Thompson Sampling for Multi-Objective Linear Contextual Bandit

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the multi-objective linear contextual bandit problem, where multiple possible conflicting objectives must be optimized simultaneously. We propose MOL-TS, the *first* Thompson Sampling algorithm with Pareto regret guarantees for this problem. Unlike standard approaches that compute an empirical Pareto front each round, MOL-TS samples parameters across objectives and efficiently selects an arm from a novel *effective Pareto front*, which accounts for repeated selections over time. Our analysis shows that MOL-TS achieves a worst-case Pareto regret bound of $\widetilde{O}(d^{3/2}\sqrt{T})$, where d is the dimension of the feature vectors, T is the total number of rounds, matching the best known order for randomized linear bandit algorithms for single objective. Empirical results confirm the benefits of our proposed approach, demonstrating improved regret minimization and strong multi-objective performance.

1 Introduction

The multi-objective multi-armed bandit (MOMAB) problem [8, 5, 15, 17, 11, 18, 10, 7] is an extension of the traditional multi-armed bandit (MAB) problem, where multiple objectives must be optimized simultaneously. For example, a decision maker chooses among actions with multiple, often contrasting objectives, making it difficult to define a single optimal action. Such problems can be generalized in the MOMAB problem, where one receives objective specific rewards from multiple objectives. This formulation makes it difficult to define the optimality, as one must balance trade-offs across multiple objectives.

To address the trade-off between different objectives, one common approach is scalarization [14, 21, 20], which transforms a multi-objective problem into a single objective by applying a scalarization function, such as weighted sum, minimax, or other nonlinear methods. However, selecting an appropriate scalarization function poses its own challenges. For example, it can be difficult to determine which scalarization is most effective. Also, an arm that is optimal under one scalarization may have significant sub-optimality under another scalarization. The other common approach uses Pareto optimality across multiple reward vectors [8, 5, 11, 18, 10, 7], where Pareto optimal arms are determined by the Pareto order. Such an approach adopts the set of multiple Pareto optimal arms, whose reward vectors are not comparable to those of any other Pareto optimal arms. Because it directly compares reward vectors objective-wise, this approach is more general than scalarization. We will therefore focus on the Pareto optimality in this paper.

Previous studies of the MOMAB have mainly analyzed the Pareto regret of upper confidence bound (UCB) algorithms [8, 11, 18, 10]. To our best knowledge, the Pareto regret of Thompson Sampling (TS) [4, 2] algorithms has not yet been explored in the existing literature of the MOMAB.

In many of the contextual and non-contextual bandit instances with a single objective, TS (and its variants) is known to be more efficient numerically compared to UCB methods [16, 6, 4, 2]. However, the worst-case analysis of TS is typically known to be more involved than that of UCB methods. Analyzing TS algorithms in the MOMAB framework is even more challenging due to the complexity that the randomized sampling creates in each of the multiple objectives. For instance, technical issues involving handling multiple sampled parameters and the possible existence of multiple Pareto optimal arms lead to a much more complicated analysis. Possibly due to these challenges, no TS methods have yet been proposed for the multi-objective bandit problem (either contextual or non-contextual).

In this paper, we propose a TS algorithm for the multi-objective linear contextual bandit problem and provide an analysis of its worst-case Pareto regret upper bound. Our algorithm samples multiple parameters for each objective and optimistically evaluate the reward vector for each arm, ensuring the theoretical probability of arm to be optimistically evaluated in all objectives.

The definition of Pareto regret directly compares the mean reward vector of each arm. However, this definition does not account for maximizing the total reward vector. For example, repeatedly choosing a single Pareto optimal arm may achieve zero regret, but it can lead to sub-optimal overall rewards if an alternative selection could yield strictly higher rewards in all objectives. Such problem can arise in algorithms that adopt the previous definition of Pareto regret (see Section 3.3).

To this end, we propose a new definition of an optimal arm that adopts the previous definition of the Pareto optimal arm and considers such problem with any number of repeated selections. Based on this, we further define a corresponding regret measure that captures the future expected regret where such problem arises with some number of repeated selections. Our proposed algorithm resolves this issue by selecting the newly defined optimal arms.

Our main contributions are summarized as follows:

- We propose an algorithm for the multi-objective linear contextual bandit problem: *Thompson sampling for Multi-objective Linear Bandit* (`MOL-TS`). To the best of our knowledge, this is the first randomized algorithm for multi-objective contextual bandits with Pareto regret guarantees. Unlike the existing multi-objective algorithms, `MOL-TS` does not explicitly compute an empirical Pareto front each round, but rather randomly selects an arm from that Pareto front, which is much more computationally efficiently.
- We propose the concept of a *effective Pareto optimal arm* (Definition 5), which satisfies the condition of Pareto optimal arm, and also the total rewards for every objective with any number of its repeated selection satisfying Pareto optimality. Any arm that is not effective Pareto optimal has an alternative selection of arms over the same total number of rounds, resulting higher total rewards in all objectives. Our proposed algorithm, `MOL-TS`, operates on this new notion of the effective Pareto front and samples an arm from the estimated effective Pareto front. As a result, `MOL-TS` produces higher cumulative rewards compared to the methods that selected from the plain Pareto front.
- We establish that `MOL-TS` is statistically efficient, achieving the Pareto regret bound of $\widetilde{O}(d^{3/2}\sqrt{T})$, where d is the dimension of the feature vectors, T is the total number of rounds. In order to ensure the provable guarantees of the randomized exploration for multiple objectives, `MOL-TS` adopts the *optimistic sampling strategy* (Section 5.3).
- Numerical experiments demonstrate the effectiveness of our proposed approach, showing improved performance in regret minimization, and objective-wise total reward maximization.

2 Related works

Multi-objective multi armed bandit setting was first explored by Drugan and Nowe [8], who proposed UCB algorithms for MOMAB by applying two representative approaches: using Pareto order and scalarized order. Subsequently, Auer et al. [5] proposed algorithms that identify all Pareto optimal arms with high probability. More recently, the upper and lower bounds of Pareto regret in the MOMAB setting have been studied in both stochastic and adversarial settings by Xu and Klabjan [18]. There are also several studies on multi-objective contextual bandits. For example, Tekin and Turğay [15] studied MOMAB in a contextual setting where a dominant objective exists, but we do not assume any dominance among objectives. Turgay et al. [17] developed the PCZ algorithm, which identifies the Pareto front using the idea of contextual zooming and proved its regret bound. However, the algorithm is complex, and the paper does not provide specific details on its implementation. Lu et al. [11] studied the multi-objective generalized linear bandit (MOGLB) problem and analyzed the upper bound of Pareto regret using the ParetoUCB algorithm. Additionally, Kim et al. [10] explored Pareto front identification in linear bandit settings. The studies mentioned thus far proposed complex algorithms that calculate the empirical Pareto front. In contrast, Zhang [20] introduced a hypervolume scalarization method in stochastic linear bandit settings, which uses random scalarization to explore the entire Pareto front.

While these studies address significant challenges in multi-objective bandits, surprisingly, although the practical effectiveness of randomized methods is widely recognized, research on randomized algorithms in multi-objective bandits has been rare. To the best of our knowledge, only Yahyaa and Manderick [19] proposed a Thompson Sampling (TS) algorithm for MOMAB, but no theoretical analysis of this approach has been conducted. Separately, there has been significant research on randomized scalarization in the multi-objective Bayesian optimization problem [14, 21], including theoretical analyses of TS algorithms. However, Zhang and Golovin [21] and Paria et al. [14] analyzed the "Bayes regret" with known Gaussian prior setting, increasing with the number of objectives.

