Provably Efficient Multi-Objective Bandit Algorithms Under Preference-Centric Customization

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
cao.1378@osu.edu, mshi24@buffalo.edu, shroff.11@osu.edu

Abstract

Multi-objective multi-armed bandit (MO-MAB) problems traditionally aim to achieve Pareto optimality. However, real-world scenarios often involve users with varying preferences across objectives, resulting in a Pareto-optimal arm that may score high for one user but perform quite poorly for another. This highlights the need for *customized learning*, a factor often overlooked in prior research. To address this, we study a *preference-aware* MO-MAB framework in the presence of explicit user preference. It shifts the focus from achieving Pareto optimality to further optimizing within the Pareto front under preference-centric customization. To our knowledge, this is the first theoretical study of customized MO-MAB optimization with explicit user preferences. Motivated by practical applications, we explore two scenarios: unknown preference and hidden preference, each presenting unique challenges for algorithm design and analysis. At the core of our algorithms are *preference estimation* and *preference-aware optimization* mechanisms to adapt to user preferences effectively. We further develop novel analytical techniques to establish near-optimal regret of the proposed algorithms. Strong empirical performance confirm the effectiveness of our approach.

Full version — https://arxiv.org/pdf/2502.13457

1 Introduction

Multi-objective multi-armed bandit (MO-MAB) is an important extension of standard MAB (Drugan and Nowe 2013). In MO-MAB problems each arm is associated with a D-dimensional reward vector. In this environment, objectives could conflict, leading to arms that are optimal in one dimension, but suboptimal in others. A natural solution is utilizing Pareto ordering to compare arms based on their rewards (Drugan and Nowe 2013). Specifically, for any arm $i \in [K]$, if its expected reward $\boldsymbol{\mu}_i$ is non-dominated by that of any other arms, arm i is deemed to be Pareto optimal. The set containing all Pareto optimal arms is denoted as Pareto front $\mathcal{O}^*$. Formally, $\mathcal{O}^* = \{i \mid \boldsymbol{\mu}_j \nsucc \boldsymbol{\mu}_i, \forall j \in [K] \setminus i\}$, where $\boldsymbol{u} \succ \boldsymbol{v}$ holds if and only if $\boldsymbol{u}(d) > \boldsymbol{v}(d), \forall d \in [D]$. The performance is then evaluated by Pareto regret, which measures the cumulative minimum distance between the learner's obtained rewards and rewards of arms within $\mathcal{O}^*$ (Drugan and Nowe 2013). However, achieving low Pareto regret alone overlooks that users ultimately seek options aligned with their individual preferences. As the example depicted in Fig. 1, given multiple Pareto optimal restaurants, one user may give a higher preference to quality, while another user may give a higher preference to affordibility. This means that *user preferences* need to be accounted for in the MO-MAB problem set up in order to choose the right solution on the Pareto front $\mathcal{O}^*$. This is the focus of this paper.

Numerous MO-MAB studies have been conducted but **most of them achieve Pareto optimality via an arm selection policy that is uniform across all users**, which we refer to as a *global policy*. One representative line of research focuses on efficiently estimating the entire Pareto front $\mathcal{O}^*$, and the action is *randomly* chosen on the estimated Pareto front (Drugan and Nowe 2013; Turgay, Oner, and Tekin 2018; Lu et al. 2019; Drugan 2018; Balef and Maghsudi 2023). Another line of research transforms the D-dimensional reward into a scalar using a scalarization function, which targets a specific Pareto optimal arm solution without the costly estimation of entire Pareto front Drugan and Nowe (2013); Busa-Fekete et al. (2017); Mehrotra, Xue, and Lalmas (2020); Xu and Klabjan (2023). These studies construct the scalarization function in a user-agnostic manner, causing the target arm solution to remain the same across different users (see Appendix A for a more detailed related work discussion). However, *simply achieving Pareto optimality using a global policy may not yield favorable outcomes, since, as mentioned earlier, users often have diverse preferences across different objectives.* Consider Fig. 1(a), where two users with different preferences interact with a conversational recommender to choose a restaurant based on multi-dimensional rewards (e.g., price, taste, service). Clearly, restaurants A, B, and C are Pareto optimal, as none of their rewards are dominated by others. Previous research using a global policy would either randomly recommend a restaurant from A, B, or C, or select one based on a fixed global criterion to achieve Pareto optimality. However, while recommending a restaurant like B might lead to positive feedback from user-1, it is likely to result in a low reward rating from user-2, who prefers an economical meal, since restaurant B is expensive. In contrast, Fig. 1(b) illustrates that when the system accurately captures user preferences (e.g., user-1 prefers a tasty meal, while user-2 prefers a cheap meal), it can select options more likely to

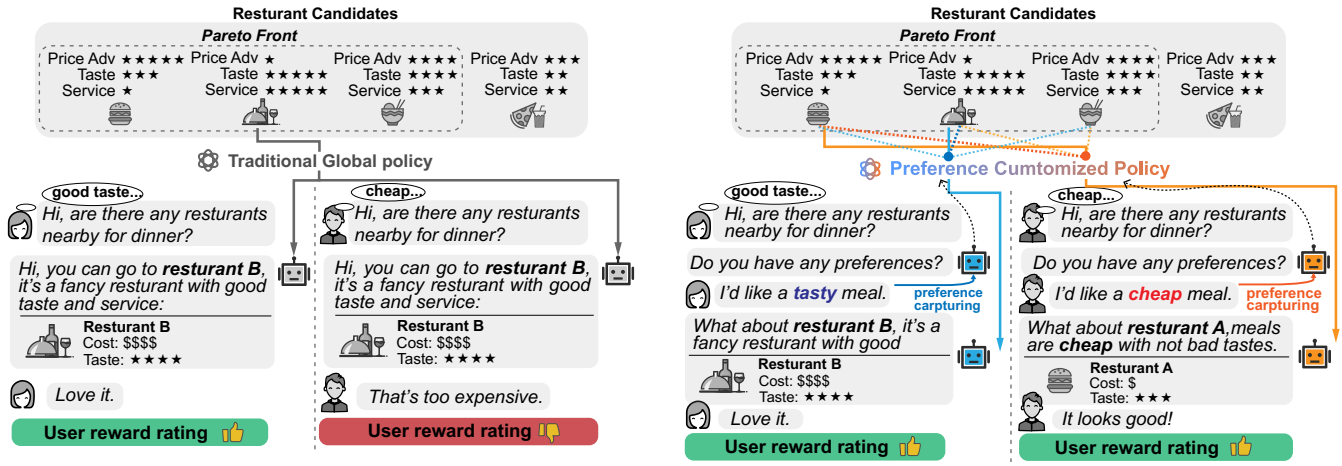


Figure 1: A scenario of users interacting with a conversational recommender for restaurant recommendation. (a) Recommender achieves Pareto optimality but receives low rating from user. (b) Recommendations with high users' ratings when the recommender captures users' preferences and aligns optimization with preferences.

receive positive reward ratings from both users. *Therefore, we argue that optimizing MO-MAB should be customized based on the user preferences rather than solely aiming for Pareto optimality with a global policy.*

To fill this gap, we introduce a formulation of MO-MAB problem, where each user is associated with a D-dimensional *preference vector*, with each element representing the user's preference for the corresponding objective. Formally, in each round t, user incurs a stochastic preference $\boldsymbol{c}_t \in \mathbb{R}^D$. The player selects an arm a_t and observes a stochastic reward $\boldsymbol{r}_{a_t,t} \in \mathbb{R}^D$. We define the scalar *overall-reward* as the inner product of arm reward $\boldsymbol{r}_{a_t,t}$ and user preference $\boldsymbol{c}_t$. The learner's goal is to maximize the overall-reward accrued over a given time horizon. We term this problem as *Preference-Aware MO-MAB* (PAMO-MAB).

