\sqrt{T \log T}}_{\textit{preference estimation error}} + \underbrace{\tfrac{\alpha B D^{3/2}}{\sqrt{\lambda}}\sqrt{KT}\log T}_{\textit{reward estimation error}} \Big).$$

The regret decomposes into two sources of uncertainty: (i) *preference uncertainty*, arising from recovering the latent preference vector, and (ii) *reward uncertainty*, arising from estimating arm utilities. The first group of terms captures the trade-off between QE anchoring and bandit-driven calibration, while the second group reflects multi-armed exploration. The parameter α controls this balance and can be tuned to optimize the overall regret.

Corollary 2.1. *By choosing $\alpha = \lambda = \Theta(\log^{-1}(T))$, all leading multiplicative factors of $\sqrt{T \log T}$ in Theorem 2 remain bounded by constants, yielding*

$$R^n(T) = O\left(\left(\sqrt{\frac{D^3}{\overline{m}_{n,T}}} + BD^{\frac{3}{2}}\sqrt{K}\right)\sqrt{T \log T}\right).$$

Aggregating all users and absorbing the problem-dependent constants of D, K yield the compact final regret

$$R(T) = \sum_{n=1}^{N} R^n(T) = O\left(N\sqrt{T \log T}\right).$$

Discussion. Compared to PRUCB [Cao et al., 2025], which attains $O(N\sqrt{T}\log T)$ regret in preference-aware MO-MAB, our proactive QE-based framework improves the logarithmic dependence on T, achieving $O(N\sqrt{T \log T})$ regret. This improvement stems from structured preference elicitation to reduce uncertainty in preference estimation.

6 LEARNING UNDER CORRUPTED PROACTIVE QUERIES

In practical deployments, proactive query elicitation is often mediated through natural-language interaction with conversational agents (e.g., large language models). Inference errors, contextual ambiguity, or imperfect alignment may cause such systems to output preference rankings that deviate from the user's true intent. To account for this realistic uncertainty, we study a more challenging setting in which the preference feedback obtained from proactive queries is partially corrupted, and analyze the robustness of MO-PQUCB under such perturbations.

6.1 CONTAMINATION MODEL

We model query corruption via an ϵ-*stochastic adversarial contamination model*. At each round t for user n, the observed top-$m_{n,t}$ list $\tilde{S}_t^n$ is generated from a Plackett–Luce (PL) subset choice model (Eq. 2) parameterized by a possibly corrupted preference vector $\tilde{\boldsymbol{c}}_{n,t}$:

With probability $1-\epsilon$, the feedback is clean and $\tilde{\boldsymbol{c}}_{n,t} = \overline{\boldsymbol{c}}_n$. With probability ϵ, the preference is corrupted as

$$\tilde{\boldsymbol{c}}_{n,t} = P_{n,t}\overline{\boldsymbol{c}}_n,$$

where $P_{n,t}$ is an arbitrary permutation matrix over the D objectives at round t for user n. Such corruption relabels objective dimensions, producing rankings misaligned with the true preference vector while preserving preference magnitudes. We allow $\{P_{n,t}\}$ to be chosen arbitrarily and potentially adaptively with respect to past observations. The history of possibly corrupted observations is denoted by $\tilde{\mathcal{S}}_t^n$.

6.2 A GENERAL LOWER BOUND

We first characterize the fundamental statistical limits of preference estimation under ϵ-corrupted PL feedback. The following result reveals a sharp transition at $\epsilon = 1/2$, beyond which consistent estimation becomes impossible.

Theorem 3. *Assume each query reveals the full objective ranking ($m = D$) and L corrupted ranking samples are observed. Then for any estimator $\hat{\boldsymbol{c}}_L$,*

$$\inf_{\hat{\boldsymbol{c}}_L} \sup_{\overline{\boldsymbol{c}}^0 \in \Theta_C^0} \mathbb{E}\big[\|\hat{\boldsymbol{c}}_L - \overline{\boldsymbol{c}}^0\|_2\big] \;\geq\; \begin{cases} \Omega(1), & \epsilon \geq \dfrac{1}{2} - \dfrac{2}{D\sqrt{L}}, \\ \Omega\Big(\dfrac{1}{(1-2\epsilon)\sqrt{DL}}\Big), & \textit{else}. \end{cases}$$

The proof is provided in Appendix I. The lower bound shows that the estimation error grows as the corruption level ϵ increases and diverges as $\epsilon \to 1/2$. This reflects a majority condition: when fewer than half of the samples are clean, the true preference vector becomes statistically indistinguishable from permuted alternatives.

This phenomenon aligns with the impossibility result of Datar et al. [2022] under the BTL model, which requires a majority of clean comparisons. Our bound extends this insight to structured top-m feedback under the PL model and explicitly characterizes the dependence on ϵ, D, and L.

6.3 ROBUST LEARNING WITH GROUP-WISE LASSO

To address corrupted proactive queries, we propose a *group-wise Lasso maximum likelihood estimator* that jointly learns the clean preference vector and identifies corrupted ranking samples. Unlike standard Lasso regularization used in

scalar reward modeling [Bukharin et al., 2024], our method operates on structured top-m rankings under the PL model and enforces group sparsity across samples.

Corruption Parameterization. Under ϵ-corrupted feedback, each observed ranking of user n at round t is generated from $\tilde{\boldsymbol{c}}_{n,t} = \overline{\boldsymbol{c}}_n + \boldsymbol{\delta}^*_{n,t}$, where $\overline{\boldsymbol{c}}$ is the true preference vector and $\boldsymbol{\delta}^*_{n,t}$ captures sample-specific corruption. For clean samples, $\boldsymbol{\delta}^*_{n,t} = \mathbf{0}$. Since the corruption is induced by a permutation, we naturally impose the following restriction:

$$\boldsymbol{\delta}^*_{n,t} \in \Theta^0_\delta = \{\boldsymbol{\delta} \in \mathbb{R}^D : \boldsymbol{\delta} \perp \mathbf{1},\ \|\boldsymbol{\delta}\|_\infty \le B\}.$$

Group-Wise Lasso Estimator. Given corrupted observations $\tilde{\mathcal{S}}^n_t$, to robustly recover the clean preference parameter $\overline{\boldsymbol{c}}_n$, we jointly estimate $(\boldsymbol{c}, \boldsymbol{\delta}_{1:t})$ by solving the penalized likelihood problem

$$\tilde{\mathcal{L}}_{\tilde{\mathcal{S}}^n_t}(\boldsymbol{c}, \boldsymbol{\delta}_{1:t}) = \sum_{j=1}^{t} \ell_{\mathrm{PL}}\big(\tilde{S}^n_j; \boldsymbol{c} + \boldsymbol{\delta}_j\big) + \eta \sum_{j=1}^{t} \|\boldsymbol{\delta}_j\|_2,$$

where $\ell_{\mathrm{PL}}(\cdot)$ denotes the PL negative log-likelihood (see Eq. (3)) and $\eta > 0$ controls the regularization strength. The robust estimator is defined as

$$(\hat{\boldsymbol{c}}^{(\widetilde{\mathrm{QE}})}_{n,t}, \hat{\boldsymbol{\delta}}_{n,1:t}) = \underset{(\boldsymbol{c}, \boldsymbol{\delta}_{1:t}) \in \Theta^0_C \times [\Theta^0_\delta]^t}{\arg\min} \tilde{\mathcal{L}}_{\tilde{\mathcal{S}}^n_t}(\boldsymbol{c}, \boldsymbol{\delta}_{1:t}). \quad (9)$$

Why Group Sparsity? Each $\boldsymbol{\delta}_j$ represents a structured corruption affecting all D objectives within ranking j. The ℓ_2 group penalty enforces sparsity across samples, encouraging $\boldsymbol{\delta}_j = \mathbf{0}$ for clean feedback while permitting nonzero corrections for corrupted rankings. This matches the ϵ-contamination model, where only a fraction of samples are adversarial.

While Lasso-regularized MLE has been studied in RLHF under the BTL model [Bukharin et al., 2024], our setting is fundamentally different. Each feedback instance is a structured top-m ranking under the more general PL model, and corruption perturbs multiple objectives jointly rather than a single scalar reward. The proposed group-wise regularization explicitly captures this structure and extends robustness guarantees to multi-objective preference learning.

Lemma 4. *For any user n, if $\eta \ge \frac{D^3}{3}$, then with probability at least $1 - \frac{(1+D+2e^2)\rho^2\pi^2}{6}$, for all $t \ge C_0 \log(T/\rho)$, there exists a $C_3 > 0$, the group-wise Lasso estimator satisfies*

$$\|\hat{\boldsymbol{c}}^{(\widetilde{QE})}_{n,t} - \overline{\boldsymbol{c}}^0_n\|_2^2 + \|\mathrm{Vec}(\hat{\boldsymbol{\delta}}_{n,1:t} - \boldsymbol{\delta}^*_{1:t})\|_2^2$$
$$\le C_3 \left(\frac{D^2\eta^2}{m^{\downarrow}_{n,t}} \Big(\epsilon + \sqrt{\frac{\epsilon \log(t/\rho)}{t}}\Big) + \frac{D^3 \log(t/\rho)}{m^{\downarrow}_{n,t} t} \right),$$

where $\mathrm{Vec}(\cdot) : \mathbb{R}^{D\times t} \mapsto \mathbb{R}^{Dt}$ denotes the vectorized flatten operation in column-major order, $m^{\downarrow}_{n,t} = \min_{j\in[t]} m_{n,j}$.

The proof is provided in Appendix J. The estimation error scales linearly with the corruption rate ϵ (up to logarithmic factors). When $\epsilon t = O(\log t)$, the leading term reduces to $O\big(\frac{\log t}{m^{\downarrow}_{n,t} t}\big)$, matching the clean-sample convergence rate. Thus the estimator remains statistically consistent below the majority corruption threshold.

Implementation. For the final bandit algorithm under corrupted queries, we adopt the same MO-PQUCB framework but replace line-17 in Algorithm 1 with Eq. (9) for preference anchor estimation, and introduce η as an additional hyperparameter. All other steps remain unchanged.

6.4 THEORETICAL GUARANTEES

We state the main compacted results under corrupted queries with dependence on key parameters; the full versions and proofs are deferred to Appendix K and L respectively.

Lemma 5. *Under the ϵ-corruption model, for any user $n \in [N]$, for all $t \ge C_0 \log(T)$, there exists a constant $C_4 > 0$ such that the estimation error of QE-anchored preference estimator with group-wise Lasso is bounded with high probability by:*

$$\|\hat{\boldsymbol{c}}^{(\mathrm{HB})}_{n,t} - \overline{\boldsymbol{c}}_n\|_{\boldsymbol{V}_{n,t}}$$
$$\le C_4 \Bigg(\underbrace{\alpha\sqrt{D^3 \log(\alpha D t)}}_{\textit{bandit noise term}} + \underbrace{D\sqrt{\frac{1-\alpha}{m^{\downarrow}_{n,t}}\big(\epsilon + \frac{D\log(t)}{t}\big)}}_{\textit{Corrupted QE noise term}} \Bigg).$$

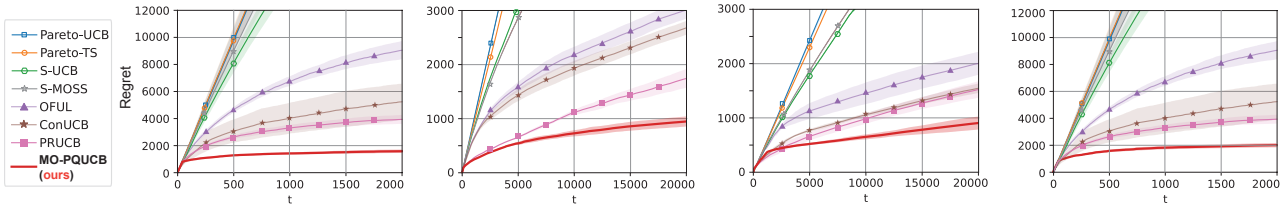
Theorem 4 (Regret under Corrupted Queries)**.** *Consider Algorithm 1 with the group-wise Lasso estimator. By setting $\alpha = \lambda = \Theta\big(\epsilon + \log^{-1}(T)\big)$, with high probability, the user-specific regret satisfies*

$$R^n(T) = O\left(\sqrt{\frac{D^3 K}{m^{\downarrow}_{n,T}} T \log T} + \sqrt{D^3 K T \epsilon}\, \log T \right).$$

Aggregating all users and absorbing the problem-dependent constants of D, K yield the compact final regret

$$R(T) = O\Big(N\big(\sqrt{T\log T} + \sqrt{\epsilon T}\, \log T\big)\Big).$$

Discussion. The regret decomposes into a clean-learning term $O(\sqrt{T \log T})$ and a corruption-induced penalty $O(\sqrt{\epsilon T} \log T)$. When $\epsilon = O(1/\log T)$, the corruption term is negligible and the regret matches the uncorrupted setting. Notably, as ϵ increases, α increases and the algorithm naturally shifts reliance from corrupted query signals to reliable bandit feedback.



(a) Synthetic data (b) TripAdvisor (c) BeerAdvocate (d) Synthetic (ϵ-Corruption)

Figure 2: Regret comparison on different datasets and query environments. (t denotes the number of interaction rounds)

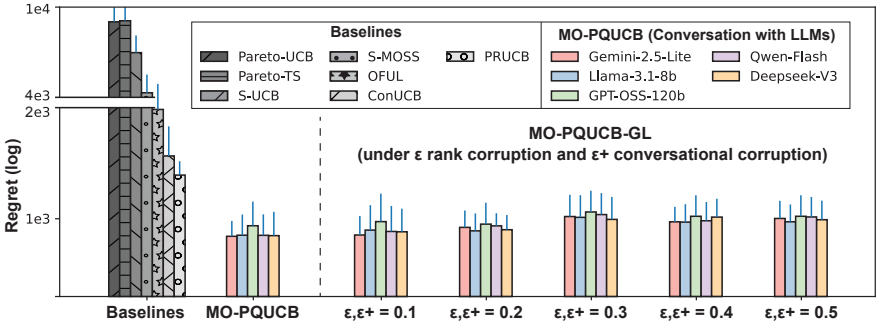


Figure 3: Cumulative regrets of hotel recommendation with LLMs as conversational agents on TripAdvisor.

7 EXPERIMENTS

We evaluate MO-PQUCB on real-world and synthetic data to validate: (i) regret reduction via proactive preference elicitation, (ii) robustness under corrupted queries, and (iii) practical integration with LLM-based conversational agents[2].

7.1 EXPERIMENTAL SETUP

Datasets. We consider: (1) a synthetic MO-MAB environment with $K = 40$, $D = 20$, $N = 20$ and $T = 2000$. (2) **TripAdvisor**: A hotel review dataset (878K reviews, 4.3K hotels). Each review contains six ($D=6$) aspect-ratings and an overall rating. After preprocessing and filtering, we set $N=10$, $K=62$ and $T = 20000$. (3) **BeerAdvocate**: A beer review dataset (1.5M reviews, 66K+ beers). Each review contains four ($D = 4$) aspects and an overall rating. After preprocessing, we set $N = 14$, $K = 167$ and $T = 20000$. Detailed descriptions and setups are in Appendix B.1.

Baselines. We compare our algorithm with the following multi-objective bandit algorithms: (1) Scalarization-based methods: **S-UCB** [Drugan and Nowe, 2013], **S-MOSS**; (2) Pareto-optimality methods: **Pareto-UCB** [Drugan and Nowe, 2013], **Pareto-TS** [Yahyaa and Manderick, 2015]; (3) Multi-objective variants of contextual and conversational bandit: **MO-OFUL** [Abbasi-Yadkori et al., 2011], **ConUCB** [Zhang et al., 2020]; (4) Preference-aware method: **PRUCB** [Cao et al., 2025]. Each experiment is repeated over 10 independent runs. By default, the query elicitation provides full-ranking feedback ($m = D$) unless stated otherwise. We provide implementation details in Appendix B.2.

[2]The code repository is provided at `https://github.com/caolinfeng/MO-PQUCB`

7.2 OVERALL PERFORMANCE

The average cumulative regrets with standard error for each algorithm on synthetic and real-world datasets are shown in Fig. 2(a), 2(b) and 2(c). MO-PQUCB consistently achieves the lowest cumulative regret and variance, outperforming all baselines. Notably, MO-PQUCB surpasses PRUCB [Cao et al., 2025], demonstrating that proactive query-based interaction effectively reduces cumulative regret through more informative preference elicitation. Although standard contextual and conversational bandit methods [Abbasi-Yadkori et al., 2011, Zhang et al., 2020] exhibit sublinear regret, they perform significantly worse than preference-aware MO-MAB approaches, highlighting the advantage of dual exploration in explicitly preference-driven environments. Furthermore, traditional scalarization and Pareto-optimality methods [Drugan and Nowe, 2013, Yahyaa and Manderick, 2015] fail to converge to personalized optimal arms, which is expected since they do not learn user preferences.

7.3 ROBUSTNESS UNDER CORRUPTED QUERIES

For the corrupted QE experiments, we introduce an additional ϵ-corruption model defined in Section 6.1 with $\epsilon = 0.2$. The preference anchor is estimated from corrupted

QE data using Eq. 9, which incorporates a group-wise Lasso penalty to enable robust preference learning. To distinguish this variant, we denote it as MO-PQUCB-GL.

In Fig. 2(d), we report the regrets of different methods. Even under corruption, our MO-PQUCB-GL consistently outperforms all baselines, demonstrating its robustness and stable convergence despite noise in preference feedback. This improvement is attributed to the algorithm's group-wise Lasso regularization and hybrid estimation scheme, which effectively suppress corrupted preference signals while still leveraging useful structural information from QE.

7.4 LLMS AS CONVERSATIONAL AGENTS

To evaluate practical deployment, we integrate MO-PQUCB into a conversational recommendation pipeline where large language models (LLMs) serve as preference interpreters between users and the bandit learner.

Protocol. We set two LLM agents as simulators. The Simulator-1 generates natural-language queries reflecting a user's latent preference vector $\overline{c}^n$. The Simulator-2 then maps each query to a ranked subset of objectives, which is provided to MO-PQUCB as proactive feedback. To model realistic conversational uncertainty, we introduce two noise sources in Simulator-1: (i) ranking corruption with probability ϵ, and (ii) ambiguous query generation with probability ϵ^+. Detailed protocol design is provided in Appendix B.3.1.

Setup. Experiments are conducted on TripAdvisor dataset with $T = 2000$. We evaluate representative LLMs including Gemini-2.5 [Comanici et al., 2025], Llama-3.1-8B [Dubey et al., 2024], GPT-OSS-120B [Agarwal et al., 2025], Qwen-Flash [Yang et al., 2025] and Deepseek-V3 [Liu et al., 2024]. The implementation details are deferred to Appendix B.3.2.

Regret. As shown in Fig. 3, all LLM-based MO-PQUCB variants achieve significantly lower regret than traditional scalarization, Pareto-optimality, and contextual bandit baselines, demonstrating the advantage of proactive query-based preference elicitation. When introducing the ϵ and ϵ^+ corruption model, the MO-PQUCB-GL variant shows stable performance even under increasing conversational and ranking noise. The gradual and limited increase in regret indicates that the group-wise Lasso regularization effectively compensates for corrupted preference signals.

8 CONCLUSION

In summary, we introduce a new proactive query-based preference elicitation framework for multi-objective bandit problems, enabling users to express high-level preferences directly through natural language. By modeling query-derived partial rankings with a PL subset choice model and integrating this into a UCB-style bandit algorithm, MO-PQUCB effectively balances structured preference signals and reward feedback. Our algorithm not only achieves provable regret guarantees but also demonstrates strong empirical performance. This work highlights the potential of proactive interaction in bridging language-driven user input with principled sequential decision-making.

Acknowledgements

This work has been supported in part by the U.S. National Science Foundation under the grants: NSF AI Institute (AI-EDGE) 2112471, CNS2312836, CNS-2225561, and CNS-2239677, by the Office of Naval Research under grant N00014-24-1-2729, and by the Army Research Laboratory under Cooperative Agreement Number W911NF-23-2-0225, by. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, pages 217–226, 2009.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Amir Rezaei Balef and Setareh Maghsudi. Piecewise-stationary multi-objective multi-armed bandit with application to joint communications and sensing. *IEEE Wireless Communications Letters*, 12(5):809–813, 2023.

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.

Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. Robust reinforcement learning from corrupted human feedback.

In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Linfeng Cao, Ming Shi, and Ness B Shroff. Provably efficient multi-objective bandit algorithms under preference-centric customization. *arXiv preprint arXiv:2502.13457*, 2025.