In the single-objective case, the theoretical analysis of Thompson sampling was first introduced by Agrawal and Goyal [3] for the MAB setting, and then was extended to the stochastic linear bandit setting in Agrawal and Goyal [4], where arms were categorized as either saturated or unsaturated to derive theoretical bounds. From a different perspective, Abeille and Lazaric [2] analyzed Thompson Sampling under the assumption of fixed probabilities for sampling optimistic parameters. Additionally, Chapelle and Li [6] showed that, although the theoretical guarantees of Thompson Sampling are weaker than those of UCB, empirical results have consistently demonstrated that TS algorithms outperform UCB algorithms in practice. However, there has been a clear gap in extending these theoretical guarantees from single objective settings to multi-objective bandits.

We provide the first theoretical analysis of a randomized algorithm in the multi-objective bandit setting. To the best of our knowledge, this is the first work to propose a TS algorithm for multi-objective linear contextual bandits and to analyze it theoretically.

3 Preliminaries

3.1 Notations

Throughout this paper, we use notations that distinguish between different objectives. For any positive integer $N \in \mathbb{N}$, let $[N] := \{1, 2, \ldots, N\}$. We denote L as the number of objectives, and for $\ell \in [L]$, any real number u corresponding to the ℓ_{th}-objective is denoted as $u^{(\ell)}$. The vector $\boldsymbol{u} \in \mathbb{R}^L$ comprises all $u^{(\ell)}$ values and is represented in bold notation, i.e., $\boldsymbol{u} = [u^{(1)}, u^{(2)}, \ldots, u^{(L)}]^\top$. Otherwise, individual features of any vector x are typically denoted as $x(i)$. For clarity, $\|\cdot\|$ denotes the Euclidean norm, and for a positive semi-definite matrix V, the norm $\|\cdot\|_V$ is defined in the inner product space with the matrix V as $\langle x, y\rangle_V = \sqrt{x^\top V y}$. Finally, we define $\mathcal{S}^n \subset \mathbb{R}^n$ as the unit $(n-1)$-simplex.

3.2 Problem settings

We consider a standard stochastic linear contextual bandit problem, extended to the multi-objective setting. Let $\mathcal{A}$ be a finite set of arms. Each arm $a \in \mathcal{A}$ corresponds to a d-dimensional context vector $x_{t,a} \in \mathbb{R}^d$. For each objective $\ell \in [L]$, there exists a fixed parameter $\theta_*^{(\ell)} \in \mathbb{R}^d$, but unknown to agent. In total, there are L parameters $\theta_*^{(1)}, \theta_*^{(2)}, \ldots, \theta_*^{(L)}$. At each round $t \in [T]$, the agent selects an arm $a_t \in \mathcal{A}$ and receives a L-dimensional reward vector $\boldsymbol{r}_{t,a_t} = [r_{t,a_t}^{(1)}, r_{t,a_t}^{(2)}, \ldots, r_{t,a_t}^{(L)}]^\top \in \mathbb{R}^L$, where the reward for each objective ℓ is given by $r_{t,a_t}^{(\ell)} = x_{t,a_t}^\top \theta_*^{(\ell)} + \xi_t^{(\ell)}$, and $\xi_t^{(\ell)}$ is a zero-mean random noise. The mean reward for objective ℓ is defined as $\mu_{t,a_t}^{(\ell)} := \mathbb{E}[r_{t,a_t}^{(\ell)}]$. And consequently, the mean reward vector of the chosen arm a_t is $\boldsymbol{\mu}_{t,a_t} = [\mu_{t,a_t}^{(1)}, \mu_{t,a_t}^{(2)}, \ldots, \mu_{t,a_t}^{(L)}]^\top \in \mathbb{R}^L$.

3.3 Pareto optimality

Definition 1 (Pareto order) *Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^L$ be two distinct vectors. We say that the vector $\boldsymbol{u}$ is dominated by the vector $\boldsymbol{v}$ (i.e., $\boldsymbol{v}$ dominates $\boldsymbol{u}$), denoted as $\boldsymbol{u} \prec \boldsymbol{v}$, if $u^{(\ell)} \leq v^{(\ell)}$ for all $\ell \in [L]$ and*

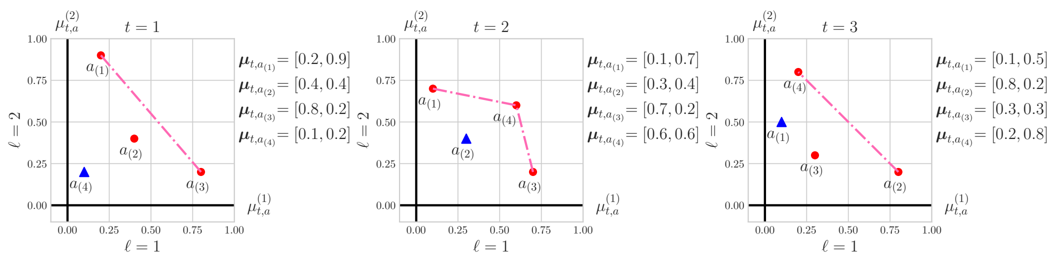

Figure 1: Example of two objectives and four arms, $a_{(1)}, a_{(2)}, a_{(3)}$, and $a_{(4)}$. Each subplot shows the mean reward vector at round t, where the horizontal and vertical axes correspond to the first and second objective, respectively. Red circles represent Pareto optimal arms, blue triangles that are not. The mean reward vectors are listed on the right, and the pink line represents the boundary of effective Pareto front (see Definition 5).

$u^{(\ell)} < v^{(\ell)}$ for some $\ell \in [L]$. Conversely, the vector $\boldsymbol{u}$ is not dominated by the vector $\boldsymbol{v}$, denoted as $\boldsymbol{u} \not\prec \boldsymbol{v}$, if there exists at least one $\ell \in [L]$ satisfying $u^{(\ell)} > v^{(\ell)}$

Definition 2 (Pareto optimal arm) *An arm is Pareto optimal if its mean reward vector is not dominated by that of any other arm. The set of all Pareto optimal arms is called Pareto Front ($\mathcal{P}_t^*$),*

$$\mathcal{P}_t^* := \{a \in \mathcal{A} \mid \boldsymbol{\mu}_{t,a} \not\prec \boldsymbol{\mu}_{t,a'}, \forall a' \in \mathcal{A}\}.$$

In linear setting, where the mean rewards of each arm remain fixed, the Pareto front is denoted as $\mathcal{P}^*$.

Definition 3 (Pareto sub-optimality gap) *The Pareto sub-optimality gap of an arm a at round t is the minimum scalar value $\epsilon \geq 0$ for a to be Pareto optimal, i.e.,*

$$\Delta_{t,a}^{PR} := \inf\{\epsilon \geq 0 \mid \boldsymbol{\mu}_{t,a} + \epsilon \mathbf{1} \not\prec \boldsymbol{\mu}_{t,a'}, \forall a' \in \mathcal{A}\}.$$

The Pareto sub-optimal gap can be defined as $\Delta_{t,a}^{PR} := \max_{a' \in \mathcal{P}^*} \min_{\ell \in [L]} \{\mu_{t,a'}^{(\ell)} - \mu_{t,a}^{(\ell)}\}$. For every Pareto optimal arm $a' \in \mathcal{P}_t^*$, the arm that maximizes $\Delta_{t,a'}^{PR}$ is a' itself, which implies $\Delta_{t,a'}^{PR} = 0$. Any other arm is dominated by at least one Pareto optimal arm, ensuring that $\Delta_{t,a}^{PR} \geq 0$.

Definition 4 (Pareto regret) *Let $a_1, a_2, \ldots, a_T$ be the sequence of arms chosen by agent. The Pareto regret up to round T is defined as $PR(T) := \sum_{t=1}^{T} \Delta_{t,a_t}^{PR}$.*

In the contextual bandit setting, the Pareto front varies dynamically depending on the given context. Hence, we can not apply the algorithm of identifying Pareto front in Auer et al. [5] and Kim et al. [10] as they remove arm from the arm set, which can be Pareto optimal in our setting.