While interactive user modeling and customized optimization cross multiple objectives present promising experimental results in some areas including recommendation (Xie et al. 2021), ranking (Wanigasekara et al. 2019), and more (Reymond et al. 2024), there are no theoretical studies on MO-MAB customization under explicit user preferences. Particularly, two open questions remain: *(1) how to develop provably efficient algorithms for customized optimization under different preference structure (e.g., unknown preference, hidden preference)? (2) how does the additional user preferences impact the overall performance?*

Our contributions are summarized as follows.

- We make the first effort to address the open questions above. Motivated by real applications, we consider PAMO-MAB under two preference structures: unknown preference with feedback, and unknown preference without feedback (hidden preference), with tailored algorithms that are proven to achieve near-optimal regret in each case. These approaches is built on a designed general algorithmic backbone that introduces two key components: preference estimation and preference-aware optimization, to enable effective learning and decision-making under preference-centric customization. The expressions of our results are in an explicit form that capture a clear dependency on preference. *To the best of our knowledge, this is the first work that explicitly showcases the fundamental impact of user preference in the regret optimization of MO-MAB problems.*
- For the general hidden preference case, we propose a novel near-optimal algorithm PRUCB-HP that addresses the unique challenges of hidden PAMO-MAB with two key designs: (1) A weighted least squares-based hidden preference learner, with weights set as the inverse squared ℓ_2-norm of reward observations, to resolve the random mapping issue caused by random preferences, and (2) A *dual-exploration* policy with novel bonus design to balance the trade-off between *local exploration* for identifying better reward arms and *global exploration* for refining preference learning. Additionally, we show that the unknown preference case can be viewed as a special instance of the hidden setting. A simplified variant, PRUCB-UP, naturally emerges under this setup, reusing the same backbone with reduced uncertainty and simplified estimation, while still achieving near-optimal regret cross all users.
- Extensive experiments consistently validate the effectiveness of our algorithms in estimating preferences and rewards online, as well as in optimizing the overall reward.

2 Problem Formulation

We consider MO-MAB with N users, K arms and D objectives. At each round $t \in [T]$, each user $n \in [N]$ is presented with an arm set $\mathcal{A}_t^n \subseteq [K]$, which may differ across users and time. The learner chooses an arm a_t^n for user n and observes a stochastic D-dimensional *reward vector* $\boldsymbol{r}_{a_t^n,t} \in \mathcal{R} \subseteq \mathbb{R}^D$, which we refer to as *reward*. For the reward, we make the following standard assumption:

Assumption 1 (Bounded stochastic reward)**.** For $i \in [K], t \in [T], d \in [D]$, each reward entry $\boldsymbol{r}_{i,t}(d)$ is independently drawn from a **fixed** but **unknown** distribution with

mean $\boldsymbol{\mu}_i(d)$ and variance $\sigma^2_{r,i,d}$, satisfying $\boldsymbol{r}_{i,t}(d) \in [0,1]$, and $\sigma^2_{r,i,d} \in [\sigma^2_{r\downarrow}, \sigma^2_{r\uparrow}]$, where $\sigma^2_{r\downarrow}, \sigma^2_{r\uparrow} \in \mathbb{R}^+$.

User preferences. At each round t, we consider each user n to be associated with a stochastic D-dimensional *preference vector* $\boldsymbol{c}^n_t \in \mathcal{C} \subseteq \mathbb{R}^D$, indicating the user preferences across the D objectives. We refer to this vector as *preference* for short. Specifically, we make the following assumptions:

Assumption 2 (Bounded stochastic preference)**.** For $t \in [T], d \in [D], n \in [N]$, each preference entry $\boldsymbol{c}^n_t(d)$ is independently drawn from a fixed distribution **(either known or unknown)** with mean $\overline{\boldsymbol{c}}^n(d)$ and variance $\sigma^2_{c^n,d}$, satisfying $\boldsymbol{c}^n(d) \geq 0$, $\|\boldsymbol{c}^n_t\|_1 \leq \delta$, $\sigma^2_{c^n,d} \in [0, \sigma^2_c]$.

Assumption 3 (Independence)**.** For any $t \in [T]$, $n \in [N]$, $i \in [K]$, $d_1, d_2 \in [D]$, $\boldsymbol{r}_{i,t}(d_1)$, $\boldsymbol{c}^n_t(d_2)$ are independent.

Assumption 3 is common in real applications since $\boldsymbol{c}_t$ and $\boldsymbol{r}_t$ are inherently determined by independent factors: user characteristics and arm properties. For example, an individual user's preferences do not influence a restaurant's location, environment, pricing level, etc., and vice versa.

Preference-aware reward. We define an *overall-reward* as the *inner product* of arm's reward and user's preference, which models the user reward rating under their preferences. In each round t, for each user n, the overall-reward score $g_{a^n_t,t}$ for the chosen arm a^n_t is defined as:

$$g_{a^n_t,t} = \boldsymbol{c}^{n\top}_t \boldsymbol{r}_{a^n_t,t}. \tag{1}$$

To evaluate the learner's performance, we define regret as the cumulative gap in overall-reward between selecting the optimal arm for each user at each round and the actual learner's policy across the entire user set:

$$R(T) = \sum_{t=1}^{T} \sum_{n=1}^{N} \overline{\boldsymbol{c}}^{\boldsymbol{n}\top} (\boldsymbol{\mu}_{a^{n*}_t} - \boldsymbol{\mu}_{a^n_t}), \tag{2}$$

$a^{n*}_t = \arg\max_{i \in \mathcal{A}^n_t} \overline{\boldsymbol{c}}^{\boldsymbol{n}\top} \boldsymbol{\mu}_i$ is the optimal arm for user n at round t. The goal is to minimize the regret $R(T)$. We term this problem as *Preference-Aware MO-MAB* (PAMO-MAB).