Ji Cheng, Bo Xue, Jiaxiang Yi, and Qingfu Zhang. Hierarchize pareto dominance in multi-objective stochastic linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11489–11497, 2024.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 815–824, 2016.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Huy Dao, Yang Deng, Dung D Le, and Lizi Liao. Broadening the view: Demonstration-augmented prompt learning for conversational recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 785–795, 2024.

Arnhav Datar, Arun Rajkumar, and John Augustine. Byzantine spectral ranking. *Advances in Neural Information Processing Systems*, 35:27745–27756, 2022.

Mădălina M Drugan. Covariance matrix adaptation for multiobjective multiarmed bandits. *IEEE Transactions on Neural Networks and Learning Systems*, 30(8):2493–2502, 2018.

Madalina M Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2013.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Matthias Ehrgott. *Multicriteria optimization*, volume 491. Springer Science & Business Media, 2005.

Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. *Advances in Neural Information Processing Systems*, 27, 2014.

Thomas P Hayes. A large-deviation inequality for vector-valued martingales. *Combinatorics, Probability and Computing*, 37, 2005.

Alihan Hüyük and Cem Tekin. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning*, 110(6):1233–1266, 2021.

Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31, 2018.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 3080–3086, 2019.

John I. Marden. *Analyzing and Modeling Rank Data*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC, Boca Raton, FL, 1996.

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.

Renting Rui, Yunjia Xi, Weiwen Liu, Jianghao Lin, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. Action first: Leveraging preference-aware actions for more effective decision-making in interactive recommender systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 53–63, 2025.

Aadirupa Saha and Aditya Gopalan. Pac battling bandits in the plackett-luce model. In *Algorithmic Learning Theory*, pages 700–737. PMLR, 2019.

Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramch, Martin J Wainwright, et al. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *Journal of Machine Learning Research*, 17(58):1–47, 2016.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12: 389–434, 2012.

Eralp Turgay, Doruk Oner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, pages 1673–1681. PMLR, 2018.

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

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhihui Xie, Tong Yu, Canzhe Zhao, and Shuai Li. Comparison-based conversational recommender system with relative bandit feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1400–1409, 2021.

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 *Forty-first International Conference on Machine Learning*, 2024.

Saba Q Yahyaa and Bernard Manderick. Thompson sampling for multi-objective multi-armed bandits problem. In *ESANN*, 2015.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Gangyi Zhang. User-centric conversational recommendation: Adapting the need of user with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1349–1354, 2023.

Xiaoying Zhang, Hong Xie, Hang Li, and John CS Lui. Conversational contextual bandit: Algorithm and application. In *Proceedings of the web conference 2020*, pages 662–672, 2020.

Canzhe Zhao, Tong Yu, Zhihui Xie, and Shuai Li. Knowledge-aware conversational preference elicitation with bandit feedback. In *Proceedings of the ACM Web Conference 2022*, pages 483–492, 2022.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 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.

A DETAILED RELATED WORK

A.1 MULTI-OBJECTIVE MULTI-ARMED BANDITS

The multi-objective multi-armed bandit (MO-MAB) framework extends the classical MAB setting by considering reward vectors over multiple objectives. Early works in this area focused on identifying the Pareto-optimal set of arms that cannot be strictly improved across all objectives [Drugan and Nowe, 2013, Turgay et al., 2018, Lu et al., 2019, Drugan, 2018, Balef and Maghsudi, 2023]. These methods aim to approximate the Pareto front and often evaluate performance in terms of Pareto regret. While effective in general-purpose scenarios, such approaches overlook the individual trade-offs that users may have across objectives, thereby failing to deliver personalized solutions.

A.2 PREFERENCE-AWARE BANDITS

Several studies have extended MO-MABs with user preferences using lexicographic orders [Ehrgott, 2005, Hüyük and Tekin, 2021, Cheng et al., 2024], which assume strict objective prioritization. However, such models fail to capture real-world trade-offs. Moreover, these methods assume the user preference to be known in advance, and the preference learning is not required. Preference-aware approaches [Cao et al., 2025] instead model utility as the inner product between a latent preference vector and objective rewards, enabling more personalized decisions and generalizing the lexicographic ordering as a special case. Our work extends this framework by incorporating proactively elicited partial preferences.

A.3 LANGUAGE-BASED PREFERENCE ELICITATION FOR BANDITS

Recent advances in instruction-tuned language models (e.g., GPT, BERT) have enabled users to express preferences via high-level natural language queries, allowing structured preference inference through models trained for instruction following [Ouyang et al., 2022, Thoppilan et al., 2022, Zhou et al., 2023]. Such capabilities open new opportunities for integrating natural language elicitation into sequential decision-making. In contrast to prior conversational bandits [Christakopoulou et al., 2016, Zhang et al., 2020, Xie et al., 2021, Zhao et al., 2022], which rely on passive, system-initiated interactions with binary feedback, our approach introduces a proactive querying protocol. Related directions in language-based RL and reward learning [Christiano et al., 2017, Wang et al., 2024] also highlight the value of natural language supervision. Our design enhances expressiveness and exploration efficiency, particularly in multi-objective settings where preference structures are more nuanced.

Table 1 summarizes the MO-MAB methods most closely related to our work. As shown, our approach uniquely integrates multi-objective optimization, preference awareness, and language-driven interaction within a unified formulation, achieving both theoretical optimality and practical adaptability in realistic conversational settings.

Table 1: Comparison of representative algorithms in regret, computational complexity, and functionality. Columns denote: **Pref.** (preference awareness), **Conv.** (conversational interaction), **Proact.** (proactive query), and **Robust.** (robustness to corrupted preference signals).

Algorithm	**Regret**	**Pref.**	**Conv.**	**Proact.**	**Robust.**
Pareto-UCB Drugan and Nowe [2013]	$O(T)$	✘	✘	✘	✘
Pareto-TS Yahyaa and Manderick [2015]	$O(T)$	✘	✘	✘	✘
S-UCB Drugan and Nowe [2013], S-MOSS	$O(T)$	✘	✘	✘	✘
OFUL Abbasi-Yadkori et al. [2011]	$O(\sqrt{T}\log T)$	✔	✘	✘	✘
ConUCB Zhang et al. [2020]	$O(\sqrt{T}\log T)$	✔	✔	✘	✘
PRUCB Cao et al. [2025]	$O(\sqrt{T}\log T)$	✔	✘	✘	✘
MO-PQUCB	$O(\sqrt{T}\log T)$	✔	✔	✔	✘
MO-PQUCB-GL	$O(\sqrt{T\log T}+\sqrt{\epsilon T}\log T)$	✔	✔	✔	✔

B EXPERIMENTS

In this section, we describe experimental results on both synthetic data and real-world data under different environments to validate our proposed algorithm.

B.1 DETAILED DATASET SETUPS

Synthetic Data. We construct synthetic environments following [Cao et al., 2025]. Specifically, we consider a MO-MAB setting with $K = 40$ arms, each associated with a $D = 20$ dimensional reward vector. For each arm $i \in [K]$, the reward on objective d is drawn from a Gaussian distribution with randomized mean $\mu_i(d) \in [0, 5]$ and variance 0.5. We simulate $N = 20$ users, where each user's mean preference vector is randomized within $[0, 5]$, and instantaneous preferences are sampled from a Gaussian distribution with variance 0.5. We set the time horizon to $T = 2000$. At each round $t \in [T]$, a user $n_t \in [N]$ is randomly selected and presented with a subset of arms $\mathcal{A}_t^n \subset [K]$ of size $|\mathcal{A}_t^n| = \lfloor K/4 \rfloor$.

TripAdvisor (Four-City Dataset[3]). The dataset contains 878,561 reviews (1.3 GB) from 4,333 hotels on TripAdvisor, with each review including an *overall rating* and six 5-scale aspect ratings: *cleanliness*, *location*, *rooms*, *service*, *sleep quality*, and *value*. We treat these six aspects as the multi-objective reward dimensions (i.e., $D = 6$) and the overall rating as the utility signal. After preprocessing by removing empty users, reviews with missing ratings, and users or hotels with fewer than 10 reviews, the dataset contains 786 reviews, 257 users, and 62 hotels. Each hotel is modeled as an arm with mean rewards given by its average aspect ratings. We select the 10 most active users ($N = 10$) as the user set, with the user preferences estimated via ridge regression between aspect ratings and overall ratings. For simulation, we generate instantaneous rewards and preferences with Gaussian noise ($\sigma^2 = 0.5$), and set the horizon to $T = 20{,}000$, where one user is randomly chosen at each step for single-city hotel recommendation.

BeerAdvocate[4]. The dataset contains over 1.5 million reviews of 66,000+ beers from various breweries. Each review provides 5-scale ratings on four aspects (*appearance*, *aroma*, *palate*, *taste*) and an *overall impression*, which we treat as the utility signal, with the four aspects as the multi-objective reward dimensions ($D = 4$). After filtering users with at least 2,000 reviews and beers with at least 50 reviews, we retain 14 users and 167 beers. Each beer is modeled as an arm with mean rewards given by average aspect ratings, and user preferences are estimated via ridge regression between aspect ratings and overall impressions. In simulation, instantaneous rewards and preferences are drawn from Gaussian distributions with variance 0.5, and the horizon is set to $T = 20{,}000$, with one user randomly selected per step for recommendation.

B.2 BASELINE DETAILS

We select the following multi-objective algorithms as baselines to compare with ours:

- **S-UCB** [Drugan and Nowe, 2013]: A scalarized UCB algorithm, which scalarizes the multi-dimensional reward by assigning weights to each objective and then employs the single objective UCB algorithm Auer et al. [2002]. Throughout the experiments, we assign each objective with equal weight.
- **S-MOSS**: A scalarized UCB algorithm, which follows the similar way with S-UCB by scalarizing the multi-dimensional reward into a single one, but uses MOSS [Audibert and Bubeck, 2009] policy for arm selection.
- **Pareto-UCB** [Drugan and Nowe, 2013]: A Pareto-based algorithm, which compares different arms by the upper confidence bounds of their expected multi-dimensional reward by Pareto order and pulls an arm uniformly from the approximate Pareto front.
- **Pareto-TS** [Yahyaa and Manderick, 2015]: A Pareto-based algorithm, which makes use of the Thompson sampling technique to estimate the expected reward for every arm and selects an arm uniformly at random from the estimated Pareto front.
- **MO-OFUL** [Abbasi-Yadkori et al., 2011]: A multi-objective variant of the classic OFUL algorithm for linear bandits. For MO-MAB environment adaptation, it estimate user preferences via ridge regression using both reward and overall utility feedback, and we replaces the input feature with empirical reward estimates, and applies an exploration rate of $\epsilon = 0.05$ to handle the uncertainty of reward estimates.
- **ConUCB** [Zhang et al., 2020]: A multi-objective variant of the traditional conversational contextual bandit algorithm based on the passive query interaction. Similar to MO-OFUL, it replaces the input feature with empirical reward estimates and uses an exploration rate of $\epsilon = 0.05$ for adaptation in preference-aware MO-MAB environment. The algorithm is developed based on a passive keyword-level cue to accelerate convergence. For the keyword implementation, we set the same number of keywords N_{kw} as the number of objectives (i.e., $N_{kw} = D$). For the i-th keyword, we relate it with 3 arms that has the largest reward value on i-th objective. When the conversation session is triggered, the system provide a

[3] https://notebook.community/melqkiades/yelp/notebooks/TripAdvisor-Datasets

[4] https://snap.stanford.edu/data/web-BeerAdvocate.html

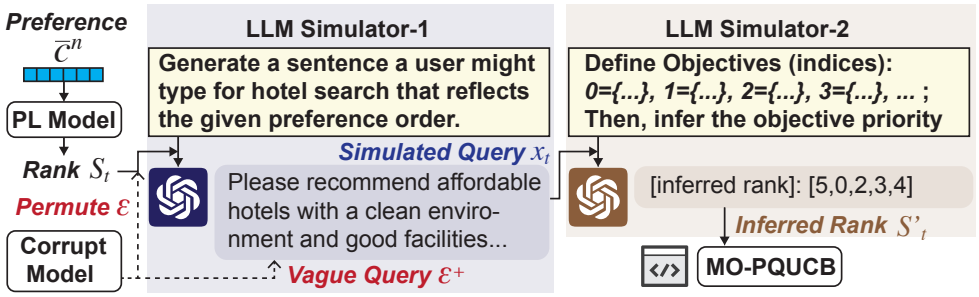


Figure 4: Experimental protocol of hotel recommendation with LLMs as conversational agents for simulation.

keyword and receive a binary feedback (e.g., like / dislike) from user to infer the user's preference and identify optimal arm. The conversational frequency is set as $5\log(T)$, which follows the optimal setting in original paper.

- **PRUCB** [Cao et al., 2025]: A recently proposed preference-aware MO-MAB algorithm based on the UCB strategy to handle the hidden user preference case.

Devices. All experiments were run on a Linux system with 2 CPUs (Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz).

B.3 EXPERIMENTAL SETUPS OF LLMS AS CONVERSATIONAL AGENTS

B.3.1 Experimental Protocol.

The experimental protocol is shown in Fig. 4. Simulator-1 generates natural language queries that reflect a user's latent preference vector $\bar{c}^n$ (e.g., "Please recommend affordable hotels with a clean environment and good facilities"). To evaluate robustness, we introduce a corruption model that simulates realistic conversational noise. The corruption acts in two ways: (1) with probability ϵ, the true ranking S_t is permuted before being passed to LLM Simulator-1; and (2) with probability ϵ^+, the LLM Simulator-1 generates vague or ambiguous natural-language queries. This setup reflects real-world interactions where users may change their minds, express conflicting priorities, or describe preferences vaguely using imprecise language, enabling systematic evaluation of MO-PQUCB's robustness to imperfect conversational inputs. Simulator-2 then receives the generated query and infers the corresponding objective ranking S'_t, which is passed to MO-PQUCB as proactive preference feedback.

B.3.2 Implementation

We employ several state-of-the-art LLMs as conversational agents, including Gemini-2.5-Lite Comanici et al. [2025], Llama-3.1-8B Dubey et al. [2024], GPT-OSS-120B Agarwal et al. [2025], Qwen-Flash Yang et al. [2025] and Deepseek-V3 Liu et al. [2024]. In each experiment, both LLM Simulator-1 and LLM Simulator-2 invoke the same model API to ensure consistent behavior between query generation and preference interpretation, with the temperature set to 0.5 to control generation randomness. We frame our experiments as a personalized hotel recommendation task using the *TripAdvisor Four-City* dataset. Specifically, we perform five independent runs, each run randomly selecting one active user and setting the interaction horizon $T = 2{,}000$ to regulate the query-calling rate.

B.3.3 Prompts and Examples in Section 7.4

This section introduces wording of prompts used and examples in the hotel recommendation simulation with LLMs as conversational agents in Section 7.4. Specifically, the prompt instructions for LLM simulator-1 to generate natural and human-like hotel recommendation queries are listed in Table 2, and the prompt instructions for LLM simulator-2 to infer user's preference rank are listed in Table 3.

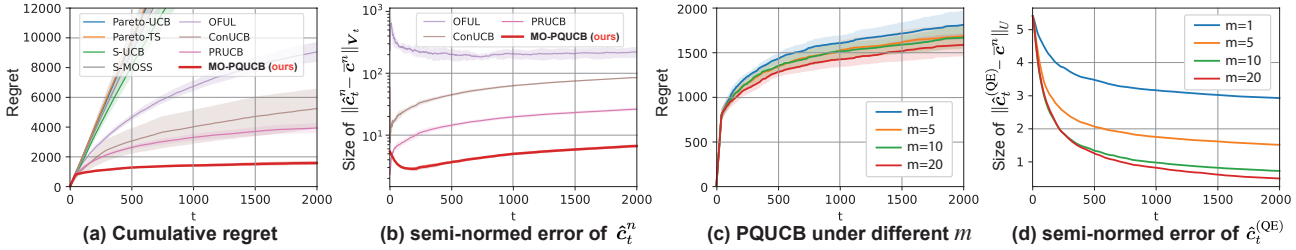


Figure 5: Experimental results on synthetic dataset.

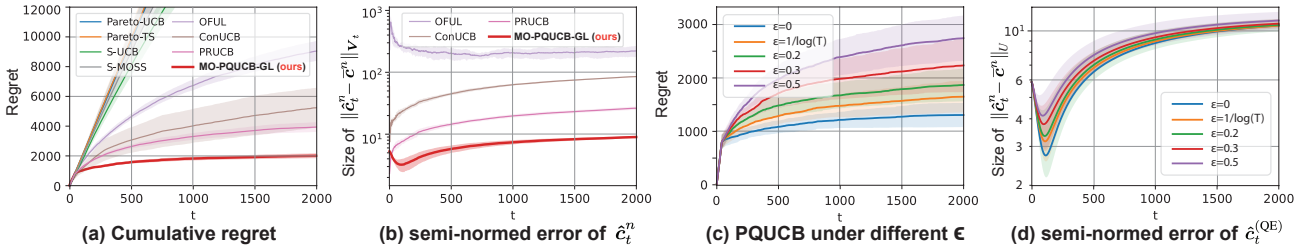


Figure 6: Experimental results on synthetic dataset under query corruption.

B.4 ADDITIONAL EXPERIMENTAL RESULTS

B.4.1 Results With Clean Queries

Semi-Normed Preference Estimation Error. We further analyze the averaged semi-normed preference estimation error, $\|\hat{c}_t^{(\mathrm{HB})} - \bar{c}\|V_t$ on synthetic data (Fig. 5 b). As shown, the proposed MO-PQUCB consistently achieves lower and more stable estimation error compared to other methods. This result confirms that incorporating proactive preference elicitation not only provides better initial guidance for learning but also enhances convergence with bandit feedback.

Effect of Query Elicitation Size. Fig. 5(c) and (d) show the cumulative regret and averaged semi-normed preference estimation error of MO-PQUCB under different query sizes m, denoting the number of top-ranked objectives observed per query. As m increases, both regret and estimation error decrease, reflecting faster convergence and more accurate preference learning. Even partial rankings ($m = 5$ or $m = 10$) markedly outperform minimal feedback ($m = 1$), while full rankings ($m = 20$) yield the best performance. These results highlight the statistical value of proactive preference elicitation—MO-PQUCB benefits consistently from more informative queries.

B.4.2 Results under ϵ-Corrupted Queries

Analysis on Corruption Rate. Fig. 6(c,d) presents the performance of MO-PQUCB under different corruption levels ϵ in the PL-based preference elicitation. As the corruption level increases, both regret and estimation error rise, reflecting the increased difficulty of accurately learning user preferences from noisy rankings. However, even at higher corruption levels ($\epsilon = O(1)$), the algorithm maintains bounded regret growth and controlled estimation error, showcasing its robustness.

Table 2: Wording of prompts used in the hotel recommendation simulation with LLMs as conversational agents. The pink-highlighted line indicates the additional instruction that is appended only when an ϵ^+-vague query corruption event is triggered.

LLM Simulator-1: Prompt Construction Overview

(a) Static Header. This prompt guides the language model to generate natural hotel-search queries that align with user-specific objective preferences. It explicitly constrains the output format to ensure structured data synthesis and prohibits extraneous text generation.

Prompt Header

You are helping with hotel recommendation data synthesis.
Objectives (indices): `0=[...], 1=[...], 2=[...], 3=[...], 4=[...], 5=[...].`
For each preference order list: Generate one natural, descriptive sentence that a user might type when searching for a hotel, reflecting the given preference order. The sentence should be vague and ambiguous.
STRICT OUTPUT: Return only JSON—a list of objects formatted as `{"id": <int>, "query": <string>}`. No extra text.

INPUT: Given preference order list [...].

(b) Dynamic Input Block. Specifies the preference ranking order data generated by the PL model under the user's true preference vector to be inserted into the header.

Example Input

```
 {"objectives": ["cleanliness","location","rooms","service","sleep quality","value"]}
{"id": 12, "preferences": [0, 4, 3, 2, 5, 1]}
```

(c) Final Assembled Prompt. The complete prompt sent to the LLM, obtained by concatenating the header and the input block.

Final Prompt Example

You are helping with hotel recommendation data synthesis.
Objectives (indices): `0=cleanliness, 1=location, 2=rooms, 3=service, 4=sleep quality, 5=value.`
For each preference order lis: Generate one natural, descriptive sentence that a user might type when searching for a hotel, reflecting the given preference order. The sentence should be vague and ambiguous.
STRICT OUTPUT: Return only JSON—a list of objects formatted as `{"id": <int>, "query": <string>}`. No extra text.
INPUT: Given preference order list: `"id": 12, "preferences": [0, 4, 3, 2, 5, 1]`

(d) Response Output Example. Example responses of both clean and ϵ^+-corrupted cases returned by the LLM following the instruction.