By using the Pareto order relationship, the definition of Pareto regret provides a general measurement of an agent's performance in a multi-objective setting. Previous studies of Pareto optimality [8, 11, 17, 10] adopt this measurement and design algorithms that randomly select arms from the Pareto front, aiming to minimize the Pareto regret. However, this definition of Pareto regret does not fully account for cumulative rewards. For example, consider a case with two objectives and four arms, as illustrated in Figure 1. Suppose two agents follow same policy that randomly selects arm from Pareto front. The first agent sequentially selects $a_{(2)}, a_{(4)}$ and $a_{(3)}$, while the second agent selects $a_{(1)}$, $a_{(4)}$ and $a_{(2)}$. Both agents selected Pareto optimal arms, resulting zero Pareto regret. But total rewards of the first agent is $\boldsymbol{\mu}_{1,a_{(2)}} + \boldsymbol{\mu}_{2,a_{(4)}} + \boldsymbol{\mu}_{3,a_{(3)}} = [1.3, 1.3]$ and the second agent is $[1.6, 1.7]$. This example highlights the limitation of Pareto regret, as it does not distinguish between policies that yield different cumulative rewards despite selecting only Pareto optimal arms. Thereby, we propose the concept of a *effective Pareto optimal* arm, whose mean reward vector is Pareto optimal and contributes to maximizing cumulative rewards across all objectives.

3.4 Effective Pareto optimality

Definition 5 (Effective Pareto optimal arm) *An arm is effective Pareto optimal (denoted a_*) if its mean reward vector is either equal to or not dominated by any convex combination of the mean* *reward vectors of the other arms. Formally, for any $\beta \in \mathcal{S}^{|\mathcal{A}|-1}$,*

$$\boldsymbol{\mu}_{t,a_*} = \sum_{a \in \mathcal{A}\setminus\{a_*\}} \beta_a \boldsymbol{\mu}_{t,a} \qquad \textit{or} \qquad \boldsymbol{\mu}_{t,a_*} \not\prec \sum_{a \in \mathcal{A}\setminus\{a_*\}} \beta_a \boldsymbol{\mu}_{t,a},$$

where $\beta = (\beta_a)_{a \in \mathcal{A}\setminus\{a_*\}}$. The set of all effective Pareto optimal arms at round t is called the effective Pareto front, denoted as $\mathcal{C}_t^*$. In the linear bandit setting, where the mean reward vectors of all arms remain fixed, the effective Pareto front is denoted as $\mathcal{C}^*$

In this paper, we refer to an arm that is not effective Pareto optimal as sub-optimal. If an arm $a' \in \mathcal{A}$ is sub-optimal, then there exists a convex combination β, such that $\boldsymbol{\mu}_{t,a'} \prec \sum_{a \in \mathcal{A}\setminus\{a'\}} \beta_a \boldsymbol{\mu}_{t,a}$.

Every effective Pareto optimal arm is also a Pareto optimal arm. This can be easily verified by restricting β to be a one-hot vector, which corresponds to comparing mean reward vectors between two individual arms. However, the converse does not hold. As can be seen from Figure 1, at the first round, arm $a_{(2)}$ is Pareto optimal, but not effective Pareto optimal, because its mean reward vector is dominated by a convex combination of two arms $a_{(1)}$ and $a_{(3)}$. Hence, for any $t \in [T]$, we have $\mathcal{C}_t^* \subset \mathcal{P}_t^*$. This shows that our definition of effective Pareto optimal arm is strictly defined than the standard definition of Pareto optimal arm.

As shown in the example in the previous subsection, two agents following the same policy randomly selected arms from the Pareto front. The second agent consistently selected arms from the effective Pareto front, while the first agent selected arms from the Pareto front but not the effective Pareto front in the first and third rounds. This difference led to the first agent achieving lower cumulative rewards than the second agent for all objectives. Importantly, this disparity becomes increasingly severe as the total number of rounds T grows, leading to significantly worse long-term performance for policies that fail to prioritize the effective Pareto front.

The intuition behind an effective Pareto optimal is that repeatedly selecting such arms leads to Pareto optimal cumulative rewards. In other words, rather than selecting a sub-optimal arm over multiple rounds, it is preferable to select effective Pareto optimal arms for the same total number of rounds, which is expected to yield strictly higher cumulative reward in some objectives without worsening the others. In summary, for large enough total number of rounds T, selecting arms from effective Pareto front $\mathcal{C}_t^*$ achieves higher total rewards than selecting arms from Pareto optimal front $\mathcal{P}_t^*$. Based on this, we propose a theorem that establishes a relationship between the newly defined effective Pareto optimality and the linear scalarization method.

Theorem 1 *For any $a_* \in \mathcal{C}^*$, there exist $\boldsymbol{w} \in \mathcal{S}^L$ satisfying $a_* = \arg\max_{a \in \mathcal{A}} \boldsymbol{w}^\top \boldsymbol{\mu}_a$. Conversely, for any $\boldsymbol{w} \in \mathcal{S}^L$, if $a_* = \arg\max_{a \in \mathcal{A}} \boldsymbol{w}^\top \boldsymbol{\mu}_a$ is unique arm, then $a_* \in \mathcal{C}^*$.*

The theorem is proved in Appendix A where we refer to the proof from Mangasarian [12]. The theorem shows a one-to-one correspondence: every effective Pareto optimal arm is optimal for some non-negative weight vector, and every non-negative weight vector guarantees to have an effective Pareto optimal arm.

Definition 6 (Effective Pareto sub-optimality gap) *Let $\beta = (\beta_a)_{a \in \mathcal{A}}$ be a vector in $\mathcal{S}^{|\mathcal{A}|}$ and define $\boldsymbol{\mu}_{t,\beta} = \sum_{a \in \mathcal{A}} \beta_a \boldsymbol{\mu}_{t,a}$. The effective Pareto sub-optimality gap for selecting arm a_t at round t is defined as*

$$\Delta_{t,a_t}^{EPR} := \inf\left\{ \epsilon \geq 0 \;\middle|\; \boldsymbol{\mu}_{t,a_t} + \epsilon \mathbf{1} \not\prec \boldsymbol{\mu}_{t,\beta}, \forall \beta \in \mathcal{S}^{|\mathcal{A}|} \right\}.$$

The effective Pareto sub-optimal gap measures the minimum value ϵ for a_t not to be dominated by any convex combination of the mean reward vectors of the other arms. In other words, the gap quantifies how close arm a_t is to being effective Pareto optimal. For any effective Pareto optimal arm, this gap is zero. Also, it is easy to verify that the effective Pareto sub-optimality gap is always greater than or equal to the standard Pareto sub-optimality gap, i.e., $\Delta_{t,a_t}^{PR} \leq \Delta_{t,a_t}^{EPR}$, since the Pareto sub-optimality gap corresponds to the special case where β is restricted to be a one-hot vector. As discussed in Section 3.3, the effective Pareto sub-optimality gap can also be expressed as

$$\Delta_{t,a_t}^{EPR} := \max_{\beta \in \mathcal{S}^{|\mathcal{C}_t^*|}} \min_{\ell \in [L]} \left\{ \left(\sum_{a_* \in \mathcal{C}_t^*} \beta_{a_*} \mu_{t,a_*}^{(\ell)} \right) - \mu_{t,a_t}^{(\ell)} \right\}, \tag{1}$$

Algorithm 1 Multi-Objective Linear TS (`MOL-TS`)