3 A Lower Bound

In the following, we develop a lower bound (Proposition 1) on the defined regret for PAMO-MAB. Such a lower bound will quantify how difficult it is to control regret without preference-adaptive policies under PAMO-MAB. Firstly, we present a definition characterizing a class of MO-MAB algorithms that are "preference-free".

Definition 1 (Preference-Free Algorithm)**.** Let $\mathbf{c}^\top = \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_t\} \in \mathbb{R}^{D \times t}$ be the preference sequence up to t episode with mean $\overline{\boldsymbol{c}}$. Let $\pi^{\mathcal{A}}_t$ be the policy of algorithm $\mathcal{A}$ at time t for selecting arm a_t. Then $\mathcal{A}$ is defined as preference-free if its policy $\pi^{\mathcal{A}}_t$ is independent of $\mathbf{c}^\top$ and $\overline{\boldsymbol{c}}$, i.e., $\mathbb{P}_{\pi^{\mathcal{A}}_t}(a_t = i \mid \mathbf{c}^\top, \overline{\boldsymbol{c}}) = \mathbb{P}_{\pi^{\mathcal{A}}_t}(a_t = i)$ for all arms $i \in [K]$ and all episodes $t \in [T]$.

To our knowledge, most existing algorithms in theoretical MO-MAB studies (Drugan and Nowe 2013; Busa-Fekete et al. 2017; Xu and Klabjan 2023; Hüyük and Tekin 2021; Cheng et al. 2024) fall within the class of preference-free algorithms, which employ a global policy for arm selection, while neglecting users' preferences.

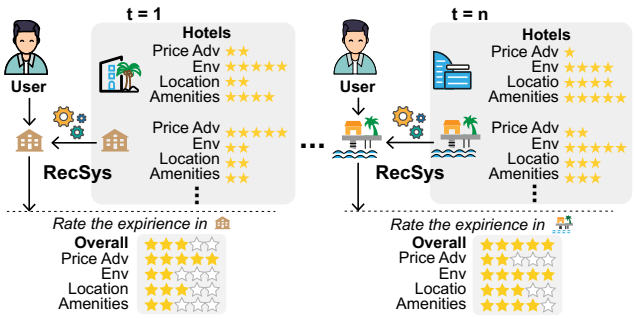


Figure 2: A scenario of user's preference feedback is not explicitly provided (hidden preference).

Proposition 1. *Assume an MO-MAB environment contains multiple objective-conflicting arms, i.e., $|\mathcal{O}^*| \geq 2$, where $\mathcal{O}^*$ is the Pareto Optimal front. Then, for any preference-free algorithm, there exists a subset of users with distinct preferences such that the regret $R(T) = \Omega(T)$.*

Proposition 1 shows that for PAMO-MAB problem with $|\mathcal{O}^*| \geq 2$, sub-linear regret is unachievable for preference-free algorithms. This is because, for any arm $i \in \mathcal{O}^*$ that is optimal in one user preference subset $\mathcal{C}^+$, there exists another user preference subset $\mathcal{C}^-$ where arm i becomes suboptimal, while preference-free algorithms cannot adapt their policies to varying preference across the entire space $\mathcal{C}$. Please see Appendix C for the detailed proof. We therefore ask the following question: ***Can we design preference-adaptive algorithms that achieve sub-linear regret for PAMO-MAB?*** The answer is **yes.** In the following, we analyze PAMO-MAB under two scenarios: hidden preference and preference feedback provided, demonstrating that with preference adaptation, sub-linear regret can indeed be achieved.

4 General Case with Hidden Preference

We first consider a more practical but more challenge scenario where only feedback on the reward and overall reward is observable, while preference feedback is hidden. For instance, in hotel surveys, customers often provide ratings on specific objectives (e.g., price, location, environment, amenities) along with an overall rating (as depicted in Fig. 2). In such cases, user preferences can be inferred from the latent relationship between the overall rating and the individual objective ratings. Formally, at each round t, the learner selects an arm $a_t \in \mathcal{A}^n_t$ for each user n, and observes the reward vector $\boldsymbol{r}_{a^n_t} \in \mathbb{R}^D$, and the overall-reward score $g_{a^n_t,t}$.

Within this framework, we adhere to the original Assumption 1 regarding rewards. It is worth noting that, in many real-world applications like hotel rating systems, the overall rating often shares the same scale as individual objective ratings. Therefore, we assume in this problem that the bound on the overall reward is identical to that of the individual rewards. This introduces one additional assumption and one revised assumption, as outlined below:

Assumption 4. For $t \in [T]$, $n \in [N]$, $a^n_t \in [K]$, the overall-reward score satisfies $g_{a^n_t,t} \in [0,1]$.

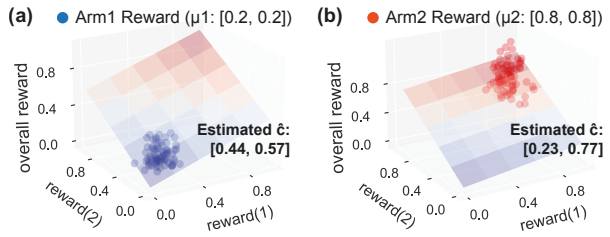


Figure 3: A 2-dimensional hidden preference PAMO-MAB toy example with mean preference $\overline{\boldsymbol{c}} = [0.5, 0.5]$, illustrating preference estimate $\hat{\boldsymbol{c}}$ via linear regression using reward data from (a) Arm-1 (dominated mean reward: $[0.2, 0.2]$) and (b) Arm-2 (Pareto-optimal mean reward: $[0.8, 0.8]$).

Assumption 5. For $t \in [T]$, $n \in [N]$, the stochastic preference is bounded and satisfies $\|\boldsymbol{c}_t^n\|_1 \leq 1$. Without loss of generality, we assume $\boldsymbol{c}_t^n(d)$ is R-sub-Gaussian[1], $\forall d \in [D]$.

As discussed in Section 3, policy adapting to user preference is crucial. To enable effective learning and decision-making in this setting, we propose a unified framework that centers around two key components:

- *Preference Estimation*: Inferring the user preference vector from the observed bandits feedback.
- *Preference-Aware Optimization*: Selecting arms based on preference estimate to align decisions with user intent.

These two components serve as the algorithmic backbone of our approach and remain consistent across both hidden and revealed preference settings. However, in the hidden preference case, each component faces unique technical challenges.

4.1 Unique Challenges

Random Mapping from $\boldsymbol{r}_t$ to g_t. In the hidden preference case, for each user $n \in [N]$, the observed overall rewards are generated through a *random mapping* of rewards. Specifically, $g_{a_t^n,t} = (\overline{\boldsymbol{c}}^{\boldsymbol{n}} + \boldsymbol{\zeta}_t^n)^\top \boldsymbol{r}_{a_t^n,t} = \overline{\boldsymbol{c}}^{\boldsymbol{n}\top} \boldsymbol{r}_{a_t^n,t} + \boldsymbol{\zeta}_t^{n\top} \boldsymbol{r}_{a_t^n,t}$, where $\boldsymbol{\zeta}_t^n = \boldsymbol{c}_t^n - \overline{\boldsymbol{c}}^n \in \mathbb{R}^D$ is an independent random noise vector. This formulation implies that the overall residual noise term $\zeta_{g,t}^n = \boldsymbol{\zeta}_t^{n\top} \boldsymbol{r}_{a_t^n,t}$ is no longer independent of the input. Consequently, standard regression models become infeasible for preference estimation, as they rely on the assumption that the residual noise in the output is independent of the input.