Response Example (Clean)

```
 { "id": 12, "query": "Please recommend a
hotel with excellent cleanliness, quiet
environment, and great service." }
```

Response Example (ϵ^+-Corrupted)

```
 { "id": 12, "query": "Please
recommend a hotel for a business
trip." }
```

Table 3: Wording of prompts used in hotel recommendation simulation with LLMs as conversational agents.

LLM Simulator-2: Prompt Construction Overview

(a) Static Header. This prompt guides the language model to infer the user's latent objective preference order ranking based on the given query. It explicitly constrains the output format to ensure structured data synthesis and prohibits extraneous text generation.

Prompt Header

Infer the user preference priority from the given hotel search query as a permutation of [0,1,2,3,4,5].
Objectives (indices): `0=[...], 1=[...], 2=[...], 3=[...], 4=[...], 5=[...].`
STRICT OUTPUT: Return only JSON—a list of objects formatted as `{"id": <int>, "rank": [i0,i1,i2,i3,i4,i5]}`. No extra text.

INPUT: Hotel search query: [...].

(b) Dynamic Input Block. Input the hotel query just generated by LLM Simulator-1.

Example Input

```
 {"objectives": ["cleanliness","location","rooms","service","sleep quality","value"]}
{ "id": 12, "query": "Please recommend a hotel with excellent cleanliness, quiet
environment, and great service." }
```

(c) Final Assembled Prompt. The complete prompt sent to the LLM, obtained by concatenating the header and the input block.

Final Prompt Example

Infer the user preference priority from the given hotel search query as a permutation of [0,1,2,3,4,5]
Objectives (indices): `0=cleanliness, 1=location, 2=rooms, 3=service, 4=sleep quality, 5=value.`
STRICT OUTPUT: Return only JSON—a list of objects formatted as "id": <int>, "rank": [i0,i1,i2,i3,i4,i5]. No extra text.
INPUT: Hotel search query: `"id": 12, "query": Please recommend a hotel with excellent cleanliness, quiet environment, and great service.`

(d) Response Output Example. An example response returned by the LLM Simulator-2 following the instruction.

Response Example

```
{ "id": 12, "rank": [0,4,3,2,1,5] }
```

C LIMITATIONS AND FUTURE WORK

While our proposed MO-PQUCB framework effectively integrates proactive preference elicitation and hybrid feedback to model user-specific preferences, it assumes a static preference vector for each user throughout the decision horizon. In many real-world applications, however, user preferences may evolve over time due to contextual changes, shifting objectives, or accumulated experiences. Our current model does not account for such dynamics, potentially limiting its adaptability in non-stationary environments. Extending the framework to accommodate temporally varying preferences remains an important direction for future work.

Besides, in this paper, we focus on the theoretical foundation of proactive preference elicitation and regret-optimal learning. While our framework assumes that the top-m preferred objectives can be extracted from user queries, the actual parsing of natural language using a language model is treated as an external module and is not the focus of our algorithmic design or analysis. In practice, parsing can be implemented via lightweight LLMs or intent classifiers, with constant overhead per round that does not scale with the number of arms K or horizon T, and can often be preprocessed or cached. Developing and integrating such a practical natural language interface with MO-PQUCB is an important and promising direction for future work.

D PROOF OF LEMMA 1

The main idea of the proof is inspired from the proof of [Hajek et al., 2014] Theorem 3. Our main results depend on the structure of a weighted undirected graph $\mathcal{G}_t$ defined as follows. For notational simplicity, we omit the user index n in this section, and present the formulation for a fixed user. All definitions naturally extend to the multi-user setting.

Definition 1 (Comparison Graph $\mathcal{G}_t$)**.** *For any $t \in [T]$, let each objective $j \in [D]$ corresponds to a vertex. For any pair of vertices $j, j' \in [D]$, there is a weighted edge between them with the weight equals to $\sum_{i\in[t]} \frac{m_i}{D^2}$. Let $L_{\mathcal{G}_t}$ denotes the Laplacian matrix of this weighted complete graph.*

Observe that for any t, $L_{\mathcal{G}_t}$ is positive semi-definite and the smallest eigenvalue of $L_{\mathcal{G}_t}$ is zero with the corresponding eigenvector given by the normalized all-one vector. Let $0 = \xi_1(\mathcal{G}_t) \leq \xi_2(\mathcal{G}_t) \leq ... \leq \xi_D(\mathcal{G}_t)$ denote the eigenvalues of $L_{\mathcal{G}_t}$ in ascending order.

Proof of Lemma 1. By the loss function for MLE optimization, for any number of QE sample t, we have

$$\begin{aligned}\mathcal{L}_{\mathcal{S}_t}(\boldsymbol{c}) &= -\sum_{i=1}^{t}\sum_{j=1}^{m_i} \log\big(\frac{\exp(\boldsymbol{c}_{S_i(j)})}{\sum_{j'=j}^{m_i}\exp(\boldsymbol{c}_{S_i(j')}) + \sum_{j'\in[D]\setminus S_i}\exp(\boldsymbol{c}_{j'})}\big)\\ &= -\sum_{i=1}^{t}\sum_{j=1}^{m_i} \log\big(\frac{\exp(\boldsymbol{c}_{\sigma_i(j)})}{\sum_{j'=j}^{D}\exp(\boldsymbol{c}_{\sigma_i(j')})}\big)\end{aligned}$$

Let $\nabla\mathcal{L}_t(c)$ as the gradient of $\mathcal{L}_{\mathcal{S}_t}(\boldsymbol{c})$ w.r.t $\boldsymbol{c}$, $H_t(c)$ as the Hessian matrix of $\mathcal{L}_{\mathcal{S}_t}(\boldsymbol{c})$ w.r.t $\boldsymbol{c}$. Define $\Delta_t = \hat{c}_t^{(\mathrm{QE})} - \overline{c}^0$. It follows from the definition that Δ_t is orthogonal to the vector $\mathbf{1}$. By the definition of MLE estimator, we have $\mathcal{L}_{\mathcal{S}_t}(\hat{c}_t^{(\mathrm{QE})}) \leq \mathcal{L}_{\mathcal{S}_t}(\overline{c}^0)$, and thus

$$\mathcal{L}_{\mathcal{S}_t}(\hat{c}_t^{(\mathrm{QE})}) - \mathcal{L}_{\mathcal{S}_t}(\overline{c}^0) - \nabla\mathcal{L}_t(\overline{c}^0)^\top\Delta_t \leq -\nabla\mathcal{L}_t(\overline{c}^0)^\top\Delta_t \leq \|\nabla\mathcal{L}_t(\overline{c}^0)\|_2\|\Delta_t\|_2,$$

where the last inequality holds due to the Cauchy-Schwartz inequality. By the Taylor expansion, there exists a $c = a\overline{c}^0 + (1-a)\hat{c}^{(\mathrm{QE})}$ for some $a \in [0,1]$ such that

$$\mathcal{L}_{\mathcal{S}_t}(\hat{c}_t^{(\mathrm{QE})}) - \mathcal{L}_{\mathcal{S}_t}(\overline{c}^0) - \nabla\mathcal{L}_t(\overline{c}^0)^\top\Delta_t = \frac{1}{2}\Delta^\top H_t(c)\Delta \geq \frac{1}{2}\xi_2(H_t(c))\|\Delta_t\|_2^2,$$

where the last inequality holds because the Hessian matrix $H_t(c)$ is positive semi-definite and $H_t(c)^\top\mathbf{1} = 0, \Delta_t^\top\mathbf{1} = 0$. Combining above results yields

$$\|\Delta_t\|_2 \leq \frac{2\|\nabla\mathcal{L}_t(\overline{c}^0)\|_2}{\xi_2(H_t(c))}. \tag{10}$$

We next present two key auxiliary results used in the proof.

Lemma 6. *For any sufficiently large number of samples $t \geq 8(2e^{2B}+1)\sqrt{2DT\log\frac{T}{\rho}}$, with probability at least $1 - \frac{D\rho^2}{t^2}$, we have*

$$\xi_2(H_t(c)) \geq \frac{t\overline{m}_t}{D(4e^{4B}+2e^{2B})}.$$

Lemma 7. *With probability at least $1 - \frac{e^2\pi^2\rho^2}{3}$, for any $t \in [T]$, we have*

$$\|\nabla\mathcal{L}_t(\overline{c}^0)\|_2 \leq 4\sqrt{\overline{m}_t t\log\frac{t}{\rho}}.$$

Combining result in Eq. 10 with above lemmas, we have with probability at least $1 - \frac{(D+2e^2)\pi^2\rho^2}{6}$, for all $t \geq 8(2e^{2B}+1)\sqrt{2DT\log\frac{T}{\rho}}$, we have

$$\|\Delta_t\|_2 \leq D(32e^{4B}+16e^{2B})\sqrt{\frac{\log(t/\rho)}{\overline{m}_t t}}.$$

□

D.1 PROOF OF LEMMA 6

Proof. Define $S^{(i,j)}$ as the set of objectives contending for the j-th position in the ranking of i-th ranking sample after higher ranking objectives have been selected. Here $j \in [1, m]_i$. Let $\mathbb{1}^{(i,j)}$ denote the indicator vector for the set $S^{(i,j)}$, and let $p^{(i,j)}$ denote the the corresponding probability column vector for the selection, then the Hessian can be written as:

$$H_t(c) = \sum_{i=1}^{t} \sum_{j=1}^{m_i} H^{i,j},$$

where $H^{i,j} = \frac{1}{2} \sum_{k,k' \in S^{(i,j)}} (e_k - e_{k'})(e_k - e_{k'})^\top p_k^{(i,j)} p_{k'}^{(i,j)} = \text{diag}(p^{(i,j)}) - p_k^{(i,j)} {p_k^{(i,j)}}^\top$, with $p_k^{(i,j)} = \frac{\mathbb{1}_k^{(i,j)} \exp(c_k)}{\sum_{k' \in S^{(i,j)}} \exp(c_{k'})}$, e_x is the x-th standard basis vector.

By Claim 1 in [Hajek et al., 2014], we have

$$e^{2B} H^{i,j} \succeq \frac{1}{2(D-j+1)^2} \sum_{k,k' \in S^{i,j}} (e_k - e_{k'})(e_k - e_{k'})^\top.$$

Summing over i and j, and note that $D - j + 1 \leq D$, we have

$$e^{2B} H_t(c) \succeq \frac{1}{2} \sum_{k,k' \in S^{i,j}} (e_k - e_{k'})(e_k - e_{k'})^\top \frac{1}{D^2} \sum_{j=1}^{m_i} \mathbb{1}_{\{\sigma_i^{-1}(k), \sigma_i^{-1}(k') \geq j\}} := \tilde{L}_t. \tag{11}$$

Additionally, Observe that

$$\begin{aligned} \sum_{j=1}^{m_i} \mathbb{P}_c[\sigma_i^{-1}(k), \sigma_i^{-1}(k') \geq j] &= 1 + \sum_{k'' \in [D]} \mathbb{1}_{\{k'' \neq k, k'\}} \frac{\exp(c_{i''})}{\exp(c_i) + \exp(c_{i'}) + \exp(c_{i''})} \\ &\geq 1 + \frac{m_i - 1}{2e^{2B} + 1} \geq \frac{m_i}{2e^{2B} + 1}. \end{aligned}$$

Recall that $L_{\mathcal{G}_t}$ is the Laplacian of $\mathcal{G}_t$ and $L_{\mathcal{G}_t} = \sum_{i=1}^{t} L_i$, we have

$$\begin{aligned} \mathbb{E}[\tilde{L}_t] &= \frac{1}{2} \sum_{i=1}^{t} \sum_{k,k' \in [D]} (e_k - e_{k'})(e_k - e_{k'})^\top \frac{1}{D^2} \sum_{j=1}^{m_i} \mathbb{P}_c[\sigma_i^{-1}(k), \sigma_i^{-1}(k') \geq j] \\ &\geq \frac{1}{2} \sum_{i=1}^{t} \sum_{k,k' \in [D]} (e_k - e_{k'})(e_k - e_{k'})^\top \frac{m_i}{D^2(2e^{2B} + 1)} = \frac{L_{\mathcal{G}_t}}{2e^{2B} + 1}, \end{aligned} \tag{12}$$

Define $a_{kk'} = \frac{1}{D^2} \sum_{j=1}^{m_i} \left(\mathbb{1}_{\left\{\sigma_i^{-1}(k), \sigma_i^{-1}(k') \geq j\right\}} - \mathbb{P}_c \left[\sigma_i^{-1}(k), \sigma_i^{-1}(k') \geq j\right] \right)$, then

$$\tilde{L} - \mathbb{E}_c[\tilde{L}] = \frac{1}{2} \sum_{i=1}^{t} \left(\sum_{k,k' \in [D]} a_{kk'} (e_k - e_{k'})(e_k - e_{k'})^\top \right) := \sum_{i=1}^{t} Y_i.$$

Observe that $|a_{kk'}| \leq \frac{m_i}{D^2}$ and therefore $-L_i \leq Y_i \leq L_i$. Furthermore, note the spectral norm $\|L_i\| = \frac{m_i}{D}$, and thus $\|Y_j\| \leq \frac{m_i}{D}$. Moreover,

$$Y_j^2 = \sum_{k,k',k'' \in [D]} a_{kk'} a_{kk''} (e_k - e_{k'})(e_k - e_{k''})^\top.$$

It follows that for any vector $x \in \mathbb{R}^D$,

$$\begin{aligned}
x^\top Y_j^2 x &= \sum_{k,k',k'' \in [D]} a_{kk'} a_{kk''} (x_k - x_{k'})(x_k - x_{k''}) \\
&\le \frac{m_i^2}{D^4} \sum_{k,k',k'' \in [D]} |x_k - x_{k'}||x_k - x_{k''}| \\
&= \frac{m_i^2}{D^4} \sum_{k \in [D]} \left(\sum_{k' \in [D]} |x_k - x_{k'}| \right)^2 \\
&\le \frac{m_i^2}{D^4} \sum_{k \in [D]} \left(D \sum_{k' \in [D]} (x_k - x_{k'})^2 \right) \qquad \text{(Cauchy–Schwarz inequality)} \\
&= \frac{m_i^2}{D^3} \sum_{k,k' \in [D]} (x_k - x_{k'})^2 = \frac{2m_i}{D} x^\top L_i x.
\end{aligned}$$

Therefore, we have $Y_i^2 \le \frac{2m_i}{D} L_i \le 2L_i$, which implies $\| \sum_{i=1}^t \mathbb{E}_c[Y_i^2] \| \le \frac{2m_i}{D} \xi_D(\mathcal{G}_t) \le 2\xi_D(\mathcal{G}_t)$.

By the matrix Bernstein inequality [Tropp, 2012], we have with probability at least $1 - \frac{D\rho^2}{t^2}$,

$$\|\tilde{L}_t - \mathbb{E}_c[\tilde{L}_t]\| \le 2\sqrt{2\xi_D(\mathcal{G}_t) \log \frac{t}{\rho}} + \frac{4}{3} \log \frac{t}{\rho}.$$

By the assumption that $\xi_D(\mathcal{G}_t) \ge C \log \frac{T}{\rho}$ for some sufficiently large C, we have

$$\|\tilde{L}_t - \mathbb{E}_c[\tilde{L}_t]\| \le 4\sqrt{2\xi_D(\mathcal{G}_t) \log \frac{t}{\rho}}.$$

Combining Eq. 11 Eq. 12 with above result yields

$$\xi_2(H_t(c)) \ge \frac{1}{e^{2B}} \xi_2(\tilde{L}_t) \ge \frac{1}{e^{2B}} \Big(\frac{1}{2e^{2B}+1} \xi_2(\mathcal{G}_t) - 4\sqrt{2\xi_D(\mathcal{G}_t) \log \frac{t}{\rho}} \Big).$$

Recall $L_{\mathcal{G}_t} = \sum_{i=1}^t L_i$, and $\overline{m}_t = \frac{1}{t} \sum_{i=1}^t m_i$, we have the complete comparison graph $\mathcal{G}_t$ with edge weight of $\frac{\overline{m}_t t}{D^2}$, which implies

$$\xi_1(\mathcal{G}_t) = 1, \qquad \xi_2(\mathcal{G}_t) = ... = \xi_D(\mathcal{G}_t) = \frac{\overline{m}_t t}{D}.$$

Putting everything together, we have for any $t \ge 8(2e^{2B}+1)\sqrt{2DT \log \frac{T}{\rho}}$, with probability at least $1 - \frac{D\rho^2}{t^2}$,

$$\begin{aligned}
\xi_2(H_t(c)) &\ge \frac{1}{e^{2B}} \Big(\frac{1}{2e^{2B}+1} \xi_2(\mathcal{G}_t) - 4\sqrt{2\xi_D(\mathcal{G}_t) \log \frac{t}{\rho}} \Big) \\
&= \frac{1}{e^{2B}} \Big(\frac{1}{2e^{2B}+1} \frac{\overline{m}_t t}{D} - 4\sqrt{2\xi_D(\mathcal{G}_t) \log \frac{t}{\rho}} \Big) \\
&\ge \frac{\overline{m}_t t}{D(4e^{4B}+2e^{2B})},
\end{aligned}$$

where the last inequality holds due to the assumption of $t \ge 8(2e^{2B}+1)\sqrt{2DT \log \frac{T}{\rho}}$. □

D.2 PROOF OF LEMMA 7

Proof. By the definition of MLE loss function, we have

$$L_{\mathcal{S}_t}(c) = -\frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{m_i}\log\left(\frac{\exp(c_{\sigma_i(j)})}{\sum_{k=j}^{D}\exp(c_{\sigma_i(k)})}\right),$$

which gives the gradient of the loss function with respect to c as

$$\begin{aligned}\nabla L_{\mathcal{S}_t}(c) &= -\frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{m_i}\frac{\partial\log\left(\frac{\exp(c_{\sigma_i(j)})}{\sum_{k=j}^{D}\exp(c_{\sigma_i(k)})}\right)}{\partial c}\\ &= -\frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{m_i}\sum_{j'=j}^{D}\left(\frac{\exp(c_{\sigma_i(j')})}{\sum_{k=j}^{D}\exp(c_{\sigma_i(k)})}\right)(e_{\sigma_i(j')}-e_{\sigma_i(j)})\end{aligned}$$

Define

$$V_i^{(jj')} = \begin{cases}\frac{\exp(\bar{c}^0_{\sigma_i(j')})}{\sum_{k=j}^{D}\exp(\bar{c}^0_{\sigma_i(k)})}, & \text{if } \sigma_i(j)<\sigma_i(j')\\ -\frac{\exp(\bar{c}^0_{\sigma_i(j)})}{\sum_{k=j}^{D}\exp(\bar{c}^0_{\sigma_i(k)})}, & \text{otherwise}\end{cases}$$

By [Zhu et al., 2023], we have:

$$\mathbb{E}[V_i^{(jj')}] = 0 \quad \text{and} \quad \nabla L_{\mathcal{S}_t}(\bar{c}^0) = \frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{m_i}\sum_{j'=j+1}^{D}V_i^{(jj')}(e_j-e_{j'})$$

Note:

$$\left\|\sum_{j'=j+1}^{D}V_i^{(jj')}(e_j-e_{j'})\right\| \le \left\|\sum_{j'=j+1}^{D}V_i^{(jj')}e_j\right\| + \left\|\sum_{j'=j+1}^{D}V_i^{(jj')}e_{j'}\right\| = 2$$

And considering t interactions and m_j rounds of repeatedly selecting the most preferred objective from remaining objectives within each step $i \in [T]$, we see that $\nabla L_{\mathcal{S}_t}(\bar{c}^0)$ is the value of a discrete-time martingale at time $\sum_{i=1}^{t} m_i$, such that the martingale has initial value 0 and increments with norm bounded by 2. By Lemma 14 [Hayes, 2005], we have:

$$\mathbb{P}\left[\left\|\nabla L_{\mathcal{S}_t}(\bar{c}^0)\right\|_2 \ge \sqrt{16t\bar{m}_t\log\tfrac{t}{\rho}}\right] \le 2e^2\exp\left(-\frac{16t\bar{m}_t\log\frac{t}{\rho}}{8\sum_{i=1}^{t}m_i}\right) = 2e^2\left(\frac{\rho}{t}\right)^2. \tag{13}$$

Summing over t from 1 to T, we have:

$$\sum_{t=1}^{T}2e^2\left(\frac{\rho}{t}\right)^2 \le 2e^2\rho^2\sum_{t=1}^{T}\frac{1}{t^2} \le \frac{e^2\pi^2\rho^2}{3}.$$

Thus, with probability at least $1-\frac{e^2\pi^2\rho^2}{3}$, for any $t \in [T]$, we have:

$$\left\|\nabla L_{\mathcal{H}_t}(\bar{c}^0)\right\|_2 \le \sqrt{16t\bar{m}\log\tfrac{t}{\rho}} \quad \text{holds.}$$

□

E PROOF OF THEOREM 1

Proof. We prove the Theorem 1 using a simple toy instance with 2 arms and 2 objectives.