Input: $\lambda, \delta, M, T > 0, c > 0$
Initialize: $V_1 = \lambda I_d$, $\hat{\theta}_1^{(\ell)}, Z_1^{(\ell)} = \mathbf{0}$ $(\forall \ell \in [L])$
for $t = 1 \to T$ **do**
  for objective $\ell = 1, 2, 3, \ldots, L$ **do**
    Sample $\tilde{\theta}_{t,1}^{(\ell)}, \tilde{\theta}_{t,2}^{(\ell)}, \ldots, \tilde{\theta}_{t,M}^{(\ell)} \sim \mathcal{N}(\hat{\theta}_t^{(\ell)}, c^2 V_t^{-1})$
    Evaluate every arm $\tilde{\mu}_{t,a}^{(\ell)}$ using Equation (2)
  end for
  Update the empirical effective Pareto front $\tilde{\mathcal{C}}_t$ using Equation (3)
  Sample arm a_t from $\tilde{\mathcal{C}}_t$ uniformly at random, play a_t, observe reward vector $\boldsymbol{r}_{t,a_t}$
  Update $V_{t+1} \leftarrow V_t + x_{t,a_t} x_{t,a_t}^\top$
  for objective $\ell = 1, \ldots, L$ **do**
    Update $Z_{t+1}^{(\ell)} \leftarrow Z_t^{(\ell)} + x_{t,a_t} r_{t,a_t}^{(\ell)}$ and $\hat{\theta}_{t+1}^{(\ell)} \leftarrow V_{t+1}^{-1} Z_{t+1}^{(\ell)}$
  end for
end for

Definition 7 (Effective Pareto regret) *The effective Pareto regret up to round T is defined as $EPR(T) := \sum_{t=1}^{T} \Delta_{t,a_t}^{EPR}$.*

4 Algorithm: `MOL-TS`

We propose a multi-objective linear Thompson Sampling algorithm, `MOL-TS`, a generic randomized algorithm designed with multiple regularized MLE, where one need not sample from an actual Bayesian posterior [2]. Our algorithm adopts an optimistic sampling strategy to avoid the theoretical challenges in worst-case regret analysis [13, 9].

Each round t, the mean reward vector for each arm is estimated based on the history of chosen arms $a_1, a_2, \ldots, a_{t-1}$, and received reward vectors $\boldsymbol{r}_{1,a_1}, \boldsymbol{r}_{2,a_2}, \ldots, \boldsymbol{r}_{t-1,a_{t-1}}$ up to round t. The true parameter for each objective $\theta_*^{(\ell)}$ is estimated by regularized least squares (RLS), denoted $\hat{\theta}_t^{(\ell)}$. Given regularizer $\lambda \in \mathbb{R}^+$, the matrix and the RLS estimator is defined as

$$V_t = \sum_{s=1}^{t-1} x_{s,a_s} x_{s,a_s}^\top + \lambda I_{d \times d}, \qquad \hat{\theta}_t^{(\ell)} = V_t^{-1} \sum_{s=1}^{t-1} x_{s,a_s} r_{s,a_s}^{(\ell)}.$$

For each objective ℓ, the parameters $(\tilde{\theta}_{t,m}^{(\ell)})_{m \in [M]}$ are sampled independently M times from Gaussian posterior distribution $\mathcal{N}(\hat{\theta}_t^{(\ell)}, c^2 V_t^{-1})$, where the tunable parameters c and M are given from the beginning. A total of ML samples are drawn. The reward for each arm and for each objective is then optimistically evaluated using the sample that yields the highest value,

$$\tilde{\mu}_{t,a}^{(\ell)} = \max\{x_{t,a}^\top \tilde{\theta}_{t,1}^{(\ell)}, x_{t,a}^\top \tilde{\theta}_{t,2}^{(\ell)}, \ldots, x_{t,a}^\top \tilde{\theta}_{t,M}^{(\ell)}\}. \tag{2}$$

The reward vector for each arm is then constructed as $\tilde{\boldsymbol{\mu}}_{t,a} = \begin{bmatrix} \tilde{\mu}_{t,a}^{(1)} & \tilde{\mu}_{t,a}^{(2)} & \ldots & \tilde{\mu}_{t,a}^{(L)} \end{bmatrix}^\top$.

The number of samples M controls the probability that the estimated rewards are optimistically evaluated. Increasing M raises the likelihood that the reward estimates are optimistic, which is crucial for ensuring a high theoretical probability of optimism across multiple objectives (see Section 5.3). We approximate the empirical effective Pareto front $\tilde{\mathcal{C}}_t$ using the estimated reward vectors, by

$$\tilde{\mathcal{C}}_t = \left\{ a \in \mathcal{A} \;\middle|\; \tilde{\boldsymbol{\mu}}_{t,a} = \sum_{a' \in \mathcal{A}} \beta_{a'} \tilde{\boldsymbol{\mu}}_{t,a'} \text{ or } \tilde{\boldsymbol{\mu}}_{t,a} \not\prec \sum_{a' \in \mathcal{A}} \beta_{a'} \tilde{\boldsymbol{\mu}}_{t,a'}, \forall \beta_{a'} \in \mathcal{S}^{|\mathcal{A}|} \right\}. \tag{3}$$

Note that this optimistic sampling strategy is different from that proposed in [13, 9]. The setting in [13] considers a dynamic assortment selection problem, and [9] considers a combinatorial selection problem. Unlike multiple arms selection problem in single objective setting, our setting considers receiving multiple rewards from single arm selection problem.

5 Regret analysis

In this section, we analyze the expected effective Pareto regret of our algorithm `MOL-TS` in the worst-case, where the expectation is taken over all sources of randomness present in the problem setup. We begin with the general assumptions widely used in the linear bandit literature. We then outline the challenges in bounding the effective Pareto regret and explain how the number of samples M affects the worst-case regret bound.

5.1 Assumptions

Let $\mathcal{F}_t = \sigma(x_{1,a_1}, \ldots, x_{t,a_t}, \boldsymbol{r}_{1,a_1}, \ldots, \boldsymbol{r}_{t-1,a_{t-1}})$ be the filtration up to round t containing all historical information about the selected arms and the received rewards. The following assumptions are commonly used in the stochastic linear bandit literature.

Assumption 1 (Boundedness) *For each arm $a \in \mathcal{A}$, $\|x_{t,a}\| \le 1$. Also, $\|\theta_*^{(\ell)}\| \le 1$ for all $\ell \in [L]$.*

Assumption 2 (Sub-Gaussian) *Each noise $\xi_t^{(\ell)}$ is conditionally R-sub-Gaussian, given the filtration $\mathcal{F}_t$ and for some $R \in \mathbb{R}^+$.*

Under the first assumption, each vector is both fixed and bounded for all rounds. If $\|x_{t,a}\| \le C$ and $\|\theta_*^{(\ell)}\| \le C$ are bounded for some constant C, then our regret bound would increase by a factor of C. Note that we do not assume any linear independence of the true parameters or noise vectors between objectives. Our assumptions are essentially the same as those used in the single objective stochastic linear bandit setting.