Additionally, the magnitude of overall residual noise is a monotonically non-decreasing function w.r.t each reward objective, i.e., $\|\zeta_{g,t}(\boldsymbol{r}_i)\| \leq \|\zeta_{g,t}(\boldsymbol{r}_j)\|$ iff $\boldsymbol{r}_i \preceq \boldsymbol{r}_j$. This property implies that arms with smaller reward vectors, which are often suboptimal in terms of utility, may paradoxically offer more reliable information for preference estimation due to their lower sensitivity to noise. This issue would also have important implications for the optimization strategy, as we will discuss in the next challenge. Thus, a tailored latent preference estimator is essential to mitigate the expanding error w.r.t the reward and ensure effective preference learning.

[1]By Hoeffding's lemma, for any $X \in [a, b]$ almost surely, X is a R-sub-Gaussian random variable with R at most $(b-a)/2$.

Algorithm 1: PRUCB with Hidden Preference (PRUCB-HP)

Parameters: α, λ, β_t, ω.
Initialization: $\hat{\boldsymbol{r}}_{i,1} \leftarrow [0]^D$, $N_{i,1} \leftarrow 0, \forall i \in [K]$;
For each user $n \in [N]$: $\hat{\boldsymbol{c}}_1^n \leftarrow [1/D]^D$, $\boldsymbol{V}_0^n \leftarrow \lambda \boldsymbol{I}$.
for $t = 1, \cdots, T$ **do**
  for user $n \in [N]$ **do**
    Compute reward bonus term:
    $B_{i,t}^{n,r} = \|\hat{\boldsymbol{c}}_t^n\|_1 \sqrt{\frac{\log t/\alpha}{\max\{N_{i,t},1\}}}, \forall i \in \mathcal{A}_t^n$.
    Compute pseudo information gain term:
    $B_{i,t}^{n,c} = \beta_t \left\| \hat{\boldsymbol{r}}_{i,t} + \sqrt{\frac{\log t/\alpha}{\max\{N_{i,t},1\}}} \boldsymbol{e} \right\|_{\boldsymbol{V}_{t-1}^{n\,-1}}, \forall i \in \mathcal{A}_t^n$.
    Pull arm $a_t \leftarrow \arg\max_{i \in \mathcal{A}_t^n} (\hat{\boldsymbol{c}}_t^n)^\top \hat{\boldsymbol{r}}_{i,t} + B_{i,t}^{n,r} + B_{i,t}^{n,c}$.
    Observe reward $\boldsymbol{r}_{a_t^n,t}$ and overall-reward $g_{a_t^n,t}$.
    Updating:
    $w_t^n = \frac{\omega}{\|\boldsymbol{r}_{a_t^n,t}\|_2^2}$, $\boldsymbol{V}_t^n = \boldsymbol{V}_{t-1}^n + w_t^n \boldsymbol{r}_{a_t^n,t} \boldsymbol{r}_{a_t^n,t}^\top$.
    $\hat{\boldsymbol{c}}_{t+1}^n = (\boldsymbol{V}_t^n)^{-1} \sum_{\ell=1}^{\top} w_\ell^n g_{a_\ell^n,\ell} \boldsymbol{r}_{a_\ell^n,\ell}$.
  end for
  Updating:
  $N_{i,t+1} = N_{i,t} + \sum_{n \in [N]} \mathbb{1}_{\{a_t^n = i\}}$.
  $\hat{\boldsymbol{r}}_{i,t+1} = \frac{\hat{\boldsymbol{r}}_{i,t} N_{i,t} + \sum_{n=1}^N \boldsymbol{r}_{a_t^n,t} \cdot \mathbb{1}_{\{a_t^n = i\}}}{N_{i,t+1}}, \forall i \in [K]$.
end for

Local Exploration *vs* Global Exploration. Unlike traditional bandit algorithms that focus on a single goal (e.g., identifying the arm with the highest reward), the uncertainty in both preference and reward, combined with the need to infer latent preference, introduces a novel trade-off challenge: balancing *global exploration* for better preference estimation and *local exploration* of arm rewards:

- *Global exploration for preferences:* Selecting arms that reduce uncertainty in poorly explored direction of the feature space, refining the model for preference learning.
- *Local exploration for rewards:* Selecting arms to reduce uncertainty for specific individual arm reward estimate, while balancing exploiting empirically high reward arms.

Note that these two learning objectives may conflict, as arms with high rewards might lack sufficient information for latent preference learning and may even worsen degrade estimation performance (as we discussed in the first challenge). This can also be verified by Fig. 3, where 80 samples of $[\boldsymbol{r}_t, g_t]$ are collected by repeatedly pulling an arm, and preference $\hat{\boldsymbol{c}}$ is estimated using linear regression. Here, $\boldsymbol{c}_t$ at each step follows a Gaussian distribution with a mean of $[0.5, 0.5]$. The results demonstrate that samples from suboptimal Arm-1 (Fig. 3a) significantly outperform those from Pareto-optimal Arm-2 (Fig. 3b) in preference estimation. This necessitates an exploration policy that effectively addresses both global and local learning objectives.

4.2 Our Algorithm

To this end, we propose a novel PRUCB-HP method (Algorithm 1) involving two key designs for both preference estimation and preference-aware optimization as follows.

Key design I: WLS-Preference Estimator. For each user $n \in [N]$, as we have seen before, the randomness of preference $\boldsymbol{c}_t^n$ leads to the overall residual noise $\zeta_{g,t}^n$ be a function w.r.t, input reward $\boldsymbol{r}_t^n$. Moreover, larger input rewards $\boldsymbol{r}_t^n$ result in greater corruption from the residual noise. To resolve this, we employ a weighted least-squares (WLS) estimator for preference learning. Specifically, our algorithm assigns a weight w_t^n to each observed sample and estimates the unknown preference using weighted ridge regression:

$$\hat{\boldsymbol{c}}_t^n \leftarrow \arg\min_{\boldsymbol{c}\in\mathbb{R}^D} \lambda\|\boldsymbol{c}\|_2^2 + \sum_{\ell=1}^{t-1} w_\ell^n(\boldsymbol{c}^\top \boldsymbol{r}_{a_\ell^n,\ell} - g_{a_\ell^n,\ell})^2,$$

where λ is the regularization parameter. Above optimization problem has a closed-form solution as:

$$\hat{\boldsymbol{c}}_t^n = (\boldsymbol{V}_{t-1}{}^n)^{-1} \sum_{\ell=1}^{t-1} w_\ell^n g_{a_\ell^n,\ell} \boldsymbol{r}_{a_\ell^n,\ell}, \tag{3}$$

where the Gram matrix $\boldsymbol{V}_{t-1}^n = \lambda \boldsymbol{I} + \sum_{\ell=1}^{t-1} w_\ell^n \boldsymbol{r}_{a_\ell^n,\ell} \boldsymbol{r}_{a_\ell^n,\ell}^\top$.