Let us consider an instance with two arms, where arm-1 has fixed reward $\boldsymbol{r}_1 = [0, 1]^\top$, arm-2 has fixed reward $\boldsymbol{r}_2 = [0.7, 0.7]^\top$. There two preference vectors $\boldsymbol{c}_1 = [0.1, 0.3]^\top$ and $\boldsymbol{c}_2 = [0.4, 0.6]^\top$. Apparently, under preference $\boldsymbol{c}_1$, arm-1 is the optimal arm, while for preference $\boldsymbol{c}_2$, arm-2 is the optimal.

Assume there exists a QE-only preference-aware algorithm $\mathcal{A}$ achieving sub-linear regret under preference $\boldsymbol{c}_1$. By Lemma 13, we have the expected number of pulls of arm-2 (suboptimal) given preference $\boldsymbol{c}_t$ is

$$\mathbb{E}_{\boldsymbol{c}_1}[N_{2,T}] = \sum_{t\in[T]} \mathbb{P}_{\pi_t^{\mathcal{A}}|\boldsymbol{c}_1}(a_t = 2 \mid \boldsymbol{c}_1) = o(T).$$

Due to the shift-invariant property of the Plackett–Luce (PL) model, any additive shift to the preference vector does not affect the induced ranking distribution. Specifically, since $\boldsymbol{c}_2 = \boldsymbol{c}_1 + 0.3 \cdot \mathbf{1}$, both $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ induce the same distribution over rankings. Consequently, a QE-based estimator that relies solely on ranking data cannot distinguish between $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, yielding the same mean-centered estimate in both cases. As a result, for any algorithm $\mathcal{A}$ interacting with a user whose true preference is $\boldsymbol{c}_2$, the probability of selecting any given arm remains identical to that under $\boldsymbol{c}_1$:

$$\mathbb{E}_{\boldsymbol{c}_2}[N_{2,T}] = \sum_{t=1}^{T} \mathbb{P}_{\pi_t^{\mathcal{A}}|\boldsymbol{c}_2}(a_t = 2) = \sum_{t=1}^{T} \mathbb{P}_{\pi_t^{\mathcal{A}}|\boldsymbol{c}_1}(a_t = 2) = \mathbb{E}_{\boldsymbol{c}_1}[N_{2,T}] = o(T).$$

However, recall that under preference $\boldsymbol{c}_2$, arm-2 is the optimal arm, which implies that the regret of $\mathcal{A}$ under preference $\boldsymbol{c}_2$ would be at least $\Omega(T)$, i.e.,

$$R_{\boldsymbol{c}_2}(T) = 1 \cdot \mathbb{E}_{\boldsymbol{c}_2}[N_{1,T}] = 1 \cdot (T - \mathbb{E}_{\boldsymbol{c}_2}[N_{2,T}]) > T - o(T) = \Omega(T).$$

□

F PROOF OF LEMMA 3

Proof. For any $i \in [K], d \in [D], t > 0$, by Hoeffding's Inequality (Lemma 15), we have:

$$\mathbb{P}\left(|\hat{\boldsymbol{r}}_{i,t}(d) - \boldsymbol{\mu}_i(d)| > \sqrt{\frac{\log(t/\rho)}{N_{i,t}}}\right) \le 2\exp\left(\frac{-2N_{i,t}^2 \log(t/\rho)}{N_{i,t}\sum_{\ell=1}^{N_{i,t}}(1-0)^2}\right) = 2\exp\left(-2\log(t/\rho)\right) = 2\left(\frac{\rho}{t}\right)^2, \tag{14}$$

Thus for any $i \in [K]$ and $t > 0$, with at least probability $1 - 2D\left(\frac{\rho}{t}\right)^2$, we have

$$\hat{\boldsymbol{r}}_{i,t} - \sqrt{\frac{\log(t/\rho)}{N_{i,t}}}\mathbf{1} \preceq \boldsymbol{\mu}_i \preceq \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}}\mathbf{1}.$$

And thus for any $n \in [N]$, we can derive

$$
\begin{aligned}
&\langle \hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}, \hat{\boldsymbol{r}}_{i,t} \rangle + B_{i,t}^{n(r)} + B_{i,t}^{n(c)} - \langle \overline{\boldsymbol{c}}_n, \boldsymbol{\mu}_i \rangle \\
&\qquad \geq \langle \hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}, \hat{\boldsymbol{r}}_{i,t} \rangle + B_{i,t}^{n(r)} + B_{i,t}^{n(c)} - \left\langle \overline{\boldsymbol{c}}_n, \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\rangle \\
&\qquad = \left\langle \hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}, \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\rangle + B_{i,t}^{n(c)} - \left\langle \overline{\boldsymbol{c}}_n, \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\rangle \\
&\qquad = \left\langle \hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})} - \overline{\boldsymbol{c}}_n, \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\rangle + B_{i,t}^{n(c)} \qquad (15) \\
&\qquad \underset{(a)}{\geq} -\|\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})} - \overline{\boldsymbol{c}}_n\|_{\boldsymbol{V}_{n,t}} \cdot \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\|_{(\boldsymbol{V}_{n,t})^{-1}} + B_{i,t}^{n(c)} \\
&\qquad \underset{(b)}{\geq} -\beta_t^n \cdot \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\|_{(\boldsymbol{V}_{n,t})^{-1}} + B_{i,t}^{n(c)} = 0, \\
&\qquad\qquad \text{(with probability at least } 1 - 2D\,(\rho/t)^2\text{)}
\end{aligned}
$$

where (a) holds by Cauchy-Schwarz inequality, (b) holds due to β_t^n as the radius of confidence region for preference estimation $\hat{\boldsymbol{c}}_{n,t}^{(\mathrm{HB})}$. By the convergence of sum of reciprocals of squares that $\sum_{t=1}^{\infty} t^{-2} = \frac{\pi^2}{6}$, we have with probability larger than $1 - D\rho^2\pi^2/6$, above result holds for all t, completing the proof. □

G PROOF OF LEMMA 2

Lemma 2 (Detailed Version)**.** *For any user $n \in [N]$, any $\rho \in (0,1)$ and any sufficiently large samples $t \geq 8(2e^{2B} + 1)\sqrt{2DT\log(\frac{T}{\rho})}$, with probability at least $1 - \frac{(1+D+2e^2)\rho^2\pi^2}{6}$, the estimated preference of Eq.5 in Algorithm 1 has the following upper-confidence bound*

$$
\begin{aligned}
\|\hat{\boldsymbol{c}}_{n,t}^{(HB)} - \overline{\boldsymbol{c}}_n\|_{\boldsymbol{V}_{n,t}} \leq\ & D(32e^{4B} + 16e^{2B})\sqrt{\frac{(1-\alpha)\log(t/\rho)}{\overline{m}_t t}} + B\sqrt{2(1-\alpha)\lambda} \\
& + \alpha\frac{BD}{2}\sqrt{D\log\big(\frac{1+\alpha D(t-1)\,(\lambda+1)}{\rho}\big) + \log(\frac{1+\lambda}{\lambda\rho})}
\end{aligned} \qquad (16)
$$

Proof. For notational simplicity, we fix a user n and omit the user index n as superscript or subscript. By Eq. 5, we have closed-form solution for preference estimation for bandit policy as:

$$
\hat{c}_t^{(\mathrm{HB})} = V_t^{-1}\left(\alpha\sum_{\tau=1}^{t-1} g_{a_\tau} r_{a_\tau,\tau} + (1-\alpha)U_\lambda \hat{c}_t^{(\mathrm{QE})}\right)
$$

Rearranging the equation yields

$$
\begin{aligned}
\hat{c}_t^{(\mathrm{HB})} &= V_t^{-1}\left(\alpha\sum_{\tau=1}^{t-1} g_{a_\tau} r_{a_\tau,\tau} + (1-\alpha)U_\lambda \hat{c}_t^{(\mathrm{QE})}\right) \\
&= V_t^{-1}\left(\alpha\sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top(\overline{c} + \xi_\tau) + (1-\alpha)U_\lambda \hat{c}_t^{(\mathrm{QE})}\right) \\
&= V_t^{-1}\left(\Big(\alpha\sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top + (1-\alpha)U_\lambda\Big)\overline{c} - (1-\alpha)U_\lambda\overline{c} + \alpha\sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top \xi_\tau + (1-\alpha)U_\lambda \hat{c}_t^{(\mathrm{QE})}\right) \\
&= \overline{c} - (1-\alpha)V_t^{-1}U_\lambda\overline{c} + \alpha V_t^{-1}\sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top \xi_\tau + (1-\alpha)V_t^{-1}U_\lambda \hat{c}_t^{(\mathrm{QE})},
\end{aligned}
$$

which implies that

$$\|\hat{c}_t^{(\mathrm{HB})} - \overline{c}\|_{V_t} = \left\|(1-\alpha)V_t^{-1}U_\lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c}) + \alpha V_t^{-1}\sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top \xi_\tau\right\|_{V_t}$$

Since V_t is PSD, we have:

$$\|\hat{c}_t^{(\mathrm{HB})} - \overline{c}\|_{V_t} \le (1-\alpha)\underbrace{\left\|V_t^{-1}U_\lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c})\right\|_{V_t}}_{A} + \alpha\underbrace{\left\|V_t^{-1}\sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top \xi_\tau\right\|_{V_t}}_{B} \tag{17}$$

For **Term A**, we have:

$$\begin{aligned} A &= \|V_t^{-1}U_\lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c})\|_{V_t} = \|U_\lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c})\|_{V_t^{-1}} \\ &\underset{(a)}{\le} \frac{1}{\sqrt{1-\alpha}}\|U_\lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c})\|_{U_\lambda^{-1}} = \frac{1}{\sqrt{1-\alpha}}\|(\hat{c}_t^{(\mathrm{QE})} - \overline{c})\|_{U_\lambda}, \end{aligned} \tag{18}$$

where (a) holds since $V_t^{-1} \preceq V_{t-1}^{-1} \preceq V_1^{-1} = ((1-\alpha)U_\lambda)^{-1}$. For $\|(\hat{c}_t^{(\mathrm{QE})} - \overline{c})\|_{U_\lambda}$ we have

$$\begin{aligned} \left\|\hat{c}_t^{(\mathrm{QE})} - \overline{c}\right\|_{U_\lambda} &= \sqrt{(\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top U_\lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c})} \\ &= \sqrt{(\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top (U+\lambda I)(\hat{c}_t^{(\mathrm{QE})} - \overline{c})} \\ &= \sqrt{\underbrace{(\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top U(\hat{c}_t^{(\mathrm{QE})} - \overline{c})}_{C} + \lambda\underbrace{(\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top(\hat{c}_t^{(\mathrm{QE})} - \overline{c})}_{D}} \end{aligned}$$

For **Term C**, we expand:

$$\begin{aligned} (\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top U(\hat{c}_t^{(\mathrm{QE})} - \overline{c}) &= \left(\hat{c}_t^{(\mathrm{QE})} - \left(\overline{c}^0 + \tfrac{\|\overline{c}\|_1}{D}\mathbf{1}\right)\right)^\top U\left(\hat{c}_t^{(\mathrm{QE})} - \left(\overline{c}^0 + \tfrac{\|\overline{c}\|_1}{D}\mathbf{1}\right)\right) \\ &= (\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0)^\top U(\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0) + \left(\tfrac{\|\overline{c}\|_1}{D}\right)^2 \mathbf{1}^\top U\mathbf{1} - \tfrac{2\|\overline{c}\|_1}{D}\mathbf{1}^\top U(\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0) \end{aligned}$$

Note $U = I - \frac{1}{D}\mathbf{1}\mathbf{1}^\top$, so:

$$\mathbf{1}^\top U = \mathbf{1}^\top - \tfrac{1}{D}\mathbf{1}^\top\mathbf{1}^\top = 0, \quad U\mathbf{1} = 0 \Rightarrow \text{(all-ones projection vanishes)}$$

This gives that,

$$(\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top U(\hat{c}_t^{(\mathrm{QE})} - \overline{c}) = (\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0)^\top U(\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0) \quad \text{(orthogonal to}\mathbf{1}\text{)}.$$

For **Term D**, we have:

$$\begin{aligned} \lambda(\hat{c}_t^{(\mathrm{QE})} - \overline{c})^\top(\hat{c}_t^{(\mathrm{QE})} - \overline{c}) &= \lambda\left(\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0 - \tfrac{\|\overline{c}\|_1}{D}\mathbf{1}\right)^\top\left(\hat{c}_t^{(\mathrm{QE})} - \overline{c}^0 - \tfrac{\|\overline{c}\|_1}{D}\mathbf{1}\right) \\ &\le \lambda\left(\tfrac{\|\overline{c}\|_1}{D}\mathbf{1} + \tfrac{\|\overline{c}\|_1}{D}\mathbf{1}\right)^\top\left(\tfrac{\|\overline{c}\|_1}{D}\mathbf{1} + \tfrac{\|\overline{c}\|_1}{D}\mathbf{1}\right) \\ &= 2\lambda\cdot\tfrac{\|\overline{c}\|_1^2}{D^2}\cdot\mathbf{1}^\top\mathbf{1} \\ &= 2\lambda\cdot\tfrac{\|\overline{c}\|_1^2}{D^2}\cdot D = 2\lambda\cdot\tfrac{\|\overline{c}\|_1^2}{D} \end{aligned}$$

Combining together we have

$$\begin{aligned}\left\|\hat{c}_t^{(\mathrm{QE})}-\overline{c}\right\|_{U_\lambda} &= \sqrt{(\hat{c}_t^{(\mathrm{QE})}-\overline{c}^0)^\top U(\hat{c}_t^{(\mathrm{QE})}-\overline{c}^0)+\lambda\frac{\|\overline{c}\|_1^2}{D}}\\ &\le \sqrt{(\hat{c}_t^{(\mathrm{QE})}-\overline{c}^0)^\top U(\hat{c}_t^{(\mathrm{QE})}-\overline{c}^0)}+B\sqrt{2\lambda D}\\ &\le D(32e^{4b}+16e^{2b})\sqrt{\frac{\log\frac{t}{\rho}}{\overline{m}_t t}}+B\sqrt{2\lambda D}.\end{aligned}$$

(with probability $\ge 1-\dfrac{(D+2e^2)\pi^2\rho^2}{6}$ for all t by Lemma 1.)

Plugging back Eq 18 gives:

$$A\le\frac{1}{\sqrt{1-\alpha}}\left\|\hat{c}_t^{(\mathrm{QE})}-\overline{c}\right\|_{U_\lambda}\le\frac{1}{\sqrt{1-\alpha}}\left(D(32e^{4b}+16e^{2b})\sqrt{\frac{\log\frac{t}{\rho}}{\overline{m}_t t}}+B\sqrt{2\lambda D}\right)\tag{19}$$

For **term B**, we first show that that $r_{a_\tau,\tau}^\top\xi_\tau$ is conditionally sub-Gaussian.

Let $\boldsymbol{a}_{1:t}=\{\boldsymbol{a}_\iota\}_{\iota=1}^t$ be the sequence of historical pulled actions within t steps, $\boldsymbol{g}_{1:t}=\{\boldsymbol{g}_{a_\iota,\iota}\}_{\iota=1}^t$ and $\boldsymbol{r}_{1:t}=\{\boldsymbol{r}_{a_\iota,\iota}\}_{\iota=1}^t$ be the sequences of historical overall scores and reward vectors within t steps respectively, and define the σ algebra $\mathcal{F}_{t-1}=\sigma(\boldsymbol{a}_{1:t-1},\boldsymbol{g}_{1:t-1},\boldsymbol{r}_{1:t})$. Note that for any $t\ge 1$,

$$\begin{aligned}\mathbb{E}[r_{a_\tau,\tau}^\top\xi_\tau\mid\mathcal{F}_{t-1}] &= \mathbb{E}[c_t^\top r_{a_t,t}\mid\mathcal{F}_{t-1}]-\mathbb{E}[\overline{c}^\top r_{a_t,t}\mid\mathcal{F}_{t-1}]\\ &\underset{(a)}{=}\overline{\boldsymbol{c}}^{t-1}\boldsymbol{r}_{a_t,t}-\overline{\boldsymbol{c}}^{t-1}\boldsymbol{r}_{a_t,t}=0,\end{aligned}$$

where (a) holds since $\boldsymbol{c}_t$ is independent of $\mathcal{F}_{t-1}$ and the conditional expectation fact that $\mathbb{E}(X\mid\mathcal{F})=X$ if $X\in\mathcal{F}$. Furthermore, by assumption of DB-bounded overall-reward, $-DB\le r_{a_\tau,\tau}^\top\xi_\tau\le DB$ holds almost surely, and hence we can conclude that $r_{a_\tau,\tau}^\top\xi_\tau$ is conditionally DB-sub-Gaussian. Also, since $\boldsymbol{r}_{a_t,t}$ is $\mathcal{F}_{t-1}$-measurable, define $r'_{a_\tau,\tau}=\sqrt{\alpha}r_{a_\tau,\tau}$, by Lemma 16, we have with probability at least $1-\rho$:

$$\left\|\sum_{\tau=1}^{t-1}r_{a_\tau,\tau}r_{a_\tau,\tau}^\top\xi_\tau\right\|_{V_t^{-1}}=\frac{1}{\sqrt{\alpha}}\left\|\sum_{\tau=1}^{t-1}r'_{a_\tau,\tau}r'^\top_{a_\tau,\tau}\xi_\tau\right\|_{V_t^{-1}}\le\frac{DB}{\sqrt{\alpha}}\sqrt{2\log\left(\frac{\det(V_t)^{1/2}\det(V_1)^{-1/2}}{\rho}\right)}.\tag{20}$$

Note $V_1=(1-\alpha)U_\lambda=(1-\alpha)\left(I-\frac{1}{D}\mathbf{1}\mathbf{1}^\top+\lambda I\right)$, it can be easily verified that:

$$\det(U_\lambda)=\lambda(1+\lambda)^{D-1}\Rightarrow(\text{eigenvalues: }\lambda\text{ with multiplicity }1,\quad 1+\lambda\text{ with multiplicity }D-1).$$

Since

$$V_t=(1-\alpha)U_\lambda+\sum_{\tau=1}^{t-1}r'_{a_\tau,\tau}r'^\top_{a_\tau,\tau}\le(1-\alpha)(1+\lambda)I+\sum_{\tau=1}^{t-1}r'_{a_\tau,\tau}r'^\top_{a_\tau,\tau},\quad\text{with }\|r'_{a_\tau,\tau}\|_2\le\sqrt{\alpha D},$$

we have

$$\det(V_t)\le((1-\alpha)(1+\lambda))^D\cdot\left(1+\tfrac{\alpha D(t-1)}{1+\lambda}\right)^D.$$

And note $\det(V_1)=(1-\alpha)^D\det(U_\lambda)=(1-\alpha)^D\lambda(1+\lambda)^{D-1}$, we have

$$\frac{\det(V_t)}{\det(V_1)}=\frac{1+\lambda}{\lambda}\cdot\left(1+\tfrac{\alpha D(t-1)}{1+\lambda}\right)^D.$$

Plugging back to Eq. 20 gives (with probability at least $1-\rho$):

$$\left\|\sum_{\tau=1}^{t-1}r_{a_\tau,\tau}r_{a_\tau,\tau}^\top\xi_\tau\right\|_{V_t^{-1}}\le BD\sqrt{D\log\left(\frac{1+\lambda}{\lambda\rho}\cdot\frac{1+\alpha D(t-1)/(1+\lambda)}{\rho}\right)}.$$

Combine A and B together, we have

$$\begin{aligned}\left\|\hat{c}_t^{(\mathrm{HB})}-\overline{c}\right\|_{V_t} &\le (1-\alpha)\left\|V_t^{-1}U_\lambda(\hat{c}_t^{(\mathrm{QE})}-\overline{c})\right\|_{V_t}+\alpha\left\|V_t^{-1}\sum_{\tau=1}^{t-1}r_{a_\tau,\tau}r_{a_\tau,\tau}^\top\xi_\tau\right\|_{V_t}\\ &\le (1-\alpha)\left(\frac{D(32e^{4b}+16e^{2b})}{\sqrt{1-\alpha}}\cdot\sqrt{\frac{\log\frac{t}{\rho}}{\overline{m}_t t}}+B\sqrt{\frac{2\lambda}{1-\alpha}}\right)\\ &\quad+\alpha BD\sqrt{D\log\left(\frac{1+\alpha D(t-1)/(1+\lambda)}{\rho}\right)+\log\left(\tfrac{1+\lambda}{\lambda\rho}\right)},\end{aligned}$$

thus completing the proof. □

H PROOF OF THEOREM 2

Before the detailed proofs, we first restate Theorem 2 with a detained version and show how the optimal choices of balance-weight α and regularization λ are derived.