5.2 Challenges in bounding the effective Pareto regret

Recall the effective Pareto regret Equation (1). For any weight vector $\boldsymbol{w} \in \mathcal{S}^L$, since $\|\boldsymbol{w}\|_1 = 1$,

$$\min_{\ell \in [L]} \left\{ \left(\sum_{a_* \in \mathcal{C}_t^*} \beta_{a_*} \mu_{t,a_*}^{(\ell)} \right) - \mu_{t,a_t}^{(\ell)} \right\} \le \boldsymbol{w}^\top \left(\left(\sum_{a_* \in \mathcal{C}_t^*} \beta_{a_*} \boldsymbol{\mu}_{t,a_*} \right) - \boldsymbol{\mu}_{t,a_t} \right).$$

The algorithm `MOL-TS` optimistically evaluates reward vector $\tilde{\boldsymbol{\mu}}_{t,a}$ for each arm, and selects arm a_t randomly from the set $\tilde{\mathcal{C}}_t$. Hence, by Theorem 1, there exist weight vector, denoted $\boldsymbol{w}_t$, satisfying $a_t = \arg\max_{a \in \mathcal{A}} \boldsymbol{w}_t^\top \tilde{\boldsymbol{\mu}}_{t,a}$. But for true mean reward vector, let $\bar{a}_* = \arg\max_{a \in \mathcal{A}} \boldsymbol{w}_t^\top \boldsymbol{\mu}_{t,a}$ be the effective Pareto optimal arm for given weight vector $\boldsymbol{w}_t$. Then we have

$$\Delta_{t,a_t}^{EPR} \le \max_{\beta \in \mathcal{S}^{|\mathcal{C}_t^*|}} \left\{ \boldsymbol{w}_t^\top \left(\left(\sum_{a_* \in \mathcal{C}^*} \beta_{a_*} \boldsymbol{\mu}_{t,a_*} \right) - \boldsymbol{\mu}_{t,a_t} \right) \right\} \le \boldsymbol{w}_t^\top (\boldsymbol{\mu}_{t,\bar{a}_*} - \boldsymbol{\mu}_{t,a_t}).$$

The key insight is that the effective Pareto regret is bounded by the weighted sum of rewards under an arbitrary weight vector, and the same holds for Pareto regret. This analysis generalizes the single objective case, which is recovered by restricting $\boldsymbol{w}_t$ to a one-hot vector.

However, since the arm a_t is randomly selected from the set $\tilde{\mathcal{C}}_t$, both the weight vector $\boldsymbol{w}_t$, and the corresponding effective Pareto optimal arm $\bar{a}_*$ are random. Due to the randomness of the vector $\boldsymbol{w}_t$ and the optimal arm $\bar{a}_*$, analyzing the worst-case regret bound of TS algorithm becomes significantly more difficult. Also, unlike the single objective setting, the multi-objective setting involves multiple true parameters and corresponding RLS estimates. This complicates the problem of ensuring optimism, as there are multiple sampled parameters in TS algorithm (see example in Section 5.3). We resolve these theoretical challenges by adopting an *optimistic sampling strategy*.

5.3 Why do we need optimistic sampling?

In this section, we explain the necessity of the number of samples M. As we discuss in Section 5.2, the challenges in the worst-case regret analysis for TS algorithms lie in the difficulty of ensuring optimism in randomly selected arm a_t. When $\boldsymbol{w}_t$ is one-hot vector, the analysis aligns with the single objective setting [4, 2]. However, since $\boldsymbol{w}_t$ is random, the analysis requires that the randomly chosen arm is optimistically evaluated under a weighted sum of rewards. This probability can become exponentially small as the number of objectives increases.

Before providing a detailed explanation, we first define the event $\hat{\mathcal{E}}_t$ that the true parameters $\theta_*^{(\ell)}$ are close enough to the RLS estimate parameters $\hat{\theta}_t^{(\ell)}$, and define $c_{1,t}(\delta)$ which is the high probability bound on the distance between the true parameter and the RLS estimate,

$$\hat{\mathcal{E}}_t := \{\forall \ell \in [L] : \|\theta_*^{(\ell)} - \hat{\theta}_t^{(\ell)}\|_{V_t} \leq c_{1,t}(\delta)\}, \quad c_{1,t}(\delta) := R\sqrt{d \log\left(\frac{1 + (t-1)/(\lambda d)}{\delta / L}\right)} + \lambda^{1/2}.$$

We define the event $\dot{\mathcal{E}}_{t,a}^{(\ell)}$ for a specific arm a and objective ℓ, where the event has at least one parameter following anti-concentration property of being optimistic, i.e.,

$$\dot{\mathcal{E}}_{t,a}^{(\ell)} := \{\exists m \in [M] : x_{t,a}^\top(\tilde{\theta}_{t,m}^{(\ell)} - \hat{\theta}_t^{(\ell)}) \geq c_{1,t}(\delta)\|x_{t,a}\|_{V_t^{-1}}\}.$$

As the algorithm `MOL-TS` optimistically evaluate each arm a with Equation (2), the probability $\mathbb{P}(\dot{\mathcal{E}}_{t,a}^{(\ell)})$ increases as M increases. Suppose, for example, one follows standard TS algorithm by setting $M = 1$, that only one parameter is sampled for each objective. Previous studies [4, 2] have shown that the probability of an arm a being optimistically evaluated is at least $\tilde{p}$, i.e.,

$$\mathbb{P}\{x_{t,a}^\top(\tilde{\theta}_{t,1}^{(\ell)} - \hat{\theta}_t^{(\ell)}) \geq c_{1,t}(\delta)\|x_{t,a}\|_{V_t^{-1}}\} \geq \tilde{p}.$$

where $\tilde{p}$ is constant probability, that depends on the choice of sampling distribution. However, since $\boldsymbol{w}_t$ is random, the probability of ensuring this optimism for every objective is at least $\tilde{p}^L$. Since this probability can become exponentially small as the number of objectives increases, the regret can become exponentially large with $(1/\tilde{p})^L$.

Optimistic sampling strategy resolves this problem as `MOL-TS` optimistically evaluates the rewards from M multiple independent parameter samples for each objective. To prevent the exponential growth of the probability of ensuring optimism, the number of samples M must depend on the number of objectives L. The next lemma shows the minimum number of samples M for ensuring the event of optimism with constant probability.

Lemma 1 (Optimistic Sampling) *For any arm $a \in \mathcal{A}$, define the event of anti-concentration property of being optimism $\dot{\mathcal{E}}_{t,a} = \bigcap_{\ell \in [L]} \dot{\mathcal{E}}_{t,a}^{(\ell)}$. Then on event $\hat{\mathcal{E}}_t$, with $p = 0.15$ and $M \geq 1 - \frac{\log L}{\log(1-p)}$, we have $\mathbb{P}(\dot{\mathcal{E}}_{t,a}) \geq p$.*

The event $\dot{\mathcal{E}}_{t,a}$ is that the arm a is optimistically evaluated for every objective. Lemma 1 shows that the lower bound on the probability that arm a being optimistically evaluated remains constant by taking optimistic sampling strategy. The proof of this lemma is provided in Appendix B.1.

5.4 Worst-case regret bound

We now present the worst-case (frequentist) regret upper bound of `MOL-TS`, where the expectation is taken over all sources of randomness present in the problem setup.

Theorem 2 (Effective Pareto regret of `MOL-TS`) *With Assumptions 1 and 2, with $c = c_{1,t}(\delta)$ and $M = \lceil 1 - \frac{\log L}{\log(1-p)} \rceil$, the effective Pareto regret of the algorithm `MOL-TS` is upper-bounded by*

$$\mathbb{E}[EPR(T)] = \widetilde{O}(d^{3/2}\sqrt{T}).$$

Corollary 1 (Pareto regret of `MOL-TS`) *With same assumptions and initialization, the Pareto regret of the algorithm `MOL-TS` is upper-bounded by $\mathbb{E}[PR(T)] = \widetilde{O}(d^{3/2}\sqrt{T})$.*

Discussions of Theorem 2 and Corollary 1. Theorem 2 establishes that the expected effective Pareto regret of `MOL-TS` is bounded above by $\widetilde{O}(d^{3/2}\sqrt{T})$, where the regret has an additional $O(\log L)$ dependence due to the number of objectives, which is minimal. Additionally, Corollary 1 holds since $\Delta_{t,a_t}^{PR} \leq \Delta_{t,a_t}^{EPR}$. The details of the proof are provided in Appendix B. To the best of our knowledge, `MOL-TS` is the first TS algorithm with the worst-case regret guarantees in both Pareto regret and effective Pareto regret.