Inspired by Zhou, Gu, and Szepesvari (2021) using the inverse of the noise variance as weight for tight variance-dependent regret guarantee, we define the weight as the inverse of squared ℓ_2-norm of the reward: $w_t^n = \omega/\|\boldsymbol{r}_{a_t^n,t}\|_2^2$, where $\omega > 0$ is a threshold parameter guaranteeing $w_t^n \geq 1$. Intuitively, it ensures samples with high rewards will be assigned smaller weights to reduce the influence of potentially large residual noises, while samples with low rewards receive larger weights to ensure their contribution to the estimation.

To see how our choice of weight can tackle the *random mapping* issue, we first define $\boldsymbol{r}'_{a_t^n,t} = \sqrt{w_t^n}\boldsymbol{r}_{a_t^n,t}$, $g'_{a_t^n,t} = \sqrt{w_t^n} g_{a_t^n,t}$, then the original formula Eq. 1 can be rewrite as

$$\sqrt{w_t^n} \cdot g_{a_t^n,t} = \sqrt{w_t^n} \cdot \boldsymbol{c}_t^{n\top} \boldsymbol{r}_{a_t^n,t} = \sqrt{w_t^n} \cdot (\overline{\boldsymbol{c}}^n + \boldsymbol{\zeta_t}^n)^\top \boldsymbol{r}_{a_t^n,t}$$
$$\implies g'_{a_t^n,t} = \overline{\boldsymbol{c}}^{n\top} \boldsymbol{r}'_{a_t^n,t} + \sqrt{w_t^n} \cdot \boldsymbol{\zeta}_t^{n\top} \boldsymbol{r}_{a_t^n,t}. \tag{4}$$

For term $\sqrt{w_t^n}\boldsymbol{\zeta}_t^\top \boldsymbol{r}_{a_t^n,t}$, we have the following lemma:

Lemma 1. *For any $n \in [N]$, the random variable $\sqrt{w_t^n}\boldsymbol{\zeta}_t^{n\top} \boldsymbol{r}_{a_t^n,t}$ is sub-Gaussian with constant $R' = \sqrt{\omega}R$.*

The proof is available in Appendix D.1. By above lemma, we observe that with the designed weight, the original random mapping regression problem is transferred into a new formula as (4). Specifically, the output $g'_{a_t^n,t}$ is mapped from $\boldsymbol{r}'_{a_t^n,t}$ via a *fixed* vector $\overline{\boldsymbol{c}}^n$ with a normed R'-sub-Gaussian residual noise, where $R' = \sqrt{\omega}R$, independent of the input.

Key design II: Dual-Exploration Policy. As discussed earlier, there is a new global-local exploration dilemma in our setting. On the one hand, the algorithm must focus on local exploration by selecting optimistically profitable arms to discover better ones. Simultaneously, it must globally explore diverse arms to gather information about the relationship between $\boldsymbol{r}_t$ and g_t for modeling the hidden preference.

To resolve this, we design an *optimistic dual-exploration policy* by incorporating a *preference-driven bonus* and *reward-driven bonus* under preference-aware optimization framework for trade-off. The optimistic policy is defined as

$$a_t \leftarrow \arg\max_{i\in\mathcal{A}_t^n} (\hat{\boldsymbol{c}}_t^n)^\top \hat{\boldsymbol{r}}_{i,t} + B_{i,t}^{n,r} + B_{i,t}^{n,c}, \tag{5}$$

where $B_{i,t}^{n,r}$ and $B_{i,t}^{n,c}$ are the dual-exploration bonus terms for user n. We detail the design of these bonus terms below and will later theoretically demonstrate in Section 4.3 how they establish a tight UCB for the expected overall reward, ensuring the effectiveness of the optimistic policy in (5).

Reward Bonus $B_{i,t}^{n,r}$. The reward bonus term explicitly encourages local exploration of arms with potentially high rewards, for the principle of optimism in face of uncertainty. Specifically, the bonus $B_{i,t}^{n,r}$ is formulated as a reward uncertainty-aware regularization term:

$$B_{i,t}^{n,r} = \rho_{i,t}^\alpha \|\hat{\boldsymbol{c}}_t^n\|_1. \tag{6}$$

$\rho_{i,t}^\alpha = \sqrt{\log(t/\alpha)/\max\{1, N_{i,t}\}}$ represents the standard Hoeffding bonus that quantifies the uncertainty in the reward estimates for each arm, ensuring that arms with higher uncertainty or lower exploration counts will be prioritized.

Preference Bonus $B_{i,t}^{n,c}$. The preference bonus term aims to encourage the exploration of arms that reduce uncertainty in preference estimation. In previous bandit studies (Abbasi-Yadkori, Pál, and Szepesvári 2011; Zhao et al. 2020; He et al. 2022) involving linear coefficient (θ^*) learning, it has been shown that $\beta\|\boldsymbol{x}_i\|_{\boldsymbol{V}^{-1}}$ provides a tight confidence bonus for the payoff of arm i, where β is the confidence set radius for coefficient estimation, and $\boldsymbol{x}_i$ is the observable arm feature. In information theory, $\|\boldsymbol{x}_i\|_{\boldsymbol{V}^{-1}}$ also reflects entropy reduction in the model posterior, and is used to measure the estimator uncertainty improvement contributed by the chosen action $\boldsymbol{x_i}$ (Li et al. 2010).

However, such design is not feasible in our setting, as the exact reward $\boldsymbol{r}_{a_t^n,t}$ is revealed only after pulling the arm a_t^n, making the actual information gain $\|\boldsymbol{r}_{a_t^n,t}\|_{(\boldsymbol{V}_{t-1}^n)^{-1}}$ from arm a_t^n unpredictable beforehand. To resolve this problem, we introduce a *pseudo information gain* term, defined as $\|\hat{\boldsymbol{r}}_{i,t} + \rho_{i,t}^\alpha \boldsymbol{e}\|_{(\boldsymbol{V}_{t-1}^n)^{-1}}$, where $\rho_{i,t}^\alpha$ is the standard Hoeffding bonus. And then the preference bonus is set as

$$B_{i,t}^{n,r} = \beta_t \cdot \|\hat{\boldsymbol{r}}_{i,t} + \rho_{i,t}^\alpha \boldsymbol{e}\|_{(\boldsymbol{V}_{t-1}^n)^{-1}}, \tag{7}$$

where β_t is the confidence radius of the preference estimate we will give in Lemma 3. Intuitively, this pseudo information gain term captures the potential improvement of preference estimator could achieve by *optimistically* selecting arm i based on its reward estimation. In this way, it explicitly encourages global exploration of arms that reduce uncertainty in preference estimation while performing local exploration.