Theorem 2 (Detailed Version)**.** *For any user $n\in[N]$, any $\rho\in(0,1)$, by setting β_t as the UCB of $\hat{\boldsymbol{c}}_t^{n(HB)}$ as Eq.* (16)*, with probability at least $1-\frac{(2+D+2e^2)\rho^2\pi^2}{6}$, we have the following regret bound*

$$R^n(T)\le\Psi_{\hat{c}}^n(T)+\Psi_{\hat{r}}^n(T)+M,\quad \textit{where}$$

$$M=\max\{\frac{4D^2K\alpha^2}{\nu_{r\downarrow}^4}\log(\frac{T}{\rho}),\frac{1}{\alpha}\log(\frac{T}{\rho}),2D(16e^{2B}+8)^2\log(\frac{T}{\rho})\},$$

$$\Psi_{\hat{c}}^n(T)=O\Big((\sqrt{\frac{D^3}{\overline{m}_T}}+B\sqrt{\frac{\lambda}{\alpha}}+BD^{\frac{3}{2}}\sqrt{\alpha\log(\frac{T}{\rho})+\alpha\log(\frac{1+\lambda}{\lambda\rho})})\sqrt{T\log(\frac{\alpha T}{\lambda})}\Big),$$

$$\Psi_{\hat{r}}^n(T)=O\Big(\sqrt{\tfrac{KD^3}{\overline{m}_T\lambda}}\log^2T+\big(BD+\tfrac{\alpha BD^{\frac{3}{2}}}{\sqrt{\lambda}}\sqrt{\log(\tfrac{T}{\rho})+\log(\tfrac{1+\lambda}{\lambda\rho})}\big)\sqrt{KT\log(\tfrac{T}{\rho})}\Big).$$

Remark 1. *With above result, we can optimize the choices of balance-weight α and regularization λ to minimize the user-specific regret. Specifically, the bound reveals the main multipliers of $\sqrt{T\log T}$. By setting $\alpha=\lambda=O(\log^{-1}(T/\rho))$, we can guarantee all such multipliers become $O(1)$, thus achieving a near optimal regret*

$$R^n(T)=O(\sqrt{T\log T}).$$

Compared to the previous PRUCB [Cao et al., 2025] in preference-aware MO-MAB, our proactive QE-based framework and algorithm achieve achieves much better performance by reducing a multiplicative $\sqrt{\log T}$ term.

Before the proof, we present some essential observations that will be used in the proof.

Claim 1. *By Hoeffding's Inequality (Lemma 15), with probability at least $1-\frac{KD\rho^2\pi^2}{3}$, we have*

$$|\boldsymbol{\mu}_i(d)-\hat{\boldsymbol{r}}_{i,t}(d)|\le\sqrt{\frac{\log(\frac{t}{\rho})}{N_{i,t}}}$$

holds for all $i\in[K],d\in[D],t\in[T]$.

Claim 2. *Since $(\alpha\boldsymbol{r}_{i,\ell}\boldsymbol{r}_{i,\ell}^\top)(m,n)\in[0,\alpha],\forall(m,n)\in[D]$ and by Hoeffding's Inequality (Lemma 15), with probability at least $1-\frac{KD(D-1)\rho^2\pi^2}{12}$, we have*

$$\mathbb{E}\left[\sum_{\ell\in\mathcal{T}_{i,t-1}}\alpha\boldsymbol{r}_{i,\ell}\boldsymbol{r}_{i,\ell}^\top\right](m,n)-\sum_{\ell\in\mathcal{T}_{i,t-1}}\left(\alpha\boldsymbol{r}_{i,\ell}\boldsymbol{r}_{i,\ell}^\top\right)(m,n)\le\alpha\sqrt{N_{i,t}\log\left(\frac{t}{\rho}\right)}$$

holds $\forall i\in[K],\forall m\in[D],\forall n\in[m,D]$.

Proof of Theorem 2. Here we focus on the user-specific regret $R^n(T)$. For notational simplicity, we fix a user n and omit the user index n as superscript or subscript. Let M be an arbitrary positive integer, we can express the user-specific regret $R^n(T)$ in a truncated form with respect to M as follows:

$$R(T)^n = \sum_{t=1}^{T} \text{regret}_t \leq O(M) + \sum_{t=M+1}^{T} \text{regret}_t, \tag{21}$$

where regret_t denotes the instantaneous regret of algorithm at step $t \in [T]$, and the last inequality holds since the fact that the instantaneous regret is bounded.

Next, we analyze the instantaneous regret over the truncated time horizon $[M+1, T]$. By Claim 1, we have

$$\boldsymbol{\mu}_{a_t^*} \preceq \hat{\boldsymbol{r}}_{a_t^*,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}}\mathbf{1}, \text{ and } \boldsymbol{\mu}_{a_t} \succeq \hat{\boldsymbol{r}}_{a_t,t} - \sqrt{\frac{\log(t/\rho)}{N_{i,t}}}\mathbf{1}. \tag{22}$$

By the definition of regret and fact above, we can derive the upper-bound of expected instantaneous regret as follows:

$$\begin{aligned}
\text{regret}_t &= \overline{\boldsymbol{c}}^\top \boldsymbol{\mu}_{a_t^*} - \overline{\boldsymbol{c}}^\top \boldsymbol{\mu}_{a_t} \\
&\underset{(a)}{\leq} \hat{\boldsymbol{c}}^\top \hat{\boldsymbol{r}}_{a_t^*,t} + B^r_{a_t^*,t} + B^c_{a_t^*,t} - \overline{\boldsymbol{c}}^\top \boldsymbol{\mu}_{a_t} \\
&\underset{(b)}{\leq} \hat{\boldsymbol{c}}^\top \hat{\boldsymbol{r}}_{a_t^*,t} + B^r_{a_t^*,t} + B^c_{a_t^*,t} - \overline{\boldsymbol{c}}^\top \left(\hat{\boldsymbol{r}}_{a_t,t} + \sqrt{\frac{\log(t/\rho)}{N_{a,t}}}\mathbf{1} \right) + 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\rho)}{N_{a,t}}} \\
&\underset{(c)}{\leq} \hat{\boldsymbol{c}}^\top \hat{\boldsymbol{r}}_{a_t,t} + B^r_{a_t,t} + B^c_{a_t,t} - \overline{\boldsymbol{c}}^\top \left(\hat{\boldsymbol{r}}_{a_t,t} + \sqrt{\frac{\log(t/\rho)}{N_{a,t}}}\mathbf{1} \right) + 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\rho)}{N_{a,t}}} \\
&= (\hat{\boldsymbol{c}} - \overline{\boldsymbol{c}})^\top \left(\hat{\boldsymbol{r}}_{a_t,t} + \sqrt{\frac{\log(t/\rho)}{N_{a,t}}}\mathbf{1} \right) + B^c_{a_t,t} + 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\rho)}{N_{a,t}}} \\
&\underset{(d)}{\leq} \|\hat{\boldsymbol{c}}_t - \overline{\boldsymbol{c}}_t\|_{\boldsymbol{V}_{t-1}} \cdot \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{a,t}}}\mathbf{1} \right\|_{\boldsymbol{V}_{t-1}^{-1}} + B^c_{a_t,t} + 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\rho)}{N_{a,t}}} \\
&\underset{(e)}{\leq} \min\left(2B^c_{a_t,t}, BD\right) + 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\rho)}{N_{a,t}}},
\end{aligned}$$

where (a) followed by Lemma 3, (b) followed by Eq. 22, (c) holds by the definition of optimization policy for arm selection, (d) holds by Cauchy-Schwarz inequality, (e) followed by the definition of $B^c_{a_t,t}$, and the fact that the instantaneous regret is at most BD. Above result implies

$$\begin{aligned}
R(T) &\leq O(M) + \sum_{t=M+1}^{T} \text{regret}_t \\
&\leq O(M) + \sum_{t=M+1}^{T} \left(\text{regret}_t^{\tilde{\boldsymbol{c}}} + \text{regret}_t^{\tilde{\boldsymbol{r}}} \right) \\
&\leq O(M) + \underbrace{\sum_{t=M+1}^{T} \min\left(2B^c_{a_t,t}, BD\right)}_{R^{\tilde{\boldsymbol{c}}}_{M+1:T}} + \underbrace{\sum_{t=M+1}^{T} 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\alpha)}{N_{a,t}}}}_{R^{\tilde{\boldsymbol{r}}}_{M+1:T}},
\end{aligned} \tag{23}$$

Next we analyze two components of $R^{\tilde{\boldsymbol{c}}}_{M+1:T}$ and $R^{\tilde{\boldsymbol{r}}}_{M+1:T}$ separately.

Step-1. (Upper-Bound over $R^{\tilde{\boldsymbol{c}}}_{M+1:T}$)

We first present two useful lemmas:

Lemma 8. *For any action sequence of $a_1, ..., a_T$ and any $M \in [0, T]$, we have*

$$\det\left(\mathbb{E}[\boldsymbol{V}_T]\right) \geq \det\left(\mathbb{E}[\boldsymbol{V}_M]\right) \prod_{t=M}^{T-1} \left(1 + \alpha \boldsymbol{\mu}_{a_t}^\top \mathbb{E}[\boldsymbol{V}_t]^{-1} \boldsymbol{\mu}_{a_t}\right).$$

Please see Appendix H.1 for the detailed proof of Lemma 8.

Lemma 9. *For any action sequence of $a_1, ..., a_T$, any weight $\alpha > 0$, and any $M \in [1, T]$, we have*

$$\log\left(\frac{\det\left(\mathbb{E}[\boldsymbol{V}_T]\right)}{\det\left(\mathbb{E}[\boldsymbol{V}_M]\right)}\right) \leq D \log\left(1 + \frac{\alpha}{\lambda(1-\alpha)}(T - M)\right).$$

Please see Appendix H.2 for the detailed proof of Lemma 9.

Under Claim 1, for any $t > 1$ and $i \in [K]$, we have

$$\left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} - \boldsymbol{\mu}_i \right\|_2 \leq \left\| 2\sqrt{\frac{\log(t/\rho)}{N_{i,t}}} \mathbf{1} \right\|_2 = 2\sqrt{D \frac{\log(t/\rho)}{N_{i,t}}} = 2\sqrt{D}\gamma_{i,t}.$$

Consequently,

$$\begin{aligned}
\left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\log(t/\rho)/N_{i,t}} \mathbf{1} \right\|_{\boldsymbol{V}_t^{-1}} - \|\boldsymbol{\mu}_i\|_{\boldsymbol{V}_t^{-1}} &\leq \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\log(t/\rho)/N_{i,t}} \mathbf{1} - \boldsymbol{\mu}_i \right\|_{\boldsymbol{V}_t^{-1}} \\
&\leq \frac{1}{\sqrt{\lambda}} \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\log(t/\rho)/N_{i,t}} \mathbf{1} - \boldsymbol{\mu}_i \right\|_2 \leq 2\sqrt{\frac{D}{\lambda}}\gamma_{i,t}, \\
\implies \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\log(t/\rho)/N_{i,t}} \mathbf{1} \right\|_{\boldsymbol{V}_t^{-1}} &\leq \|\boldsymbol{\mu}_i\|_{\boldsymbol{V}_t^{-1}} + 2\sqrt{\frac{D}{\lambda}}\gamma_{i,t}. \qquad (24)
\end{aligned}$$

Define $M = \left\lfloor \min\left\{t' \mid t\nu_{r\downarrow}^2 + \lambda \geq 2D\alpha\sqrt{Kt\log\frac{t}{\rho}}, \forall t \geq t'\right\} \right\rfloor$. Please note that for $\nu_{r\downarrow}^2 > 0$, we have $\lim_{t\to\infty} \frac{2D\alpha\sqrt{Kt\log\frac{t}{\rho}}}{\nu_{r\downarrow}^2 t} = \lim_{t\to\infty} C_1\sqrt{\frac{\log(t)-C_2}{t}} = 0$ since as t increase, $\sqrt{\log t}$ grows much slowly compared to $\sqrt{t}$. Hence for sufficiently large t', the inequality $t\nu_{r\downarrow}^2 + \lambda \geq 2D\alpha\sqrt{Kt\log\frac{t}{\rho}}, \forall t \geq t'$ holds, which implies that such an M does indeed exist. By Lemma 18, for any $t \in [M+1, T]$, we have

$$\begin{aligned}
R_{M+1:T}^{\tilde{\mathbf{c}}} &= \sum_{t=M+1}^{T} \min\left(2B_{a_t,t}^c, BD\right) \\
&= \sum_{t=M+1}^{T} \min\left(2\beta_t \left\| \hat{\boldsymbol{r}}_{a_t,t} + \sqrt{\log(t/\rho)/N_{a_t,t}} \mathbf{1} \right\|_{\boldsymbol{V}_t^{-1}}, BD\right) \\
&\underset{(a)}{\leq} \sum_{t=M+1}^{T} \min\left(2\beta_t \left(\|\boldsymbol{\mu}_{a_t}\|_{\boldsymbol{V}_t^{-1}} + \frac{1}{\sqrt{\lambda}}\gamma_{a_t,t} \right), BD\right) \qquad (25) \\
&\underset{(b)}{\leq} \sum_{t=M+1}^{T} \min\left(2\beta_t \left(\sqrt{2}\|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_t]^{-1}} + \frac{1}{\sqrt{\lambda}}\gamma_{a_t,t} \right), BD\right) \\
&\underset{(c)}{\leq} 2\sqrt{2} \underbrace{\sum_{t=M+1}^{T} \min\left(\beta_t \|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{BD}{2\sqrt{2}}\right)}_{\text{Reg}^{(1)}} + 4 \underbrace{\sum_{t=M+1}^{T} \beta_t \sqrt{\frac{D}{\lambda}}\gamma_{a_t,t}}_{\text{Reg}^{(2)}},
\end{aligned}$$

Step-1.1. (Upper-Bound over $\text{Reg}_t^{(1)}$)

Plugging the $\beta_t = D(32e^{4B} + 16e^{2B})\sqrt{\frac{(1-\alpha)\log(t/\rho)}{\overline{m}_t t}} + B\sqrt{2(1-\alpha)\lambda} + \alpha\frac{BD}{2}\sqrt{D\log\left(\frac{1+\alpha D(t-1)/(\lambda+1)}{\rho}\right) + \log(\frac{1+\lambda}{\lambda\rho})}$ (defined in Lemma 2), we have

$$\begin{aligned}\text{Reg}^{(1)} &\leq D(32e^{4B} + 16e^{2B})\sqrt{\frac{(1-\alpha)}{\overline{m}_t}}\sum_{t=M+1}^{T}\min\left(\sqrt{\frac{\log(t/\rho)}{t}}\|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{BD}{2\sqrt{2}}\right)\\ &+\left(B\sqrt{2\lambda(1-\alpha)} + \alpha\frac{BD}{2}\sqrt{D\log\left(\frac{1+\alpha DT/(\lambda+1)}{\rho}\right) + \log(\frac{1+\lambda}{\lambda\rho})}\right)\sum_{t=M+1}^{T}\min\left(\|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{BD}{2\sqrt{2}}\right)\end{aligned}$$

By Cauchy–Schwarz inequality, we have

$$\begin{aligned}\text{Reg}^{(1)} &\leq D(32e^{4B} + 16e^{2B})\sqrt{\frac{(1-\alpha)}{\overline{m}_t}}\sqrt{\sum_{t=M+1}^{T}1\cdot\underbrace{\sum_{t=M+1}^{T}\min\left(\frac{\log(t/\rho)}{t}\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)}_{I_1}}\\ &+\left(B\sqrt{2\lambda(1-\alpha)} + \alpha\frac{BD}{2}\sqrt{D\log\left(\frac{1+\alpha DT/(\lambda+1)}{\rho}\right) + \log(\frac{1+\lambda}{\lambda\rho})}\right)\sqrt{\sum_{t=M+1}^{T}1\cdot\underbrace{\sum_{t=M+1}^{T}\min\left(\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)}_{I_2}}.\end{aligned} \tag{26}$$

For $\boldsymbol{I_1}$, note for sufficiently large enough iterations $t \geq \frac{\log(T/\rho)}{\alpha}$, we have $\frac{\log(T/\rho)}{t} \leq \alpha$, thus

$$\begin{aligned}\sum_{t=M+1}^{T}\min\left(\frac{\log(t/\rho)}{t}\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right) &\leq \sum_{t=M+1}^{T}\min\left(\alpha\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)\\ &\underset{(a)}{\leq} c_0\sum_{t=1}^{T}\log\left(1+\alpha\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}\right)\\ &\underset{(b)}{\leq} c_0 D\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right),\end{aligned} \tag{27}$$

where (a) holds by using the inequality $c_0\log(1+x) \geq \frac{B^2D^2}{8\log(1+\frac{B^2D^2}{8})}\log(1+x) \geq x$, with some constant c_0 for $x \in [0, \frac{B^2D^2}{8}]$, last inequality (b) holds by chaining the results of Lemma 8 and Lemma 9, and plugging back to the result above.

For $\boldsymbol{I_2}$, similarly, we have:

$$\begin{aligned}\sum_{t=M+1}^{T}\min\left(\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right) &\leq \frac{1}{\alpha}\sum_{t=M+1}^{T}\min\left(\alpha\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)\\ &\leq \frac{c_0}{\alpha}\sum_{t=1}^{T}\log\left(1+\alpha\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}\right)\\ &\leq \frac{c_0 D}{\alpha}\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right).\end{aligned} \tag{28}$$

Combining Eq. 27 and Eq. 28 yields for $t \geq \frac{\log(T/\rho)}{\alpha}$,

$$\begin{aligned}\text{Reg}^{(1)} &\leq D(32e^{4B}+16e^{2B})\sqrt{\frac{c_0 D(1-\alpha)}{\overline{m}_t}}\sqrt{(T-M)\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right)} \\ &+\left(B\sqrt{\frac{2\lambda(1-\alpha)c_0 D}{\alpha}}+\frac{BD}{2}\sqrt{\alpha c_0 D}\sqrt{D\log\big(\frac{1+\alpha DT/(\lambda+1)}{\rho}\big)+\log(\frac{1+\lambda}{\lambda\rho})}\right)\sqrt{(T-M)\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right)}.\end{aligned} \tag{29}$$

Step-1.2. (Upper-Bound over $\text{Reg}_t^{(2)}$)

Plugging the $\beta_t = D(32e^{4B}+16e^{2B})\sqrt{\frac{(1-\alpha)\log(t/\rho)}{\overline{m}_t t}} + B\sqrt{2(1-\alpha)\lambda} + \alpha\frac{BD}{2}\sqrt{D\log\big(\frac{1+\alpha D(t-1)/(\lambda+1)}{\rho}\big)+\log(\frac{1+\lambda}{\lambda\rho})}$ (defined in Lemma 2), we have

$$\begin{aligned}\text{Reg}^{(2)} &\leq D^2(32e^{4B}+16e^{2B})\sqrt{\frac{(1-\alpha)}{m_T^{\downarrow}\lambda}}\sum_{t=M+1}^{T}\sqrt{\frac{\log^2(t/\rho)}{tN_{a_t,t}}} \\ &+\left(B\sqrt{2\lambda(1-\alpha)}+\alpha\frac{BD}{2}\sqrt{D\log\big(\frac{1+\alpha DT/(\lambda+1)}{\rho}\big)+\log(\frac{1+\lambda}{\lambda\rho})}\right)\sqrt{\frac{D}{\lambda}}\sum_{t=M+1}^{T}\sqrt{\frac{\log(t/\rho)}{N_{a_t,t}}}.\end{aligned}$$

By Lemma 19 and Lemma 20, we have

$$\begin{aligned}\text{Reg}^{(2)} &\leq O\left(D(32e^{4B}+16e^{2B})\sqrt{\frac{D(1-\alpha)}{\lambda}}\sqrt{K\log^2(\frac{T}{\rho})}\right) \\ &+O\left(\left(B\sqrt{2D(1-\alpha)}+\alpha\frac{BD}{2}\sqrt{\frac{D^2}{\lambda}\log\big(\frac{1+\alpha DT/(\lambda+1)}{\rho}\big)+\frac{D}{\lambda}\log(\frac{1+\lambda}{\lambda\rho})}\right)\sqrt{KT\log(\frac{T}{\rho})}\right).\end{aligned} \tag{30}$$

Step-2. (Upper-Bound over $R^{\tilde{\boldsymbol{r}}}_{M+1:T}$)

$$\sum_{t=M+1}^{T}2\|\overline{\boldsymbol{c}}\|_1\sqrt{\frac{\log(t/\alpha)}{N_{a,t}}} \leq 2BD\sum_{t=M+1}^{T}\sqrt{\frac{\log(t/\alpha)}{N_{a,t}}} \leq 4BD\sqrt{K(T-M)\log(\frac{T}{\rho})} \tag{31}$$

where the final inequality holds by Lemma 19.