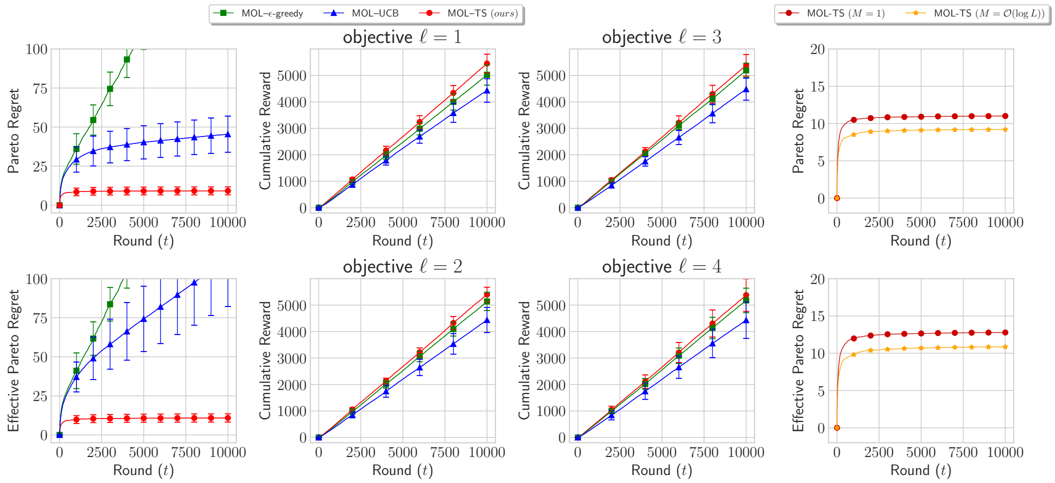

Figure 2: Experimental results with 4 objectives ($L = 4$). Plots in the left three columns measure the performances of `MOL-TS` and the others. Two plots in the first column measure the Pareto regret and the effective Pareto regret. Four plots in the second and third columns measure the cumulative reward for each objective. Plots in the right most column measure the performances of `MOL-TS` with $M = 1$ and $M = O(\log L)$. The error bars represent the 1-sigma standard deviation over 10 instances.

6 Experiments

In this section, we empirically evaluate the performance of our algorithm. We measure the Pareto regret and effective Pareto regret over $T = 10000$ rounds. Each experimental setup contains 10 different instances with fixed number of arms K, objectives L, and feature dimension d. We demonstrate the case where $K = 50$, $d = 5$, $L = 4$. The parameter vector for each objective $\theta_*^{(\ell)}$ has a norm of 1. Each round, d-dimensional context vectors are revealed for every arm, bounded by 1 in Euclidean norm. Upon playing an arm, the agent receives a reward vector with an additional noise term, where the noise values are sampled from a zero mean Gaussian distribution with $\sigma = 1$.

We compare the performance of `MOL-TS` with those basic novel algorithms : the Upper Confidence Bound algorithm, and ϵ-Greedy algorithm. The Upper Confidence Bound algorithm is `MOGLM-UCB` (represented as `MOL-UCB` in our experiments), from Lu et al. [11] in linear bandit setting and ϵ-Greedy algorithm is basic MOMAB algorithm with $\epsilon = 0.05$. Other algorithms cannot be applied in contextual setting, as they remove sub-optimal arm from the arm set. We also compare the performance of `MOL-TS` with and without the optimistic sampling. As shown in Figure 2, our proposed algorithm `MOL-TS` shows greater performance compared to other algorithms, with minimizing the Pareto regret and effective Pareto regret, and maximizing cumulative rewards in all objectives. Additionally, `MOL-TS` with optimistic sampling performs better in minimizing regret. Additional experiments with different settings of K, d and L are left in Appendix F, with additional algorithm `PFIwR` from Kim et al. [10] in linear bandit setting.

7 Discussions

In this paper, we study the multi-objective linear contextual bandit problem, where multiple conflicting objectives must be optimized simultaneously. We define the effective Pareto regret, whose definition considers the Pareto optimality of cumulative reward vectors. We propose a Thompson Sampling algorithm with optimistic sampling strategy, `MOL-TS`, that achieves the Pareto regret and effective Pareto regret of $\widetilde{O}(d^{3/2}\sqrt{T})$, matching the best known order for randomized linear bandit algorithms for single objective setting. Empirical results confirm the benefits of our proposed approach, demonstrating improved regret minimization and strong multi-objective performance.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

[3] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[4] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

[5] Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Artificial intelligence and statistics*, pages 939–947. PMLR, 2016.

[6] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

[7] 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.

[8] Madalina M Drugan and Ann Nowe. Designing multi-objective multi-armed bandits algorithms: A study. In *The 2013 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2013.

[9] Taehyun Hwang, Kyuwook Chai, and Min-hwan Oh. Combinatorial neural bandits. In *International Conference on Machine Learning*, pages 14203–14236. PMLR, 2023.

[10] Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. Learning the pareto front using bootstrapped observation samples. *arXiv preprint arXiv:2306.00096*, 2023.

[11] Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. *arXiv preprint arXiv:1905.12879*, 2019.

[12] Olvi L Mangasarian. *Nonlinear programming*. SIAM, 1994.

[13] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[14] Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, pages 766–776. PMLR, 2020.

[15] Cem Tekin and Eralp Turğay. Multi-objective contextual multi-armed bandit with a dominant objective. *IEEE Transactions on Signal Processing*, 66(14):3799–3813, 2018.

[16] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[17] 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.

[18] Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, pages 38499–38517. PMLR, 2023.

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

[20] Qiuyi Zhang. Optimal scalarizations for sublinear hypervolume regret. *arXiv preprint arXiv:2307.03288*, 2023.

[21] Richard Zhang and Daniel Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *International conference on machine learning*, pages 11096–11105. PMLR, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose the first Thompson Sampling algorithm with Pareto regret guarantees in multi-objective linear contextual bandit. Our contributions are clearly summarized in Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Appendix D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Theorems are presented with assumptions in detail (see Section 5). Details and explanations of the proofs are given in Appendix A and Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All codes of algorithms and experiments are provided in a ZIP file. Experimental results are provided in Section 6 and Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All codes of algorithms and experiments are provided in a ZIP file. Experimental results are provided in Section 6 and Appendix F.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The details of experimental settings are provided in Section 6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In both Section 6 and Appendix F, all the bars in regret plots and reward plots represent the 1-sigma standard deviation of experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include information about the computing environment used to run experiments in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research in our paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our paper has no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We clearly mention the sources of the comparator algorithms in Section 6.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper clearly describes our proposed algorithm in Section 4. The codes of our proposed algorithm are provided in a ZIP file.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our core method development in this research does not involve any usage of LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Proof of Theorem 1

In this section we present the proof for the necessary properties of effective Pareto optimal arms. Before proving this theorem, we use the following definition that denotes the convex hull of arbitrary set.

Definition 8 *Let $\mathcal{M} \subset \mathbb{R}^L$ be a set of mean reward vector $\boldsymbol{\mu}_a$. For all $a \in \mathcal{A}$. Define the convex hull of a set $\mathcal{M}$ as* $\mathbf{Conv}(\mathcal{M})$.

By definition, for any $\beta = (\beta_a)_{a\in\mathcal{A}} \in \mathcal{S}^{|\mathcal{A}|}$, we have

$$\sum_{a\in\mathcal{A}} \beta_a \boldsymbol{\mu}_a \in \mathbf{Conv}(\mathcal{M})$$

This convex hull covers all the convex combination of mean reward vectors of every arms. We note that, in the convex set $\mathbf{Conv}(\mathcal{M})$, the mean reward vector of the effective Pareto optimal arm satisfies the Pareto optimality in $\mathbf{Conv}(\mathcal{M})$. In other words, for all $a_* \in \mathcal{C}^*$, we have

$$\boldsymbol{\mu}_{a_*} \not\prec \boldsymbol{\mu}, \qquad \forall \boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M}) \setminus \{\boldsymbol{\mu}_{a_*}\} \tag{4}$$

However, for the arm $a \in \mathcal{A} \setminus \mathcal{C}^*$, there exists $\boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M})$ satisfying

$$\boldsymbol{\mu}_a \prec \boldsymbol{\mu}$$

The arm that is effective Pareto optimal, satisfies the Pareto optimality in $\mathbf{Conv}(\mathcal{M})$ (see Definition 5).