4.3 Theoretical Results

In this section, we provide theoretical guarantees for the PRUCB-HP algorithm. We first characterize the estimation error of $\hat{\boldsymbol{c}}_t$ w.r.t $\overline{\boldsymbol{c}}$ by WLS preference estimator below.

Lemma 2. *Under Assumption 5, for any user $n \in [N]$, $0 < \alpha < 1$, $\omega > 0$, with a probability at least $1-\vartheta$, the preference estimator $\hat{c}_t^n$ in Algorithm 1 verifies for all $t \in [1, T]$:*

$$\|\hat{\boldsymbol{c}}_t^n - \overline{\boldsymbol{c}}^n\|_{\boldsymbol{V}_{t-1}^n} \leq R\sqrt{\omega D \log\left((1+\omega t/\lambda)/\vartheta\right)} + \sqrt{\lambda}.$$

Please see Appendix D.2 for the proof. This estimate confidence bound essentially implies the effectiveness of WLS preference estimator with the designed weight to handle the random-mapping issue in our problem. With this in hand, we can then derive an upper confidence bound for the expected overall reward for each user.

Lemma 3. *Set* $\beta_t = \sqrt{\omega D \log\big((1+\omega t/\lambda)/\vartheta\big)} + \sqrt{\lambda}$, *then for any* $n \in [N]$, $i \in [K]$ *and* $t > 0$, *with probability at least equal to* $1 - \vartheta - D\alpha^2/t^2$, *we have*

$$\overline{\boldsymbol{c}}^{n\top}\boldsymbol{\mu}_i \le \hat{\boldsymbol{c}}_t^{n\top}\hat{\boldsymbol{r}}_{i,t} + B_{i,t}^{n,r} + B_{i,t}^{n,c}, \tag{8}$$

with $B_{i,t}^{n,r} = \rho_{i,t}^{\alpha}\|\hat{\boldsymbol{c}}_t^n\|_1$ *and* $B_{i,t}^{n,c} = \beta_t \left\|\hat{\boldsymbol{r}}_{i,t} + \rho_{i,t}^{\alpha}\boldsymbol{e}\right\|_{(\boldsymbol{V}_{t-1}^n)^{-1}}$ *as the bonus terms for dual-exploration.*

Please see Appendix D.3 for the proof. Lemma 3 essentially suggests an upper confidence bound of the expected reward for each user, as adopted in our dual-exploration policy (Eq. 5). Notably, two bonus terms strike a balance between local and global explorations while guaranteeing optimization under the principle of optimism in face of uncertainty.

Theorem 1. *For PAMO-MAB with hidden preference, for any* $\lambda > 0$, *by setting* $\alpha = \sqrt{\frac{12\vartheta}{KD(D+3)\pi^2}}$, $\omega \ge D$, $\beta_t = \sqrt{\omega D \log\big((1+\omega T/\lambda)/\alpha\big)} + \sqrt{\lambda}$, *with probability greater than* $1 - 2\vartheta$, *Algorithm 1 has*

$$\begin{aligned} R(T) =& O\Big(DRN\sqrt{\omega T \log^2\big((1+\omega T/\lambda)/\vartheta\big)} \\ &+ \frac{DRN}{\sqrt{\lambda}}\sqrt{\omega DKT\log^2\big((1+\omega T/\lambda)/\vartheta\big)} \\ &+ N\sqrt{KT\log(T/\vartheta)} + MN\Big), \end{aligned}$$

with $M = \Big\lfloor \min\big\{t' \mid t\sigma_{r\downarrow}^2 + \lambda \ge 2D\omega\sqrt{Kt\log\tfrac{t}{\alpha}}, \forall t \ge t'\big\}\Big\rfloor$.[2] *Here the first term represents regret from preference estimation error, the third from reward estimation error, and the second from the combined error of both.*

Please see Appendix D.5 for the proof. The key difficulty of the proof is to upper-bound the accumulative preference bonus $\sum_t B_{i,t}^c$. Specifically, we need to quantify the weighted ℓ_2-norm of the empirical estimation $\hat{\boldsymbol{r}}_{a_t,t}$ with weighting matrix $\boldsymbol{V}_t$ constructed by the true reward $\boldsymbol{r}_{a_t,t}$ instead. This inconsistency renders the classical induction method (Abbasi-Yadkori, Pál, and Szepesvári 2011) for deriving $\log(\frac{\det \boldsymbol{V}_T}{\det \boldsymbol{V}_0})$ infeasible for upper-bounding $\sum_t \|\hat{\boldsymbol{r}}_{a_t,t}\|^2_{\boldsymbol{V}_{t-1}^{-1}}$. To resolve this, we first transfer $\|\hat{\boldsymbol{r}}_{a_t,t}\|_{\boldsymbol{V}_{t-1}^{-1}}$ to $\|\boldsymbol{\mu}_{a_t}\|_{\boldsymbol{V}_{t-1}^{-1}}$. Then we show that for sufficiently large t, $a\|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_{t-1}]^{-1}}$ serves as an upper bound for $\|\boldsymbol{\mu}_{a_t}\|_{\boldsymbol{V}_{t-1}^{-1}}$ with constant $a > 1$ (see Lemma 8). This allows us to use $a\|\boldsymbol{\mu}_{a_t}\|_{\mathbb{E}[\boldsymbol{V}_{t-1}]^{-1}}$ as an upper-bound, where a new recursion relationship between $\mathbb{E}[\boldsymbol{V}_{t-1}]$ and $\boldsymbol{\mu}_{a_t}$ can be guaranteed, enabling us to bound $\log(\frac{\det \mathbb{E}[\boldsymbol{V}_T]}{\det \mathbb{E}[\boldsymbol{V}_0]})$ via induction, which can further be bounded by slightly modifying existing techniques in linear bandits.

Remark 1. Theorem 1 shows that, even without explicit preference feedback, PRUCB-HP achieves sub-linear regret through carefully designed mechanisms for preference adaptation. In particular, for $t \ge M$, where M is a constant independent of T, the regret asymptotically scales as $\tilde{O}(D\sqrt{T})$.

[2] Since $\sigma_{\boldsymbol{r}\downarrow}^2 \in \mathbb{R}^+$, $\lim_{t\to\infty} 2D\omega\sqrt{Kt\log\frac{t}{\alpha}}/(\sigma_{r\downarrow}^2 t) = \lim_{t\to\infty} C_1\sqrt{(\log t - C_2)/t} = 0$, as $\sqrt{\log t}$ grows much slowly compared to $\sqrt{t}$. Hence M exists for sufficiently large t'.