Step-3. (Combine the Bounds) Summing up the results of Eq. 29, Eq. 30 and Eq. 31 together complete the proof. □

H.1 PROOF OF LEMMA 8

Proof of Lemma 8. Let $\boldsymbol{r}'_{a_t,t} = \sqrt{\alpha}\boldsymbol{r}_{a_t,t}$, $\boldsymbol{\mu}'_{a_t} = \sqrt{\alpha}\boldsymbol{\mu}_{a_t}$. For $\boldsymbol{V}_t$ and $\mathbb{E}[\boldsymbol{V}_t]$, by definition,

$$\boldsymbol{V}_{t+1} = \boldsymbol{V}_t + \alpha\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top \quad \text{and} \quad \boldsymbol{V}_1 = \lambda\boldsymbol{I}+\boldsymbol{U}$$
$$\mathbb{E}[\boldsymbol{V}_{t+1}] = \mathbb{E}[\boldsymbol{V}_t + \alpha\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top] = \mathbb{E}[\boldsymbol{V}_t + \boldsymbol{r}'_{a_t,t}\boldsymbol{r}'^\top_{a_t,t}] = \mathbb{E}[\boldsymbol{V}_t] + \boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^\top_{a_t} + \alpha\Sigma_{r,a_t}.$$

Since $\mathbb{E}[\boldsymbol{V}_t]$ is symmetric and positive definite, we have

$$\begin{aligned}\det\left(\mathbb{E}[\boldsymbol{V}_{t+1}]\right) &= \det\left(\mathbb{E}[\boldsymbol{V}_t] + \boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^\top_{a_t} + \alpha\Sigma_{r,a_t}\right) \\ &= \det\left(\mathbb{E}[\boldsymbol{V}_t]^{\frac{1}{2}}\left(\boldsymbol{I}+\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\left(\boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^\top_{a_t}+\alpha\Sigma_{r,a_t}\right)\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)\mathbb{E}[\boldsymbol{V}_{t-1}]^{\frac{1}{2}}\right) \\ &= \det\left(\mathbb{E}[\boldsymbol{V}_{t-1}]\right)\det\left(\boldsymbol{I}+\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\left(\boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^\top_{a_t}+\alpha\Sigma_{r,a_t}\right)\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right) \\ &\underset{(a)}{\geq} \det\left(\mathbb{E}[\boldsymbol{V}_t]\right)\left(\det\left(\boldsymbol{I}+\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^\top_{a_t}\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)+\det\left(\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\alpha\Sigma_{r,a_t}\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)\right)\end{aligned} \tag{32}$$

where (a) holds since $\left(\boldsymbol{I}+\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^{\top}_{a_t}\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)$ and $\left(\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\alpha\Sigma_{r,a_t}\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)$ are positive definite and applying Lemma 17 yields the result.

Let $\mathbb{E}[\boldsymbol{V}_{t-1}]^{-\frac{1}{2}}\boldsymbol{\mu}'_{a_t}=\boldsymbol{v}_t$, and we observe that

$$\left(\boldsymbol{I}+\boldsymbol{v}_t\boldsymbol{v}_t^{\top}\right)\boldsymbol{v}_t=\boldsymbol{v}_t+\boldsymbol{v}_t\left(\boldsymbol{v}_t^{\top}\boldsymbol{v}_t\right)=\left(1+\boldsymbol{v}_t^{\top}\boldsymbol{v}\right)\boldsymbol{v}_t.$$

Hence, $1+\boldsymbol{v}_t^{\top}\boldsymbol{v}$ is an eigenvalue of $\boldsymbol{I}+\boldsymbol{v}_t\boldsymbol{v}_t^{\top}$. And since $\boldsymbol{v}_t\boldsymbol{v}_t^{\top}$ is a rank-1 matrix, all other eigenvalue of $\boldsymbol{I}+\boldsymbol{v}_t\boldsymbol{v}_t^{\top}$ equal to 1, implying

$$\begin{aligned}\det\left(\boldsymbol{I}+\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\boldsymbol{\mu}'_{a_t}\boldsymbol{\mu}'^{\top}_{a_t}\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)&=\det\left(\boldsymbol{I}+\boldsymbol{v}_t\boldsymbol{v}_t^{\top}\right)\\&=1+\boldsymbol{v}_t\boldsymbol{v}_t^{\top}\\&=1+\left(\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\boldsymbol{\mu}'_{a_t}\right)^{\top}\left(\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\boldsymbol{\mu}'_{a_t}\right)\\&=1+\boldsymbol{\mu}'^{\top}_{a_t}\mathbb{E}[\boldsymbol{V}_t]^{-1}\boldsymbol{\mu}'_{a_t}.\end{aligned}\tag{33}$$

Combining Eq. 32 and Eq. 33, we have

$$\begin{aligned}\det\left(\mathbb{E}[\boldsymbol{V}_{t+1}]\right)&\geq\det\left(\mathbb{E}[\boldsymbol{V}_t]\right)\left(1+\boldsymbol{\mu}'^{\top}_{a_t}\mathbb{E}[\boldsymbol{V}_t]^{-1}\boldsymbol{\mu}'_{a_t}+\det\left(\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\Sigma_{r,a_t}\mathbb{E}[\boldsymbol{V}_t]^{-\frac{1}{2}}\right)\right)\\&\geq\det\left(\mathbb{E}[\boldsymbol{V}_t]\right)\left(1+\boldsymbol{\mu}'^{\top}_{a_t}\mathbb{E}[\boldsymbol{V}_t]^{-1}\boldsymbol{\mu}'_{a_t}\right)\end{aligned}$$

The solution of Lemma 8 follows from induction. □

H.2 PROOF OF LEMMA 9

Proof. By the definition of $\boldsymbol{V}_t$, we have

$$\begin{aligned}\log\left(\frac{\det\left(\mathbb{E}[\boldsymbol{V}_T]\right)}{\det\left(\mathbb{E}[\boldsymbol{V}_M]\right)}\right)&=\log\left(\det\left(\frac{\mathbb{E}[\boldsymbol{V}_M]+\sum_{t=M}^{T-1}\mathbb{E}[\alpha\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^{\top}]}{\mathbb{E}[\boldsymbol{V}_M]}\right)\right)\\&\underset{(a)}{\leq}\log\left(\det\left(\left((1-\alpha)\boldsymbol{U}_\lambda+\sum_{t=M}^{T-1}\mathbb{E}[\alpha\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^{\top}]\right)\frac{1}{1-\alpha}\boldsymbol{U}_\lambda^{-1}\right)\right)\\&=\log\left(\det\Big(\boldsymbol{I}+\frac{\alpha}{1-\alpha}\sum_{t=M}^{T-1}\mathbb{E}[\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^{\top}]\boldsymbol{U}_\lambda^{-1}\Big)\right),\end{aligned}\tag{34}$$

where (a) holds since $\det(\mathbb{E}[\boldsymbol{V}_M])\geq\det(\mathbb{E}[\boldsymbol{V}_1])=\boldsymbol{U}_\lambda$. By Sherman–Morrison formula, we have:

$$\begin{aligned}U_\lambda^{-1}&=\left(\lambda I+I-\tfrac{1}{D}\mathbf{1}\mathbf{1}^{\top}\right)^{-1}=\left((\lambda+1)I-\tfrac{1}{D}\mathbf{1}\mathbf{1}^{\top}\right)^{-1}\\&=\frac{1}{\lambda+1}I+\frac{\frac{1}{D(\lambda+1)^2}\mathbf{1}\mathbf{1}^{\top}}{1-\frac{1}{D}\cdot\frac{1}{\lambda+1}\cdot\mathbf{1}^{\top}\mathbf{1}}=\frac{1}{\lambda+1}I+\frac{\frac{1}{D(\lambda+1)^2}\mathbf{1}\mathbf{1}^{\top}}{1-\frac{1}{\lambda+1}}\end{aligned}\tag{35}$$

Let $\xi_1,...,\xi_D$ denote the eigenvalues of $\sum_{t=M}^{T-1}\mathbb{E}[\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^{\top}]\boldsymbol{U}_\lambda^{-1}$, and note:

$$
\begin{aligned}
\sum_{d=1}^{D} \xi_d &= \text{Trace}\left(\sum_{t=M}^{T-1} \mathbb{E}[\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top]\boldsymbol{U}_\lambda^{-1}\right) \\
&\underset{(a)}{=} \frac{\sum_{t=M}^{T-1}\mathbb{E}\left[\text{Trace}\left(\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top\right)\right]}{\lambda+1} + \frac{\sum_{t=M}^{T-1}\text{Trace}\left(\mathbb{E}[\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top]\mathbf{1}\mathbf{1}^\top\right)}{D(\lambda+1)\lambda} \\
&\underset{(b)}{=} \frac{\sum_{t=M}^{T-1}\mathbb{E}\left[\|\boldsymbol{r}_{a_t,t}\|_2^2\right]}{\lambda+1} + \frac{\sum_{t=M}^{T-1}\mathbb{E}[\|\boldsymbol{r}_{a_t,t}\|_1^2]}{D(\lambda+1)\lambda} \\
&\underset{(c)}{\leq} \frac{D(T-M)}{\lambda+1} + \frac{D^2(T-M)}{D(\lambda+1)\lambda} \\
&= \frac{D}{\lambda}(T-M).
\end{aligned} \tag{36}
$$

where (a) holds by Eq. 35, (c) holds since we assume $\boldsymbol{r}_t \leq [1]^D$, (b) holds since

$$
\begin{aligned}
\text{Trace}\left(\mathbb{E}[\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top]\mathbf{1}\mathbf{1}^\top\right) &= \mathbb{E}[\text{Trace}\left(\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top\mathbf{1}\mathbf{1}^\top\right)] = \mathbb{E}[(\boldsymbol{r}_{a_t,t}^\top\mathbf{1})\text{Trace}\left(\boldsymbol{r}_{a_t,t}\mathbf{1}^\top\right)] \\
&= \mathbb{E}[(\boldsymbol{r}_{a_t,t}^\top\mathbf{1})^2] = \mathbb{E}[\|\boldsymbol{r}_{a_t,t}\|_1^2].
\end{aligned}
$$

Combining Eq. 34 and Eq. 36 implies

$$
\begin{aligned}
\log\left(\frac{\det\left(\mathbb{E}[\boldsymbol{V}_T]\right)}{\det\left(\mathbb{E}[\boldsymbol{V}_M]\right)}\right) &\leq \log\left(\det\Big(\boldsymbol{I} + \frac{\alpha}{1-\alpha}\sum_{t=M}^{T-1}\mathbb{E}[\boldsymbol{r}_{a_t,t}\boldsymbol{r}_{a_t,t}^\top]\boldsymbol{U}_\lambda^{-1}\Big)\right) \\
&= \log\left(\prod_{i=1}^{D}\left(1+\frac{\alpha}{1-\alpha}\xi_i\right)\right) \\
&= D\log\left(\prod_{i=1}^{D}\left(1+\frac{\alpha}{1-\alpha}\xi_i\right)\right)^{\frac{1}{D}} \\
&\underset{(a)}{\leq} D\log\left(\frac{1}{D}\sum_{i=1}^{D}\left(1+\frac{\alpha}{1-\alpha}\xi_i\right)\right) \\
&\underset{(b)}{\leq} D\log\left(1+\frac{\alpha D(T-M)}{D\lambda(1-\alpha)}\right),
\end{aligned}
$$

where (a) follows from the inequality of arithmetic and geometric means, and (b) follows from Eq. 36. □

I PROOF OF THEOREM 3

Proof. We prove the result with a constructed instance.

Construct instances with a pair of two different zero-mean preference vectors:

$$
c = \begin{cases} \frac{\Delta}{2}, & i \leq \lfloor \frac{D}{2} \rfloor \quad \text{(upper list)} \\ -\frac{\Delta}{2}, & \lfloor \frac{D}{2} \rfloor < i \leq D \text{ (lower list)} \end{cases}, \quad c' = \begin{cases} -\frac{\Delta}{2}, & i \leq \lfloor \frac{D}{2} \rfloor \quad \text{(upper list)} \\ \frac{\Delta}{2}, & \lfloor \frac{D}{2} \rfloor < i \leq D \text{ (lower list)} \end{cases}
$$

Let $P = \frac{e^{\Delta/2}}{e^{\Delta/2}+e^{-\Delta/2}} = \sigma(\Delta)$ denote the probability of objectives with preference value $\frac{\delta}{2}$ preferred over objectives with preference of $-\frac{\delta}{2}$. Under ϵ-corruption model with arbitrary flipping, we can easily verify that the corrupted pairwise comparison probability with largest distribution shift is caused by entries swaps between upper and lower lists, resulting the pairwise comparison probability between objectives in upper list and lower list as:

$$
\widetilde{P}_{ij}^{(c)} = (1-\epsilon)P + \epsilon(1-P) = (1-2\epsilon)P + \epsilon
$$

$$\widetilde{P}_{ij}^{(c')} = (1-\epsilon)(1-P) + \epsilon P = (2\epsilon - 1)P - \epsilon + 1$$

We leverage the ranking-breaking strategy that reduces the full ranking σ into independent pairwise comparisons, where each compared objective being preferred follows a Bernoulli distribution shown above. Let $\sigma = (i_1 \succ i_2 \succ \cdots \succ i_D)$ be a full ranking over D items. We model the likelihood as:

$$P(\sigma \mid c) = \prod_{(i \succ j) \in \sigma} \mathrm{Ber}(1; \widetilde{P}_{ij}^{(c)})$$

where each pairwise outcome Y_{ij} is drawn independently as:

$$Y_{ij} \sim \mathrm{Ber}(\widetilde{P}_{ij}^{(c)})$$

Thus, the full-data distribution becomes a product of independent Bernoulli comparisons:

$$P(\{Y_{ij}\} \mid c) = \prod_{(i,j) \in S} \mathrm{Ber}(Y_{ij}; \widetilde{P}_{ij}^{(c)})$$

Now consider set S of all (i, j) pairs where $i \in [\lfloor \frac{D}{2} \rfloor]$, $j \in [\lfloor \frac{D}{2} \rfloor + 1, D]$ with $|S| = \frac{D^2}{4}$. Then the KL divergence between two corrupted models with underlying c and c' is given as

$$\mathrm{KL}(P^{(c)} \| P^{(c')}) = \sum_{(i,j) \in S} \mathrm{KL}(\mathrm{Ber}(\widetilde{P}_{ij}^{(c)}) \| \mathrm{Ber}(\widetilde{P}_{ij}^{(c')}))$$

Define $\delta = \frac{1}{2} \left| \widetilde{P}_{ij}^{(c)} - \widetilde{P}_{ij}^{(c')} \right| = (1-2\epsilon)(P - \frac{1}{2}) = (1-2\epsilon)\tanh(\Delta/2)$ for $\epsilon < \frac{1}{2}$.

Apply standard bound for Bernoulli KL divergence (since $\widetilde{P}_{ij}^{(c)} + \widetilde{P}_{ij}^{(c')} = 1$):

$$\mathrm{KL}\left(\mathrm{Ber}(\widetilde{P}_{ij}^{(c)}) \middle\| \mathrm{Ber}(\widetilde{P}_{ij}^{(c')})\right) \leq 4\delta^2 = 4(1-2\epsilon)^2 \tanh^2(\Delta/2)$$

Thus we have,

$$\mathrm{KL}(P^{(c)} \| P^{(c')}) \leq |S| \cdot 4(1-2\epsilon)^2 \tanh^2(\Delta/2) = D^2(1-2\epsilon)^2 \tanh^2(\Delta/2).$$

Using Pinsker's inequality, for L ranking samples with $\epsilon < \frac{1}{2}$, we have:

$$\mathrm{TV}((P^{(c)})^L \| (P^{(c')})^L) = \sqrt{\frac{L}{2} \mathrm{KL}(P^{(c)} \| P^{(c')})} \leq \sqrt{\frac{L}{2}} D(1-2\epsilon)\tanh(\frac{\Delta}{2})$$

Solving for $\mathrm{TV} = \frac{1}{2}$ gives,

$$\tanh(\frac{\Delta}{2}) = \frac{1}{\sqrt{2L}D(1-2\epsilon)} \implies \Delta = c_0 \cdot 2\sqrt{2} \tanh^{-1}(\frac{2}{(1-2\epsilon)D\sqrt{L}}),$$

with some constant $c_0 > 0$. Finally, let $\hat{c}_L$ be any estimator with L samples, by Le Cam's lemma we have:

$$\begin{aligned} \inf_{\hat{\boldsymbol{c}}_L} \sup_{\overline{\boldsymbol{c}}^0 \in \{c, c'\}} \mathbb{E}\big[\|\hat{\boldsymbol{c}}_L - \overline{\boldsymbol{c}}^0\|_{\mathbf{2}}\big] &\geq \frac{\|c - c'\|_2}{2}(1 - \mathrm{TV}((P^{(c)})^L \| (P^{(c')})^L)) \\ &= \frac{\sqrt{D}\Delta}{4} = c_0 \sqrt{\frac{D}{2}} \tanh^{-1}\left(\frac{2}{(1-2\epsilon)D\sqrt{L}}\right), \end{aligned}$$

with the valid region of $\frac{2}{(1-2\epsilon)D\sqrt{L}} < 1$, yields $\epsilon \leq \frac{1}{2} - O(\frac{1}{D\sqrt{L}})$. Additionally, we further have $\|\hat{\boldsymbol{c}}_L - \overline{\boldsymbol{c}}^0\|_2$ be bounded by $2B\sqrt{D}$ due to the bounded region of c. For $\epsilon > \frac{1}{2}$, it can be easily verify that $P^{(c)}$ and $P^{(c')}$ are distinguishable, leading to a constant estimator error gap. □

J PROOF OF LEMMA 4

Proof. For notational simplicity, we fix a user n and omit the user index n as superscript or subscript. Note for any number of sample t, we define the error components as $[\delta_i]_t = [\delta_1, \ldots, \delta_t] \in \mathbb{R}^{D\times t}$, where $\delta_i \in \mathbb{R}^D$. Recall we have the solution for the estimator

$$(\hat{c}_t^{(\mathrm{QE})}, [\hat{\delta}_i]_t) = \arg \min_{c\in\Theta_c^0, [\hat{\delta}_i]_t} \hat{\mathcal{L}}_{\tilde{S}_t}(c, \delta) + \lambda \sum_{j=1}^{t} \|\delta_j\|_2$$

where

$$\hat{\mathcal{L}}_{\tilde{S}_t}(c, [\hat{\delta}_i]_t) = -\frac{1}{t}\sum_{i=1}^{t}\sum_{j=1}^{m_i} \log\left(\frac{\exp(c_{\sigma_i(j)} + \delta_{i,\sigma_i(j)})}{\sum_{k=d}^{D}\exp(c_{\sigma_i(k)} + \delta_{i,\sigma_i(k)})}\right)$$

Define $\Delta_c = \hat{c}^{(\mathrm{QE})} - \overline{c}^0$, $\Delta_\delta = [\hat{\delta}_i]_t - [\delta_i^*]_t \in \mathbb{R}^{D\times T}$ and $\Delta_\delta^{\text{flat}} = \mathrm{vec}(\Delta_\delta) \in \mathbb{R}^{DT}$, where $\mathrm{vec}(\cdot)$ denotes column-major flattening. By the definition of the MLE estimator, we have:

$$\hat{\mathcal{L}}_{\tilde{S}_t}(c, [\hat{\delta}_i]_t) + \lambda\sum_{i=1}^{t}\|\hat{\delta}_i\|_2 \le \hat{\mathcal{L}}_{\tilde{S}_t}(c, [\delta_i^*]_t) + \lambda\sum_{i=1}^{t}\|\delta_i^*\|_2$$