Other than the Pareto optimality, the next definition describes the optimality of one specific objective, with constraint on the other objectives.

Definition 9 ($\boldsymbol{\epsilon}$-constraint optimal) *Let $\boldsymbol{\epsilon}_\ell$ be $L-1$ dimensional arbitrary constraint vector*

$$\boldsymbol{\epsilon}_\ell = \begin{bmatrix}\epsilon^{(1)} & \dots & \epsilon^{(\ell-1)} & \epsilon^{(\ell+1)} & \dots & \epsilon^{(L)}\end{bmatrix}^\top \in \mathbb{R}^{L-1}$$

Given the constraint vector $\boldsymbol{\epsilon}_\ell$, the $\boldsymbol{\epsilon}_\ell$-constraint optimal vector among the set $\mathbf{Conv}(\mathcal{M})$*, denoted $\boldsymbol{\mu}_{\ell_*}$, is defined by*

$$\boldsymbol{\mu}_{\ell_*} = \underset{\boldsymbol{\mu}\in\mathbf{Conv}(\mathcal{M})}{\operatorname{argmax}} \{\mu^{(l)} \mid \mu^{(k)} \geq \epsilon^{(k)} \textit{ for all } k = 1, 2, ..., L, k \neq \ell\}$$

The vector is $\boldsymbol{\epsilon}$-constraint optimal (denoted $\boldsymbol{\mu}_*$) if, for all $\ell \in [L]$, there exists constraint vector $\boldsymbol{\epsilon}_\ell$ such that the vector $\boldsymbol{\mu}_*$ is $\boldsymbol{\epsilon}_\ell$-constraint optimal vector.

The constraint vector $\boldsymbol{\epsilon}_\ell$ is the lower bound values, such that the vector $\boldsymbol{\mu}$ dominates the constraint vector, except objective ℓ. Among those vector $\boldsymbol{\mu}$ satisfying constraint, the $\boldsymbol{\epsilon}_\ell$-constraint optimal vector is the one that has maximum value in objective ℓ. The $\boldsymbol{\epsilon}$-constraint optimal vector is such constraint vector $\boldsymbol{\epsilon}_\ell$ exists, as to be $\boldsymbol{\epsilon}_\ell$ constraint optimal, for all $\ell \in [L]$.

The next lemma shows the equivalence between $\boldsymbol{\epsilon}$-constraint optimality and Pareto optimality.

Lemma 2 *The vector $\boldsymbol{\mu}_* \in$* $\mathbf{Conv}(\mathcal{M})$ *is Pareto optimal if and only if it is $\boldsymbol{\epsilon}$-constraint optimal.*

Proof. ($\Longrightarrow$) Let $\boldsymbol{\mu}_*$ be Pareto optimal. Assume it is not $\boldsymbol{\epsilon}_\ell$-constraint optimal for some ℓ. Let the constraint vector be $\epsilon^{(k)} = \mu_*^{(k)}$ for $k = 1, ..., L, k \neq \ell$. Since it is not $\boldsymbol{\epsilon}_\ell$-constraint optimal, then there exists vector $\dot{\boldsymbol{\mu}}$ such that $\mu_*^{(k)} \leq \dot{\mu}^{(k)}$ for $k = 1, ..., L$ and $\mu_*^{(\ell)} < \dot{\mu}^{(\ell)}$. Since $\dot{\boldsymbol{\mu}}$ exists and dominates $\boldsymbol{\mu}_*$, this contradicts the definition of Pareto optimality.

($\Longleftarrow$) Let $\boldsymbol{\mu}_*$ be $\boldsymbol{\epsilon}$-constraint optimal. Suppose the constraint vector is defined as $\epsilon^{(\ell)} = \mu_*^{(\ell)}$ for all $\ell \in [L]$. Since $\boldsymbol{\mu}_*$ is $\boldsymbol{\epsilon}_\ell$-constraint optimal for every $\ell = 1, ..., L$, there is no other $\boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M})$ satisfying $\boldsymbol{\mu}_*^{(\ell)} < \dot{\boldsymbol{\mu}}^{(\ell)}$ and $\boldsymbol{\mu}_*^{(k)} \leq \dot{\boldsymbol{\mu}}^{(k)}$ when $k \neq \ell$, for every $\ell = 1, ..., L$. This holds the definition of Pareto optimality. ∎

Above lemma demonstrates that every Pareto optimal arm is also $\boldsymbol{\epsilon}$-constraint optimal. This equivalence also holds for non-convex set. But in the convex set $\mathbf{Conv}(\mathcal{M})$, the mean reward vector effective Pareto optimal arm satisfies the Pareto optimality in $\mathbf{Conv}(\mathcal{M})$, hence, it also satisfies $\boldsymbol{\epsilon}$-constraint optimal. It is enough to show that, for any $\boldsymbol{\epsilon}$-constraint optimal vector $\boldsymbol{\mu}_* \in \mathbf{Conv}(\mathcal{M})$, there exists weight vector $\boldsymbol{w} \in \mathcal{S}^L$ satisfying

$$\boldsymbol{\mu}_* = \underset{\boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M})}{\operatorname{argmax}} \ \boldsymbol{w}^\top \boldsymbol{\mu}$$

To prove above equation, we prove some of the lemmas that are useful for our proof.

Lemma 3 *let Ω be non-empty convex set in $\mathbb{R}^L$, not containing origin. Then there exists a vector $\boldsymbol{w} \in \mathcal{S}^L$ such that $\boldsymbol{w}^\top \boldsymbol{\mu} \geq 0$ holds for all $\boldsymbol{\mu} \in \Omega$.*

Proof. for $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_m \in \Omega$, define matrix and arbitrary vector as

$$M = [\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2 \quad ... \quad \boldsymbol{\mu}_m]^\top \in \mathbb{R}^{m \times L}, \quad \beta \in \mathcal{S}^m.$$

by convexity of set Ω, we have $M^\top \beta \in \Omega$, but $0 \notin \Omega$. So, there is no solution β, satisfying

$$M^\top \beta = 0, \quad \beta \in \mathcal{S}^m.$$

The solution still do not exist even if we remove the constraint $\|\beta\|_1 = 1$. By Proposition 1, the second condition of Gordan's theorem does not hold. Hence, there exist L-dimensional vector $\boldsymbol{w}$ that $\boldsymbol{w}^\top \boldsymbol{\mu}_i > 0$ holds for all $i = 1, ..., m$. Since $\boldsymbol{w}$ is non-zero vector, we can take $\boldsymbol{w}$ as $\sum_{\ell \in [L]} |w^{(\ell)}| = 1$. Define the set

$$V_{\boldsymbol{\mu}_i} = \{\boldsymbol{w} \in \mathbb{R}^L \mid \sum_{\ell \in [L]} |w^{(\ell)}| = 1, \boldsymbol{w}^\top \boldsymbol{\mu}_i \geq 0\}.$$