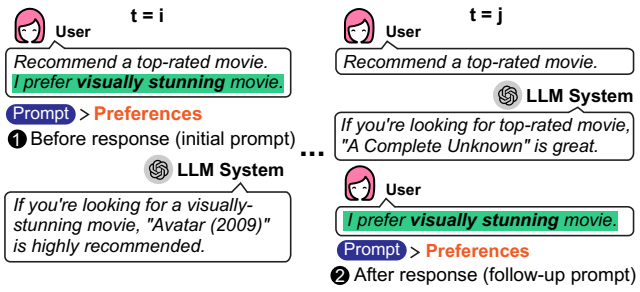


Figure 4: An example where user explicitly provides her preference to LLM system before (❶) or after (❷) the response of movie recommendation (decision making).

To the best of our knowledge, this is the first result characterizing the performance of PAMO-MAB with hidden preference.

5 A Special Case with Provided Preference

We next consider another practically important case where the user's preferences are explicitly provided to the agent either before or after decision making. This setup is prevalent in numerous real-world applications. Many online systems now allow users to express their preferences through interactive techniques, such as conversations and prompt design, either before or after taking action. For example, as shown in Fig. 4, a user can explicitly share her movie preferences with an LLM-based chat system either prior to receiving a recommendation (❶ in the initial input) or after receiving one (❷ in the follow-up conversation).

We observe that when the preference vector $\boldsymbol{c}t$ is known before decision-making, the problem reduces to a standard single-objective MAB. Instead, we focus on the more challenging case where the preference is unknown during action selection but revealed afterward. The known-before-action case can be treated as a special instance of this setting and is deferred to Appendix E.2. The goal remains to maximize the cumulative overall-reward over all users. Formally, at each round t, the learner selects an arm a_t, then observes both the reward $\boldsymbol{r}\_a_t, t$ and user's preference $\boldsymbol{c}_t$, with both adhering to the original Assumptions 1 and 2. This setting avoids the challenges of preference inference and stochastic reward mapping, and can be viewed as a special case of the hidden preference setting with zero noise in preference feedback.

To address this case, we propose a simplified algorithm: PRUCB-UP (Algorithm 2) which builds on the same Preference-Aware framework used in the hidden case, with two direct adaptations:

Preference estimation. Due to the explicitly provided preference feedback $\boldsymbol{c}_t^n$, we can directly leverage the empirical average of historical feedback as the preference estimate for each user n. For $t \ge 1$, preference estimate is updated as

$$\hat{\boldsymbol{c}}_{t+1}^n = \big((t-1)\hat{\boldsymbol{c}}_t^n + \boldsymbol{c}_t^n\big)/t. \tag{9}$$

Similarly, the reward estimate $\hat{\boldsymbol{r}}_{i,t}$ is defined as empirical estimation. For $t \in [1, T]$, it is updated as:

$$N_{i,t+1} = N_{i,t} + \textstyle\sum_{n\in[N]} \mathbb{1}_{\{a_t^n = i\}}, \tag{10}$$

Algorithm 2: PRUCB Unknown Preference (PRUCB-UP)

Initialize: α, $N_{i,1} \leftarrow 0$, $\hat{\boldsymbol{r}}_{i,1} \leftarrow [0]^D, \forall i \in [K]$;
$\hat{\boldsymbol{c}}_1^n \leftarrow [0]^D, \forall n \in [N]$.
for $t = 1, \cdots, T$ **do**
  for user $n \in [N]$ **do**
    Draw arm a_t^n by (11),
    Observe reward $\boldsymbol{r}_{a_t^n,t}$ and user preference $\boldsymbol{c}_t^n$.
    Update preference estimate $\hat{\boldsymbol{c}}_{t+1}^n$ by (9).
  end for
  Update $N_{i,t+1}, \hat{\boldsymbol{r}}_{i,t+1}, \forall i \in [K]$ by (10).
end for

$$\hat{\boldsymbol{r}}_{i,t+1} = (\hat{\boldsymbol{r}}_{i,t} N_{i,t} + \textstyle\sum_{n\in[N]} \boldsymbol{r}_{a_t^n,t} \cdot \mathbb{1}_{\{a_t^n = i\}})/N_{i,t+1},$$

with $N_{i,1} \leftarrow 0, \hat{\boldsymbol{r}}_{i,1} \leftarrow [0]^D, \forall i \in [K]$.

Preference-aware optimization. We adopt the same UCB framework as in hidden preference case for preference-aware optimization, but with key simplifications. Since the preference $\boldsymbol{c}_t$ is observed after action selection and is independent of the chosen arm, its estimation does not involve sequential decision-making—hence no confidence term is needed for $\hat{\boldsymbol{c}}_t$. In contrast, reward estimates $\hat{\boldsymbol{r}}_t$ remain action-dependent, requiring reward bonuses to ensure sufficient exploration.

To this end, the arm selection policy is designed as:

$$a_t^n = \arg\max_{i \in \mathcal{A}_t^n} (\hat{\boldsymbol{c}}_t^n)^\top (\hat{\boldsymbol{r}}_{i,t} + \rho_{i,t}^\alpha \boldsymbol{e}). \tag{11}$$

where $\rho_{i,t}^\alpha = \sqrt{\log(t/\alpha)/\max\{1, N_{i,t}\}}$ is standard Hoeffding bonus. We characterize the regret of PRUCB-UP below.

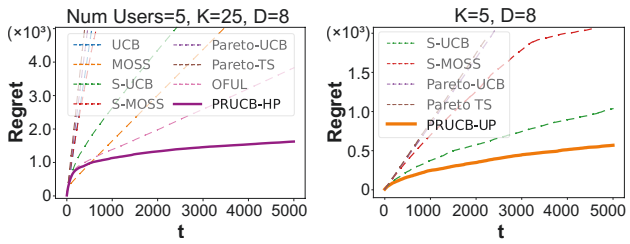
Theorem 2. *Let* $\mathcal{T}_i^n = \{t \in [T] \mid i \in \mathcal{A}_t^n, i \neq a_t^{n*}\}$, $\eta_i^{n\uparrow} = \max_{t\in\mathcal{T}_i^n} \overline{\boldsymbol{c}}^{n\top}(\boldsymbol{\mu}_{a_t^{n*}} - \boldsymbol{\mu}_i)$, $\eta_i^{n\downarrow} = \min_{t\in\mathcal{T}_i^n} \overline{\boldsymbol{c}}^{n\top}(\boldsymbol{\mu}_{a_t^{n*}} - \boldsymbol{\mu}_i)$, $\|\Delta_i^{n\uparrow}\|_2 = \max_{\{t,j\}\in\mathcal{T}_i^n \times \mathcal{A}_t^n} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2$, *Algorithm 2 has the regret R(T) of*

$$O\Bigg(\sum_{n\in[N]}\sum_{i\in[K]}\Big(\underbrace{\frac{\delta^2 \log T}{\eta_i^{n\downarrow}} + D\pi^2\alpha^2\eta_i^{n\uparrow}}_{R_T^r} + \underbrace{\frac{D^2\|\Delta_i^{n\uparrow}\|_2^2\delta^2}{\eta_i^{n\downarrow}}}_{R_T^c}\Big)\Bigg),$$

where R_T^r *and* R_T^c *refer to the regrets caused by reward estimate error and preference estimate error respectively.*