Due to strong convexity of $\hat{\mathcal{L}}_{\tilde{S}_t}$ (by Lemma 10), we can rewrite as:

$$\begin{aligned}
\eta\sum_{i=1}^{t}\|\delta_i^*\|_2 - \eta\sum_{i=1}^{t}\|\hat{\delta}_i\|_2 &\ge \hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}_t^{(QE)}, [\hat{\delta}_i]_t) - \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t) \\
&= \hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}_t^{(QE)}, [\delta_i^*]_t + \Delta_\delta) - \hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}_t^{(QE)}, [\delta_i^*]_t) \\
&\quad + \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0 + \Delta_c, [\delta_i^*]_t) - \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t) \\
&\ge \nabla_c\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)^\top\Delta_c + \nabla_\delta\hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}_t^{(QE)}, [\delta_i^*]_t)^\top\Delta_\delta^{\text{flat}} \\
&\quad + \frac{t\overline{m}_t}{D(4e^{4B}+2e^B)}\|\Delta_c\|_2^2 + \frac{m_t^{\downarrow}}{D(4e^{4B}+2e^B)}\|\Delta_\delta^{\text{flat}}\|_2^2.
\end{aligned}$$

Reorganizing above result gives

$$\begin{aligned}
&\frac{t\overline{m}_t}{D(4e^{4B}+2e^B)}\|\Delta_c\|_2^2 + \frac{m_t^{\downarrow}}{D(4e^{4B}+2e^B)}\|\Delta_{\mathcal{S}}^{\text{flat}}\|_2^2 \\
&\le \eta\sum_{i=1}^{t}\|\hat{\delta}_i^*\|_2 - \eta\sum_{i=1}^{t}\|\hat{\delta}_i\|_2 - \nabla_c\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)^\top\Delta_c - \nabla_{\mathcal{S}}\hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}_t^{\mathrm{QE}}, [\delta_i^*]_t)^\top\Delta_\delta^{\text{flat}}.
\end{aligned}$$

Let $\mathcal{T}_t^c$ denote the set of time steps where ranking is corrupted within t samples, and $|\mathcal{T}_t^c| = N_t^c$, let $\Delta_{\delta+}^{\text{flat}} = \Delta_\delta^{\text{flat}}[\mathcal{T}_t^c]$ denote the gap of estimated δ over corrupted samples. By Lemma 11, with $\eta = \frac{2D+9p+2p^2D^2+3+6(D^2+t)}{27}$, we have:

$$\begin{aligned}
&\frac{t\overline{m}_t}{D(4e^{4B}+2e^B)}\|\Delta_c\|_2^2 + \frac{m_t^{\downarrow}}{D(4e^{4B}+2e^B)}\|\Delta_\delta^{\text{flat}}\|_2^2 \\
&\quad\le 2\sqrt{N_t^c}\eta\|\Delta_\delta^{\text{flat}}\|_2 + \|\nabla_c\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)\|_2\cdot\|\Delta_c\|_2 \\
&\quad\le \left(2\sqrt{N_t^c}\eta + \sqrt{\frac{1}{t}}\|\nabla_c\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)\|_2\right)\left(\|\Delta_{\delta+}^{\text{flat}}\|_2 + \sqrt{t}\|\Delta_c\|_2\right) \\
&\quad\le \left(2\sqrt{N_t^c}\eta + \sqrt{16D\log\tfrac{t}{\rho}}\right)\sqrt{2(\|\Delta_{\delta+}^{\text{flat}}\|_2^2 + t\|\Delta_c\|_2^2)},
\end{aligned}$$

where the last inequality holds by Eq. 13 and Cauchy–Schwarz inequality. This implies

$$\sqrt{2}D(4e^{4B}+2e^B)\left(2\sqrt{N_t^c}\eta + \sqrt{16D\log\tfrac{1}{\rho}}\right) \ge \frac{t\overline{m}_t\|\Delta_c\|_2^2 + m_t^{\downarrow}\|\Delta_\delta^{\text{flat}}\|_2^2}{\sqrt{t\|\Delta_c\|_2^2 + \|\Delta_{\delta+}^{\text{flat}}\|_2^2}}$$

$$\geq \sqrt{m_t^{\downarrow}} \cdot \sqrt{t\|\Delta_c\|_2^2 + \|\Delta_{\delta+}^{\text{flat}}\|_2^2}.$$

Note that within t steps, by Chernoff bound on Bernoulli distribution (Lemma 21) of corruption $B(\varepsilon)$, the number of corruption time satisfies:

$$\mathbb{P}\left(N_t^c \geq \varepsilon t + 2\sqrt{\varepsilon t \log\left(\frac{t}{\rho}\right)}\right) \leq \left(\frac{\rho}{t}\right)^2$$

$$\Rightarrow \quad N_t^c \geq \varepsilon t + 2\sqrt{\varepsilon t \log\left(\frac{t}{\rho}\right)} \quad \text{with high probability } (1 - \tfrac{\rho^2}{t^2}).$$

Finally, for any $t \geq 2D(16e^{2B}+8)^2 \log(\frac{t}{\rho})$, by setting $\eta \geq \frac{2D+9D+(2D^2+6)\sqrt{D^2+3}}{27}$, we have:

$$\begin{aligned}\|\Delta_c\|_2^2 + \|\Delta_{\delta+}^{\text{flat}}\|_2^2 &\leq \frac{2D^2(4e^{4B}+2e^B)^2}{tm_t^{\downarrow}}\left(2\eta\sqrt{\varepsilon t + 2\left(\varepsilon t \log(\tfrac{t}{\rho})\right)^{1/2}} + \sqrt{16D\log\tfrac{t}{\rho}}\right)^2 \\ &\leq \frac{4D^2(4e^{4B}+2e^B)^2}{m_t^{\downarrow}}\left(4\eta^2(\varepsilon + 2\sqrt{\frac{\varepsilon}{t}\log(\tfrac{t}{\rho})}) + 16D\log\left(\tfrac{t}{\rho}\right)\right)\end{aligned}$$

holds with probability at least $1 - (1 + D + 2e^2) \cdot \frac{\rho^2}{t^2}$, which complete the proof □

Lemma 10 (Strong Convexity)**.** *The loss function $\hat{\mathcal{L}}_{\tilde{S}_t}$ is strongly convex in the following sense:*

(i) *$\hat{\mathcal{L}}_{\tilde{S}_t}$ is strongly convex with respect to c at $(\overline{c}^0, [\delta_i^*]_t)$:*

$$\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0 + \Delta_c, [\delta_i^*]_t) - \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t) - \nabla_c \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)^\top \Delta_c \geq \frac{t\overline{m}_t}{D(4e^{4B}+2e^B)}\|\Delta_c\|_2^2$$

(ii) *$\hat{\mathcal{L}}_{\tilde{S}_t}$ is strongly convex with respect to δ at $(\hat{c}^{(QE)}, [\delta_i^*]_t)$:*

$$\hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}^{(QE)}, [\delta_i^*]_t + \Delta_\delta) - \hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}^{(QE)}, [\delta_i^*]_t) - \nabla_\delta \hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}^{(QE)}, [\delta_i^*]_t)^\top \Delta_\delta^{\textit{flat}} \geq \frac{m_t^{\downarrow}}{D(4e^{4B}+2e^B)}\|\Delta_\delta^{\textit{flat}}\|_2^2$$

The proof is provided in Appendix J.1.

Lemma 11. *Let*

$$\eta \geq \frac{2D + 9p + 2p^2(D^2+3) + 6(D^2+3)}{27},$$

then the following inequality holds:

$$\begin{aligned}\lambda\sum_{i=1}^t \|\delta_i^*\|_2 - \lambda\sum_{i=1}^t \|\hat{\delta}_i\|_2 - \nabla_c \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)^\top \Delta_c - \nabla_\delta \hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c}^{(QE)_t}, [\delta_i^*]_t)^\top \Delta_\delta^{\textit{flat}} \\ \leq 2\sqrt{N_t^c}\eta\|\Delta_{\delta+}^{\textit{flat}}\|_2 + \|\nabla_c \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t)\|_2 \cdot \|\Delta_c\|_2.\end{aligned} \tag{37}$$

The proof is provided in Appendix J.2.

J.1 PROOF OF LEMMA 10

Proof. Without loss of generality, we fixed the total number of sample as t and omit the index of t for the simplicity.

(i) Strongly Convex w.r.t c at $(\overline{c}^0, [\delta_i^*]_t)$

We follow some results of QE preference estimator in uncorrupted case for proof. Specifically, let

$$p_k^{(i,j)} = \frac{\mathbb{1}_k^{(i,j)} \exp(c_k + \delta_{i,k})}{\sum_{k' \in S^{(i,j)}} \exp(c_{k'} + \delta_{i,k'})}$$

denote the probability for selection in $S^{i,j}$. The Hessian with respect to c can be written as:

$$H_c(c,[\delta]_t) = \sum_{i=1}^{t}\sum_{j=1}^{m_i} H^{(i,j)}, \quad \text{where}$$

$$H^{(i,j)} = \frac{1}{2}\sum_{j,j'\in S^{i,j}} (e_k - e_{k'})(e_k - e_{k'})^\top p_k^{(i,j)} p_{k'}^{(i,j)} = \mathrm{diag}(p^{(i,j)}) - p^{(i,j)}(p^{(i,j)})^\top.$$

By Lemma 6, we have

$$\xi_2(H_c(c,[\delta]_t)) \geq \frac{t\overline{m}_t}{D(4e^{4B}+2e^{2B})}.$$

Thus for any $u \in \Theta_c^0$, we have

$$u^\top H_c(c,[\delta]_t)u \geq \xi_2(H_c(c,[\delta]_t))\|u\|_2^2 \geq \frac{t\overline{m}_t}{D(4e^{4B}+2e^{2B})}\|u\|_2^2,$$

which implies the strong convexity of $\hat{\mathcal{L}}_{\tilde{S}_t}$ w.r.t c at $(\overline{c}^0, [\delta_i^*]_t)$.

(ii) Strongly Convex w.r.t δ at $(\hat{c}^{(\mathrm{QE})}, [\delta_i^*]_t)$

Since each ranking sample σ_i is only dependent on the model of $PL(c+\delta_i)$, we have the gradient $\nabla_\delta \hat{\mathcal{L}}_{\tilde{S}_t}(c,[\delta_i^*]_t)$ splits *sample-by-sample*. In this case, the Hessian $H_\delta(c,[\delta]_t)$ is *block-diagonal*, i.e., $D \times D$ blocks for each ranking sample σ_i. And thus, we only need to check the convexity *sample-by-sample*.

Specifically, for each sample σ_i, the Hessian w.r.t δ_i is

$$H_{\delta_i}^{(i)}(c,[\delta]_t) = \sum_{j=1}^{m_i} H^{(i,j)}, \quad \text{where} \quad H^{(i,j)} = \mathrm{diag}(p^{(i,j)}) - p^{(i,j)}(p^{(i,j)})^\top \in \mathbb{R}^{D\times D}.$$

And it gives the final Hessian w.r.t $[\delta_i]_t$ as:

$$H_\delta(c,[\delta]_t) = \begin{bmatrix} H_{\delta_1}^{(1)} & & & \\ & H_{\delta_1}^{(1)} & & \\ & & \ddots & \\ & & & H_{\delta_t}^{(t)} \end{bmatrix} \in \mathbb{R}^{Dt\times Dt}.$$

Recall that each δ_i is induced by swaps, for any $i \in [t]$, we have $\delta_i \perp \mathbf{1}$, and by Lemma 6, we have

$$\xi_2(H_{\delta_i}^{(i)}(c,[\delta]_t)) \geq \frac{m_i}{D(4e^{4B}+2e^{2B})}. \tag{38}$$

Thus for any $\delta_i \in \Theta_\delta$ and $U = \begin{bmatrix} u_1 \\ \vdots \\ u_t \end{bmatrix} \in \mathbb{R}^{D\cdot t}$ with $u_i \in \mathbb{R}^D$ and $u_i \perp \mathbf{1}, \forall i \in [t]$, we have

$$u^\top H_{\delta_i}^{(i)} u \geq \xi_2(H_{\delta_i}^{(i)})\|u\|_2^2 \geq \frac{m_i}{D((4e^{4B}+2e^{2B})}\|u\|_2^2,$$

which implies the strong convexity of $H_{\delta_i}^{(i)}$, and for the final Hessian, we have

$$\begin{aligned} U^\top H_\delta(c,[\delta]_t)U &= [u_1^\top H_{\delta_1}^{(1)}, u_2^\top H_{\delta_2}^{(2)}, ..., u_t^\top H_{\delta_t}^{(t)}]U = \sum_{i=1}^{t} u_i H_{\delta_i}^{(i)} u_i \\ &\geq \sum_{i=1}^{t} \xi_2\left(H_{\delta_i}^{(i)}\right)\|u_i\|_2^2 \geq \min_{i\in[t]} \xi_2\left(H_{\delta_i}^{(i)}\right)\sum_{i=1}^{t}\|u_i\|_2^2 = \frac{m_t^\downarrow}{D(4e^{4B}+2e^{2B})}\|U\|_2^2, \end{aligned}$$

where the last equality holds by Eq. 38 and $\sum_{i=1}^{t}\|u_i\|_2^2 = \|U\|_2^2$. Above result implies the strong convexity of $\hat{\mathcal{L}}_{\tilde{S}_t}$ w.r.t $[\delta_i]_t$ at $(\overline{c}^0, [\delta_i^*]_t)$. □

J.2 PROOF OF LEMMA 11

We first present a useful lemma that will be used for proof.

Lemma 12. *For any ranking sample δ_i under any corrupted Plackett-Luce model $\mathrm{PL}(\delta_i)$, we have:*

$$\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta^*]_t)\right\|_2 \le \frac{2D^3+9p+2p^2\sqrt{D^2+3}+6\sqrt{D^2+3}}{27}$$

The proof is provided in Appendix J.3

Proof of Lemma 11. Let $\mathcal{T}_t^c$ denote the set of time steps where ranking is corrupted within t samples, and $|\mathcal{T}_t^c| = N_t^c$, let $\Delta^{\text{flat}}_{\delta+} = \Delta^{\text{flat}}_{\delta}[\mathcal{T}_t^c]$ denote the gap of estimated δ over corrupted samples. we have

$$\begin{aligned}
\sum_{i=1}^t \|\delta_i^*\|_2 - \sum_{i=1}^t \|\hat{\delta}_i\|_2 &= \sum_{i\in\mathcal{T}_t^c}\Big(\|\delta_i^*\|_2 - \|\hat{\delta}_i\|_2\Big) + \sum_{i\in[t]\setminus\mathcal{T}_t^c}\Big(\|\delta_i^*\|_2 - \|\hat{\delta}_i\|_2\Big)\\
&= \sum_{i\in\mathcal{T}_t^c}\Big(\|\delta_i^*\|_2 - \|\hat{\delta}_i\|_2\Big) - \sum_{i\in[t]\setminus\mathcal{T}_t^c}\|\hat{\delta}_i\|_2\\
&\le \sum_{i\in\mathcal{T}_t^c}\|\delta_i^* - \hat{\delta}_i\|_2 - \sum_{i\in[t]\setminus\mathcal{T}_t^c}\|\hat{\delta}_i\|_2\\
&\le \sqrt{N_t^c}\Big(\sum_{i\in\mathcal{T}_t^c}\|\delta_i^* - \hat{\delta}_i\|_2^2\Big)^{1/2} - \sum_{i\in[t]\setminus\mathcal{T}_t^c}\|\hat{\delta}_i\|_2\\
&= \sqrt{N_t^c}\cdot\|\Delta^{\text{flat}}_{\delta+}\|_2 - \sum_{i\in[t]\setminus\mathcal{T}_t^c}\|\hat{\delta}_i\|_2
\end{aligned} \tag{39}$$

On the other hand, we have

$$\begin{aligned}
&\nabla_\delta\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)^\top\Delta_\delta\\
&= \sum_{i\in\mathcal{T}_t^c}\left|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)^\top(\delta_i^* - \hat{\delta}_i)\right| + \sum_{i\in[t]\setminus\mathcal{T}_t^c}\left|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)^\top(\delta_i^* - \hat{\delta}_i)\right|\\
&\ge -\sum_{i\in\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\left\|\delta_i^* - \hat{\delta}_i\right\|_2 - \sum_{i\in[t]\setminus\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\left\|\hat{\delta}_i\right\|_2\\
&\ge -\max_{i\in\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\sum_{i\in\mathcal{T}_t^c}\left\|\delta_i^* - \hat{\delta}_i\right\|_2 - \max_{i\in[t]\setminus\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\sum_{i\in[t]\setminus\mathcal{T}_t^c}\left\|\hat{\delta}_i\right\|_2\\
&\ge -\max_{i\in\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\sqrt{N_t^c}\cdot\left\|\Delta^{(\text{flat})}_{\delta_t}\right\|_2 - \max_{i\in[t]\setminus\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\sum_{i\in[t]\setminus\mathcal{T}_t^c}\left\|\hat{\delta}_i\right\|_2
\end{aligned} \tag{40}$$

Combining above results together yields

$$\begin{aligned}
&\eta\sum_{i=1}^t\|\delta_i^*\|_2 - \eta\sum_{i=1}^t\|\hat{\delta}_i\|_2 - \nabla_c\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)^\top\Delta_c - \nabla_\delta\hat{\mathcal{L}}_{\tilde{S}_t}(\hat{c},\delta^*)^\top\Delta^{\text{flat}}_\delta\\
&\qquad\le \sqrt{N_t^c}\left(\max_{i\in\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2 + \eta\right)\left\|\Delta^{\text{flat}}_{\delta+}\right\|_2\\
&\qquad+\left(\max_{i\in[t]\setminus\mathcal{T}_t^c}\left\|\nabla^{(i)}_{\delta_i}\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2 - \eta\right)\sum_{i\in[t]\setminus\mathcal{T}_t^c}\|\hat{\delta}_i\|_2\\
&\qquad+\left\|\nabla_c\hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0,[\delta_i^*]_t)\right\|_2\cdot\|\Delta_c\|_2
\end{aligned} \tag{41}$$

By Lemma 12, setting

$$\eta \geq \frac{2D^3 + 9p + 2p^2\sqrt{D^2+3} + 6\sqrt{D^2+3}}{27} \geq \max_{i\in[t]} \left\| \nabla_{\delta_i}^{(i)} \hat{\mathcal{L}}_{\tilde{S}_t}(\overline{c}^0, [\delta_i^*]_t) \right\|_2$$

completes the proof.

□

J.3 PROOF OF LEMMA 12

Proof. For each sample δ_i, the gradient with respect to δ_i is:

$$\nabla_{\delta_i}^{(i)} \tilde{\mathcal{L}}_{\tilde{S}_t}(c, [\delta]_t) = -\sum_{j=1}^{m_i} \frac{2}{D} \left[\log \left(\frac{\exp(c_i(j) + \delta_i(j))}{\sum_{k=j}^{D} \exp(c_i(k) + \delta_i(k))} \right) \right].$$

Using the gradient of softmax log-likelihood, we get:

$$\nabla_{\delta_i}^{(i)} \tilde{\mathcal{L}}_{\tilde{S}_t}(c, [\delta]_t) = -\sum_{j=1}^{m_i} \sum_{i'=i+1}^{D} \left(\frac{\exp(c_i(j) + \delta_i(j))}{\sum_{k=i}^{D} \exp(c_i(k) + \delta_i(k))} (e_i(j) - e_i(j')) \right).$$

From this, we can bound each coordinate:

$$\left| \nabla_{\delta_i}^{(i)} \tilde{\mathcal{L}}_{\tilde{S}_t}(c, [\delta]_t)(j) \right| \leq \begin{cases} 1 - \frac{D}{\sum_{k=j}^{D} \exp(c_i(k)+\delta_i(k))} \exp(c_i(j) + \delta_i(j)) \leq 1, & j \leq m_i \\ \sum_{j'=1}^{m_i} \frac{\exp(c_i(j)+\delta_i(j))}{\sum_{k=j}^{D} \exp(c_i(k)+\delta_i(k))} \leq m_i, & j > m_i \end{cases}$$

This implies the total ℓ_2 norm is bounded by:

$$\left\| \nabla_{\delta_i}^{(i)} \tilde{\mathcal{L}}_{\tilde{S}_t}(c, [\delta]_t) \right\|_2 \leq \sqrt{m_i + m_i^2(D - m_i)} \leq \frac{2D^3 + 9p + 2p^2\sqrt{D^2+3} + 6\sqrt{D^2+3}}{27},$$

where the final inequality is the closed-form solution of

$$\max_{m_i \in [0,D]} \sqrt{m_i + m_i^2(D - m_i)}.$$

□

K PROOF OF LEMMA 5

Before the proof, we first restate Lemma 5 with a detained version.