Remark 2. Theorem 2 shows that with preference feedback, PRUCB-UP achieves a regret of $O(KN\delta \log T + KN\delta D^2)$, demonstrating near-optimal performance. Notably, the regret caused by additional preference estimation error is bounded by a constant related to objective dimension D and preference ℓ_1-norm bound δ. This implies that the impact of preference estimation error on the regret is small. Additional, we show the preference known case can be solved by PRUCB-UP as a special case, achieving regret of $O(K\delta \log T)$, please see Appendix E.2 for details.

To prove Theorem 2, the main difficulty lies in decoupling and capturing the effect of the joint error from both reward estimation and preference estimation on the final regret. To address this, we introduce a tunable parameter ϵ to quantify the error of preference estimate $\hat{\boldsymbol{c}}_t$, and decompose suboptimal actions into two disjoint sets: (1) suboptimal pulls under sufficiently precise preference estimate; (2) suboptimal pulls under imprecise preference estimate, which we show to be more analytically tractable. Please see Appendix E.1 for the full proof.



(a) Hidden preference case (b) Unknown preference case

Figure 5: Regret comparison of our proposed PRUCB with other benchmarks under different preference environments, where our methods outperforms other methods significantly.

6 Numerical Analysis

We evaluate the performance of PRUCB-UP and PRUCB-HP in unknown and hidden preference environments, respectively. The PAMO-MAB instance includes N users, K arms and D objectives, with preferences and rewards following Gaussian distributions with randomly initialized means. (Detailed settings refer to Appendix B.2). We introduce a user-switching protocol to simulate practical scenarios. The environment features multiple users, each exposed to a block of arms (5 in our setup) per round. Only arms within the current block can be selected for that user. In the next round, the arm block rotates to a another user. The objective is to maximize cumulative overall ratings across all users. A more detailed illustration is provided in Appendix B.2.

We compare our results with other baselines including S-UCB, Pareto-UCB (Drugan and Nowe 2013), S-MOSS, Pareto-TS (Yahyaa and Manderick 2015)), UCB (Auer, Cesa-Bianchi, and Fischer 2002), MOSS (Audibert and Bubeck 2009) and OFUL (Abbasi-Yadkori, Pál, and Szepesvári 2011). The regret is averaged across 10 trials with round $T = 5000$. Figure 5 shows that our proposed algorithms significantly outperform other competitors under both environments. It is worth noting that for all the preference-free competitors exhibit linear regret, aligning with Proposition 1, demonstrating that approaches agnostic to user preferences cannot align their outputs with user preferences, even if they achieve Pareto optimality. For more comprehensive experimental analyses, please refer to Appendix B.

7 Conclusion

In this paper, we make the first effort to theoretically explore the explicit user preferences-aware MO-MAB. Motivated by real-world applications, we provide a comprehensive analysis of this problem under unknown preference and hidden preference environments, with tailored algorithms achieving provably near-optimal regrets.

Acknowledgments

This work has been supported in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0225, by the U.S. National Science Foundation under the grants: NSF AI Institute (AI-EDGE) 2112471, CNS-2312836, CNS-2225561, and CNS-2239677, and Office of Naval Research under grant N00014-24-1-2729. 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

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24.

Audibert, J.-Y.; and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits. In *Conference on Learning Theory*, 217–226.

Audibert, J.-Y.; Munos, R.; and Szepesvári, C. 2007. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory*, 150–165. Springer.

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

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

Busa-Fekete, R.; Szörényi, B.; Weng, P.; and Mannor, S. 2017. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, 625–634. PMLR.

Cheng, J.; Xue, B.; Yi, J.; and Zhang, Q. 2024. Hierarchize Pareto Dominance in Multi-Objective Stochastic Linear Bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11489–11497.

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

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

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

Fulton, W. 2000. Eigenvalues, invariant factors, highest weights, and Schubert calculus. *Bulletin of the American Mathematical Society*, 37(3): 209–249.

He, J.; Zhou, D.; Zhang, T.; and Gu, Q. 2022. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35: 34614–34625.

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

Jun, K.-S.; Li, L.; Ma, Y.; and Zhu, J. 2018. Adversarial attacks on stochastic bandits. *Advances in neural information processing systems*, 31.

Lamprier, S.; Gisselbrecht, T.; and Gallinari, P. 2018. Profile-based bandit with unknown profiles. *Journal of Machine Learning Research*, 19(53): 1–40.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 661–670.

Lu, S.; Wang, G.; Hu, Y.; and Zhang, L. 2019. Multi-objective generalized linear bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3080–3086.

Mehrotra, R.; Xue, N.; and Lalmas, M. 2020. Bandit based optimization of multiple objectives on a music streaming platform. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3224–3233.

Reymond, M.; Bargiacchi, E.; Roijers, D. M.; and Nowé, A. 2024. Interactively Learning the User's Utility for Best-Arm Identification in Multi-Objective Multi-Armed Bandits. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 1611–1620.

Turgay, E.; Oner, D.; and Tekin, C. 2018. Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, 1673–1681. PMLR.

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

Wanigasekara, N.; Liang, Y.; Goh, S. T.; Liu, Y.; Williams, J. J.; and Rosenblum, D. S. 2019. Learning Multi-Objective Rewards and User Utility Function in Contextual Bandits for Personalized Ranking. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 19, 3835–3841.

Xie, R.; Liu, Y.; Zhang, S.; Wang, R.; Xia, F.; and Lin, L. 2021. Personalized approximate pareto-efficient recommendation. In *Proceedings of the Web Conference 2021*, 3839–3849.

Xu, M.; and Klabjan, D. 2023. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, 38499–38517. PMLR.

Yahyaa, S. Q.; Drugan, M. M.; and Manderick, B. 2014. Knowledge Gradient for Multi-objective Multi-armed Bandit Algorithms. In *International Conference on Agents and Artificial Intelligence*, 74–83.

Yahyaa, S. Q.; and Manderick, B. 2015. Thompson Sampling for Multi-Objective Multi-Armed Bandits Problem. In *ESANN*.

Zhao, P.; Zhang, L.; Jiang, Y.; and Zhou, Z.-H. 2020. A simple approach for non-stationary linear bandits. In *International Conference on Artificial Intelligence and Statistics*.

Zhou, D.; Gu, Q.; and Szepesvari, C. 2021. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, 4532–4576. PMLR.