Lemma 5 (Detailed Version)**.** *Following the settings in Lemma 4, for any user $n \in [N]$, any $\rho \in (0,1)$, with probability at least $1 - \frac{(1+D+2e^2)\rho^2\pi^2}{6}$, for all sufficiently large $t \geq 2D(16e^{2B} + 8)^2 \log(\frac{T}{\rho})$, the estimated preference of Eq.5 satisfies*

$$\begin{aligned} \|\hat{\boldsymbol{c}}_{n,t}^{(HB)} - \overline{\boldsymbol{c}}_n\|_{\boldsymbol{V}_{n,t}} \leq\ & \alpha \frac{BD}{2} \sqrt{D \log\big(\frac{1 + \alpha D(t-1)\,(\lambda+1)}{\delta}\big) + \log\big(\frac{1+\lambda}{\lambda\delta}\big)} \\ & + \sqrt{1-\alpha} \left(\frac{D(8e^{4B} + 4e^{2B})}{\sqrt{m_t^{\downarrow}}} \sqrt{\eta^2 \Big(\epsilon + 2\big(\frac{\epsilon \log(t/\rho)}{t}\big)^{1/2}\Big) + \frac{4D\log(t/\rho)}{t}} + B\sqrt{2\lambda} \right). \end{aligned}$$

Proof. The proof can be greatly simplified by applying the results in our previous proofs.

Specifically, by Eq. 17 in uncorrupted case, we have

$$\|\hat{c}_t^{(\text{HB})} - \overline{c}\|_{V_t} \leq (1-\alpha) \underbrace{\left\| V_t^{-1} U_\lambda (\hat{c}_t^{(\text{QE})} - \overline{c}) \right\|_{V_t}}_{A} + \alpha \underbrace{\left\| V_t^{-1} \sum_{\tau=1}^{t-1} r_{a_\tau,\tau} r_{a_\tau,\tau}^\top \xi_\tau \right\|_{V_t}}_{B},$$

where the term B is exactly same with the uncorrupted case since the independence with the corruption ϵ. For term A, by the derivation in Eq. 19, we have

$$A \leq \frac{1}{\sqrt{1-\alpha}} \left\| \hat{c}_t^{(\text{QE})} - \overline{c} \right\|_{U_\lambda} \leq \frac{1}{\sqrt{1-\alpha}} \sqrt{(\hat{c}_t^{(\text{QE})} - \overline{c}^0)^\top U (\hat{c}_t^{(\text{QE})} - \overline{c}^0)} + B\sqrt{2\lambda D}.$$

Note the only difference is the new QE-estimator error bound induced by the ϵ-corruption model, as has been characterized in Lemma 5. Pluging the new QE-estimator error bound of Lemma 5 yields

$$\begin{aligned} \left\| \hat{c}_t^{(\text{HB})} - \overline{c} \right\|_{V_t} \leq\ & \alpha \frac{BD}{2} \sqrt{D \log\big(\frac{1+\alpha D(t-1)\,(\lambda+1)}{\delta}\big) + \log\big(\frac{1+\lambda}{\lambda\delta}\big)} \\ & + \sqrt{1-\alpha}\Big(\frac{D(8e^{4B}+4e^{2B})}{\sqrt{m_t^\downarrow}} \sqrt{\eta^2\big(\epsilon + 2\big(\frac{\epsilon \log(t/\rho)}{t}\big)^{1/2}\big) + \frac{4D\log(t/\rho)}{t}} + B\sqrt{2\lambda D}\Big), \end{aligned}$$

completing the proof.

□

L PROOF OF THEOREM 4

Before the detailed proofs, we first restate Theorem 2 with a detained version and show how the optimal choices of balance-weight α and regularization λ are derived.

Theorem 4 (Detailed Version)**.** *For Algorithm 1 with group-wise Lasso MLE, by setting β_t as the UCB of $\hat{\mathbf{c}}_t^{n(BF)}$ in Lemma 4, setting η as Lemma 4, then for any user $n \in [N]$, any sparse corruption rate $\epsilon \in [0,1]$, with probability at least $1 - \frac{(2+D+2e^2)\rho^2\pi^2}{6}$, we have*

$$R^n(T) \leq \Psi^{n(\hat{c})}(T) + \Psi^{n(\hat{r})}(T) + M, \quad \textit{where}$$

$$M = \max\Big\{ \frac{4D^2K\alpha^2}{\nu_{r\downarrow}^4} \log(\frac{T}{\rho}), \frac{4\epsilon\eta^4}{\alpha^2} \log(\frac{T}{\rho}), \frac{4D}{\alpha} \log(\frac{T}{\rho}), 2D(16e^{2B}+8)^2 \log(\frac{T}{\rho}) \Big\}$$

$$\Psi^{n(\hat{c})}(T) = O\Big(\Big(\frac{D^2\sqrt{D+\eta\epsilon/\alpha}}{\sqrt{m_T^\downarrow}} + BD\sqrt{\frac{\lambda}{\alpha}} + BD^{\frac{3}{2}} \sqrt{\alpha \log(\frac{T}{\rho}) + \alpha \log(\frac{1+\lambda}{\lambda\rho})} \Big) \sqrt{T \log(\frac{\alpha T}{\lambda})} \Big);$$

$$\Psi^{n(\hat{r})}(T) = O\Big(\sqrt{\frac{KD^4}{m_T^\downarrow \lambda}} \log^2 T + \Big(BD + \eta \sqrt{\frac{D^3\epsilon}{m_T^\downarrow \lambda}} + \frac{\alpha B D^{\frac{3}{2}}}{\sqrt{\lambda}} \sqrt{\log(\frac{T}{\rho}) + \log(\frac{1+\lambda}{\lambda\rho})} \Big) \sqrt{KT \log(\frac{T}{\rho})} \Big).$$

Remark 2. *The parameters α and λ can be optimally tuned. By setting $\alpha = \lambda = O(\epsilon + \log^{-1}(T/\rho))$, we have*

$$R^n(T) \leq O\Big(\Big(D^2 \sqrt{\frac{D+\eta}{m_T^\downarrow}} + BD + \eta \sqrt{\frac{D^3K}{m_T^\downarrow}} \Big) \sqrt{T\log(T)} + B \sqrt{D^3K(\epsilon + \frac{1}{\log(T/\rho)})T \cdot \log(\frac{T}{\rho})} \Big)$$

Proof. The proof can be greatly simplified by applying the results in our previous proofs. Specifically, by Eq. 23 and Eq. 25, we have the regret as following formula.

$$
\begin{aligned}
R(T) &\le O(M) + \underbrace{\sum_{t=M+1}^{T} \min\left(2B^c_{a_t,t}, BD\right)}_{R^{\tilde{\boldsymbol{c}}}_{M+1:T}} + \underbrace{\sum_{t=M+1}^{T} 2\|\overline{\boldsymbol{c}}\|_1 \sqrt{\frac{\log(t/\alpha)}{N_{a,t}}}}_{R^{\tilde{\boldsymbol{r}}}_{M+1:T}} \\
&\le O(M) + 2\sqrt{2} \underbrace{\sum_{t=M+1}^{T} \min\left(\beta_t \|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{BD}{2\sqrt{2}}\right)}_{\text{Reg}^{(1)}} + 4 \underbrace{\sum_{t=M+1}^{T} \beta_t \sqrt{\frac{D}{\lambda}} \gamma_{a_t,t}}_{\text{Reg}^{(2)}} + R^{\tilde{\boldsymbol{r}}}_{M+1:T}.
\end{aligned} \tag{42}
$$

Note that the only difference is the new β_t induced by additional corruption component, leading to different $\text{Reg}^{(1)}$ and $\text{Reg}^{(2)}$. In the following, we show the derivation of these two terms while omit others since they are exactly the same with uncorrupted case.

Step-1.1. (Upper-Bound over $\text{Reg}^{(1)}_t$)

Plugging the $\beta_t = \sqrt{1-\alpha}\Big(\frac{D(8e^{4B}+4e^{2B})}{\sqrt{m^{\downarrow}_t}}\sqrt{\eta^2\big(\epsilon + 2\big(\frac{\epsilon\log(t/\rho)}{t}\big)^{1/2}\big) + \frac{4D\log(t/\rho)}{t}} + B\sqrt{2\lambda D}\Big) + \alpha\frac{BD}{2}\sqrt{D\log\big(\frac{1+\alpha D(t-1)\ (\lambda+1)}{\delta}\big) + \log(\frac{1+\lambda}{\lambda\delta})}$ (defined in Lemma 5), and by Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
\text{Reg}^{(1)} &\le D(32e^{4B}+16e^{2B})\sqrt{\frac{D(1-\alpha)}{m^{\downarrow}_t}} \sqrt{\sum_{t=M+1}^{T} 1 \cdot \underbrace{\sum_{t=M+1}^{T} \min\left(\frac{\log(t/\rho)}{t}\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)}_{I_1}} \\
&+ D(32e^{4B}+16e^{2B})\sqrt{\frac{(1-\alpha)}{m^{\downarrow}_t}} \sqrt{\sum_{t=M+1}^{T} 1 \cdot \underbrace{\sum_{t=M+1}^{T} \min\left(2\eta^2\sqrt{\frac{\epsilon\log(t/\rho)}{t}}\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)}_{I_*}} \\
&+ \left(\eta D(32e^{4B}+16e^{2B})\sqrt{\frac{(1-\alpha)\epsilon}{m^{\downarrow}_T}} + B\sqrt{2\lambda(1-\alpha)} + \alpha\frac{BD}{2}\sqrt{D\log\big(\frac{1+\alpha DT/(\lambda+1)}{\rho}\big) + \log(\frac{1+\lambda}{\lambda\rho})}\right) \\
&\quad \times \sqrt{\sum_{t=M+1}^{T} 1 \cdot \underbrace{\sum_{t=M+1}^{T} \min\left(\|\boldsymbol{\mu}_{a_t}\|^2_{\mathbb{E}[\boldsymbol{V}_t]^{-1}}, \frac{B^2D^2}{8}\right)}_{I_2}}.
\end{aligned} \tag{43}
$$

Note for I_1 and I_2, we have obtained the results in previous uncorrupted case. Next we show the bound of term I_*. When $\frac{4\epsilon\eta^4\log\frac{t}{\rho}}{\alpha^2} \le t$, we have $2\eta^2\sqrt{\frac{\epsilon\log\frac{t}{\rho}}{t}} \le \alpha$, which gives

$$
\begin{aligned}
I_* &= \sum_{t=1}^{T} \min\left(2\eta^2\sqrt{\frac{\epsilon\log(t/\rho)}{t}}\|x\|^2_{V_t^{-1}}, \frac{B^2D^2}{8}\right) \\
&\le \sum_{t=1}^{T} \min\left(\alpha\|x\|^2_{V_t^{-1}}, \frac{B^2D^2}{8}\right) \\
&\le c_0 D\log\left(1 + \frac{\alpha}{\lambda(1-\alpha)}(T-M)\right),
\end{aligned}
$$

where the last inequality holds by Eq. 27. Putting together, we have for $t \ge \max\{\frac{\log(T/\rho)}{\alpha}, \frac{4\epsilon\eta^4\log\frac{T}{\rho}}{\alpha^2} \le t\}$,

$$
\begin{aligned}
\text{Reg}^{(1)} \leq\ & D(32e^{4B}+16e^{2B})(D+\sqrt{D})\sqrt{\frac{c_0(1-\alpha)}{m_T^{\downarrow}}}\sqrt{(T-M)\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right)} \\
&+\eta D(32e^{4B}+16e^{2B})\sqrt{\frac{(1-\alpha)\epsilon}{\alpha m_T^{\downarrow}}}\sqrt{(T-M)\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right)} \\
&+\left(B\sqrt{\frac{2\lambda(1-\alpha)c_0D}{\alpha}}+\frac{BD}{2}\sqrt{\alpha c_0 D}\sqrt{D\log\big(\frac{1+\alpha DT/(\lambda+1)}{\rho}\big)+\log(\frac{1+\lambda}{\lambda\rho})}\right) \\
&\times\sqrt{(T-M)\log\left(1+\frac{\alpha}{\lambda(1-\alpha)}(T-M)\right)}.
\end{aligned}
\tag{44}
$$

Step-1.2. (Upper-Bound over $\text{Reg}_t^{(2)}$)

Plugging the new β_t (defined in Lemma 5), we have

$$
\begin{aligned}
\text{Reg}^{(2)} \leq\ & D^2(32e^{4B}+16e^{2B})\sqrt{\frac{(1-\alpha)}{m_T^{\downarrow}\lambda}}\sum_{t=M+1}^{T}\sqrt{\frac{\log^2(t/\rho)}{tN_{a_t,t}}} \\
&+D(16e^{4B}+8e^{2B})\sqrt{\frac{D(1-\alpha)}{m_T^{\downarrow}\lambda}}\eta\sum_{t=M+1}^{T}\underbrace{\sqrt{\epsilon+2\big(\frac{\epsilon\log(t/\rho)}{t}\big)^{1/2}}}_{A}\sqrt{\frac{\log(t/\rho)}{N_{a_t,t}}} \\
&+\left(B\sqrt{2\lambda(1-\alpha)}+\alpha\frac{BD}{2}\sqrt{D\log\big(\frac{1+\alpha DT/(\lambda+1)}{\rho}\big)+\log(\frac{1+\lambda}{\lambda\rho})}\right)\sqrt{\frac{D}{\lambda}}\sum_{t=M+1}^{T}\sqrt{\frac{\log(t/\rho)}{N_{a_t,t}}}.
\end{aligned}
\tag{45}
$$

Note term A is the additional term introduced by corruption. We can see that $O(1)\leq O(A)\leq O(\frac{\log(t/\rho)}{t})$, thus we have the second term can be bounded as

$$
O\left(D(16e^{4B}+8e^{2B})\sqrt{\frac{D\epsilon(1-\alpha)}{m_T^{\downarrow}\lambda}}\eta\sum_{t=M+1}^{T}\sqrt{\frac{\log(t/\rho)}{N_{a_t,t}}}\right)=O\left(\eta D\sqrt{\frac{D\epsilon}{\lambda}}\sqrt{KT\log(\frac{T}{\rho})}\right).
$$

Pluging above result back to Eq. 45, and combining Eq. 45, Eq. 45, Eq. 42 and Eq. 31 conclude the proof of $R(T)$. □

M AUXILIARY LEMMAS

Lemma 13 (Variant of Lemma 7 in [Jun et al., 2018]). *Assume that a bandit algorithm enjoys a sub-linear regret bound, then $\mathbb{E}[N_{i,T}]=o(T),\forall i\neq a^*$.*

Lemma 14 (Theorem 1.8 of [Hayes, 2005]). *Suppose that $S_r=\sum_{t=1}^{r}X_t$ is a martingale where $X_1,X_2,\ldots,X_m$ take values in $\mathbb{R}^n$ and are such that $\mathbb{E}[X_t]=0$ and $\|X_t\|_2\leq D$ for all t, for $D>0$. Then, for every $x>0$,*

$$
\mathbb{P}\left[\|S_r\|_2>x\right]\leq 2e^2\exp\left(-\frac{x^2}{2rD^2}\right).
$$

Lemma 15 (Hoeffding's inequality for general bounded random variables, Theorem 2.2.6 of [Vershynin, 2018]). *Given independent random variables $\{X_1,...,X_m\}$ where $a_i\leq X_i\leq b_i$ almost surely (with probability 1) we have:*

$$
\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}X_i-\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[X_i]\geq\epsilon\right)\leq\exp\left(\frac{-2\epsilon^2m^2}{\sum_{i=1}^{m}(b_i-a_i)^2}\right).
$$

Lemma 16 (Self-Normalized Bound for Vector-Valued Martingales [Abbasi-Yadkori et al., 2011])**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and conditionally R-sub-Gaussian for some $R \geq 0$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda \eta_t} \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^\top, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

$$\|S_t\|^2_{\overline{V}_t^{-1}} \leq 2R^2 \log\left(\frac{\det(\overline{V}_t)^{1/2} \det(V)^{-1/2}}{\delta}\right).$$

Lemma 17 (Determinant of Symmetric PSD Matrices Sum)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric and positive definite matrix, and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be a symmetric and positive (semi-) definite matrix. Then we have*

$$\det(\boldsymbol{A}+\boldsymbol{B}) \geq \det(\boldsymbol{A}) + \det(\boldsymbol{B})$$

Proof.

$$\det(\boldsymbol{A}+\boldsymbol{B}) = \det(\boldsymbol{A}) \det\left(\boldsymbol{I} + \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right). \tag{46}$$

Let $\lambda_1, ..., \lambda_n$ be the eigenvalues of $\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}$. Since $\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}$ is positive (semi-) definite, we have $\lambda_i \geq 0, \forall i \in [n]$, which implies

$$\det\left(\boldsymbol{I} + \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right) = \prod_{i=1}^{n}(1+\lambda_i) \geq 1 + \prod_{i=1}^{n}\lambda_i = \det(\boldsymbol{I}) + \det\left(\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right). \tag{47}$$

Combining Eq.46 with Eq. 47 concludes the proof.

□

Lemma 18 (Lemma 10 in [Cao et al., 2025])**.** *Let $M = \left\lfloor \min\left\{t' \mid t\nu^2_{r^\downarrow} + \lambda \geq 2D\omega\sqrt{Kt\log\frac{t}{\alpha}}, \forall t \geq t'\right\}\right\rfloor$, under the conditions of Claim 1 and Claim 2, for any $t \geq M+1$, and any $\boldsymbol{\mu} \in \mathbb{R}^D$, we have*

$$\boldsymbol{\mu}^\top \boldsymbol{V}_{t-1}^{-1}\boldsymbol{\mu} \leq 2\boldsymbol{\mu}^\top \mathbb{E}[\boldsymbol{V}_{t-1}]^{-1}\boldsymbol{\mu}.$$

Lemma 19 (Lemma 12 in [Cao et al., 2025])**.** *For any $M \geq 0$, we have*

$$\sum_{t=M+1}^{T}\sqrt{\frac{\log\left(\frac{t}{\rho}\right)}{N_{a_t,t}}} \leq 2\sqrt{K(T-M)\log\left(\frac{T}{\rho}\right)}.$$

Lemma 20. *For any $M \geq 0$, we have*

$$\sum_{t=M+1}^{T}\sqrt{\frac{\log^2(t/\rho)}{tN_{a,t}}} = \mathcal{O}\left(\sqrt{K}\cdot\log^2(T/\rho)\right).$$

Proof. By the Cauchy-Schwarz inequality, we have:

$$\sum_{t=M+1}^{T}\sqrt{\frac{\log^2(t/\rho)}{tN_{a_t,t}}} \leq \sqrt{\sum_{t=M+1}^{T}\frac{\log^2(t/\rho)}{t}}\cdot\sqrt{\sum_{t=M+1}^{T}\frac{1}{N_{a_t,t}}}.$$

We bound the first sum term by an integral:

$$\sum_{t=M+1}^{T} \frac{\log^2(t/\rho)}{t} \leq \int_{M}^{T} \frac{\log^2(x/\rho)}{x}\, dx.$$

Let $u = \log(x/\rho) \Rightarrow x = \rho e^u, dx = \rho e^u du$. Then:

$$\int_{M}^{T} \frac{\log^2(x/\rho)}{x} dx = \int_{\log(M/\rho)}^{\log(T/\rho)} u^2\, du = \left[\frac{u^3}{3}\right]_{\log(M/\rho)}^{\log(T/\rho)}.$$

So:

$$\sum_{t=M+1}^{T} \frac{\log^2(t/\rho)}{t} = \mathcal{O}\left(\log^3(T/\rho)\right).$$

We then bound the second term. Note that a_t is the selected arm at time t, and for each time t, $N_{a_t,t}$ is incremented by 1. Thus:

$$\sum_{t=1}^{T} \frac{1}{N_{a_t,t}} \leq \sum_{i=1}^{K} \sum_{s=1}^{N_{i,T}} \frac{1}{s} \leq \sum_{i=1}^{K} \log N_{i,T} \leq K \log T,$$

which gives:

$$\sum_{t=M+1}^{T} \frac{1}{N_{a_t,t}} \leq \mathcal{O}(K \log T).$$

Putting both bounds together:

$$\sum_{t=M+1}^{T} \sqrt{\frac{\log^2(t/\rho)}{tN_{a_t,t}}} \leq \sqrt{\mathcal{O}(\log^3(T/\rho))} \cdot \sqrt{\mathcal{O}(K \log T)} = \mathcal{O}(\sqrt{K} \cdot \log^2(T/\rho)).$$

□

Lemma 21 (Standard Chernoff Bound (Additive Form))**.** *Let S_T be the sum of t i.i.d. Bernoulli random variables with mean ε. Then for any $\rho > 0$, the following inequality holds:*

$$\mathbb{P}\left(S_t \geq \varepsilon t + \sqrt{2\varepsilon t \log\left(\tfrac{1}{\rho}\right)}\right) \leq \rho.$$