Then we can write

$$\bigcap_{i=1,...,m} V_{\boldsymbol{\mu}_i} \neq \emptyset.$$

Each set $V_{\boldsymbol{\mu}_i}$ is closed and bounded, hence, it is compact set. Since $\boldsymbol{\mu}_i$ was arbitrary chosen, the collection $(V_{\boldsymbol{\mu}})_{\boldsymbol{\mu}}$ satisfies finite intersection property. So, we have

$$\bigcap_{\boldsymbol{\mu} \in \Omega} V_{\boldsymbol{\mu}} \neq \emptyset.$$

■

Lemma 4 *Let Ω be non-empty convex set in $\mathbb{R}^L$, such that the vector $\boldsymbol{\mu} \in \Omega$ with all negative entries do not exist. Then, there exist vector $\boldsymbol{w} \in S^L$ such that $\boldsymbol{w}^\top \boldsymbol{\mu} \geq 0$ holds for all $\boldsymbol{\mu} \in \Omega$.*

Proof. For a vector $\boldsymbol{\mu} \in \Omega$, define the set

$$\mathcal{B}_{\boldsymbol{\mu}} = \{\boldsymbol{y} \in \mathbb{R}^L | y^{(\ell)} > \mu^{(\ell)}, \forall \ell \in [L]\},$$
$$\mathcal{B} = \bigcup_{\mu \in \Omega} \mathcal{B}_{\mu}.$$

If origin is in $\mathcal{B}$, then there exists $\boldsymbol{\mu}$ that $0 > \mu^{(\ell)}$ holds for all $\ell \in [L]$, which contradicts the assumption. If $\boldsymbol{y}_1 \in \mathcal{B}_{\boldsymbol{\mu}_1}$, $\boldsymbol{y}_2 \in \mathcal{B}_{\boldsymbol{\mu}_2}$, we have

$$\gamma \boldsymbol{y}_1 + (1-\gamma)\boldsymbol{y}_2 \in \mathcal{B}_{\gamma \boldsymbol{\mu}_1 + (1-\gamma)\boldsymbol{\mu}_2} \subset \mathcal{B}$$

for $\gamma \in [0, 1]$. Hence, $\mathcal{B}$ is convex set. By Lemma 3, there exist vector $\boldsymbol{w}$, satisfying $\boldsymbol{w}^\top \boldsymbol{y} \geq 0$ for all $\boldsymbol{y} \in \mathcal{B}$. If the vector has negative entry $w^{(\ell)} < 0$, we can choose $\boldsymbol{y} \in \mathcal{B}$ with large $y^{(\ell)}$ so that $\boldsymbol{w}^\top \boldsymbol{y} < 0$. Hence, we must have $w^{(\ell)} \geq 0$ for all $\ell \in [L]$. Also, since $\boldsymbol{w}$ is non-zero vector, we can restrict the vector in unit L-dimensional simplex. We now prove $\boldsymbol{w}^\top \boldsymbol{\mu} \geq 0$ for all $\boldsymbol{\mu} \in \Omega$. For any positive real $\epsilon > 0$, we have $\boldsymbol{\mu} + \epsilon \mathbf{1} \in \mathcal{B}$. If there exist $\delta > 0, \boldsymbol{\mu}$ with $\boldsymbol{w}^\top \boldsymbol{\mu} = -\delta$, we can choose $\epsilon < \delta$, so that

$$\boldsymbol{w}^\top (\boldsymbol{\mu} + \epsilon \mathbf{1}) = -\delta + \epsilon < 0.$$

Hence, we must have vector $\boldsymbol{w}$ that satisfies $\boldsymbol{w}^\top \boldsymbol{\mu} \geq 0$ for all $\boldsymbol{\mu} \in \Omega$, and $\boldsymbol{w} \in \mathcal{S}^L$ ■

The next lemma is the revision of Generalized Gordan's Theorem.

Lemma 5 (Generalized Gordan's Theorem) *Let Ω be non-empty convex set. Either one of the following statements holds, but not both.*

1. *There exists $\boldsymbol{\mu} \in \Omega$ whose entries are all negative.*
2. *There exists L-dimensional vector $\boldsymbol{w} \in \mathcal{S}^L$ satisfying $\boldsymbol{w}^\top \boldsymbol{\mu} \geq 0$ for all $\boldsymbol{\mu} \in \Omega$.*

Proof. ($\bar{1} \implies 2$) proof follows by Lemma 4.
($2 \implies \bar{1}$) If $\boldsymbol{\mu}$ is negative vector, we must have $\boldsymbol{w}^\top \boldsymbol{\mu} < 0$ for all $\boldsymbol{w} \in \mathcal{S}^L$. ■

Lemma 6 *For any $\boldsymbol{w} \in \mathcal{S}^L$, let $\mathcal{A}^*_{\boldsymbol{w}}$ be set of optimal arm, that has optimal weight sum reward given weight vector $\boldsymbol{w}$, i.e.,*

$$\mathcal{A}^*_{\boldsymbol{w}} = \arg\max_{a \in \mathcal{A}} \boldsymbol{w}^\top \boldsymbol{\mu}_a.$$

Then there exists effective Pareto optimal arm a_*, such that $a_* \in \mathcal{A}^*_w$.

Proof. Suppose there exist $\boldsymbol{w}$ that $a_* \notin \mathcal{A}^*_{\boldsymbol{w}}$ for any $a_* \in \mathcal{C}^*$. Let $\bar{a} \in \mathcal{A}^*_{\boldsymbol{w}}/\mathcal{C}^*$ that maximizes $\boldsymbol{w}^\top \boldsymbol{\mu}_a$. Since $\bar{a} \notin \mathcal{C}^*$, for every effective Pareto optimal arm $a_* \in \mathcal{C}^*$, there exist $\beta_{a_*} \geq 0$ satisfying

$$\boldsymbol{\mu}_{\bar{a}} \prec \sum_{a_* \in \mathcal{C}^*} \beta_{a_*} \boldsymbol{\mu}_{a_*}, \quad \sum_{a_* \in \mathcal{C}^*} \beta_{a_*} = 1.$$

Since $\boldsymbol{w}$ is vector with non-negative entries, we have

$$\boldsymbol{w}^\top \boldsymbol{\mu}_{\bar{a}} = \sum_{a_* \in \mathcal{C}^*} \beta_{a_*} \boldsymbol{w}^\top \boldsymbol{\mu}_{\bar{a}} \leq \sum_{a_* \in \mathcal{C}^*} \beta_{a_*} \boldsymbol{w}^\top \boldsymbol{\mu}_{a_*}.$$

We have, at least, one Pareto optimal arm with

$$\boldsymbol{w}^\top \boldsymbol{\mu}_{\bar{a}} \leq \boldsymbol{w}^\top \boldsymbol{\mu}_{a_*}$$

Such existence of a_* is guaranteed with existence of $\beta_{a_*} > 0$. ■

Now, we begin the proof of Theorem 1

Proof. Suppose a_* is effective Pareto optimal. By Equation (4), for any $\boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M}) \setminus \{\boldsymbol{\mu}_{a_*}\}$, we have $\boldsymbol{\mu}_{a_*} \not\prec \boldsymbol{\mu}$, hence, the vector $\boldsymbol{\mu}_{a_*}$ satisfies the Pareto optimality in $\mathbf{Conv}(\mathcal{M})$. By Lemma 2, $\boldsymbol{\mu}_{a_*}$ is also $\boldsymbol{\epsilon}$-constraint optimal, with $\epsilon^{(\ell)} = \mu^{(\ell)}_{a_*}$, such that no vector $\boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M}) \setminus \{\boldsymbol{\mu}_{a_*}\}$ satisfies $\mu^{(k)}_{a_*} - \mu^{(k)}_h \leq 0$ for $k = 1, ..., L$ and $\mu^{(\ell)}_{a_*} - \mu^{(\ell)}_h < 0$ for some ℓ. By Lemma 5, since the set $\mathbf{Conv}(\mathcal{M})$ is convex, the first statement does not hold. There exists L-dimensional vector $\boldsymbol{w} \in \mathcal{S}^L$ satisfying $\boldsymbol{w}^\top(\boldsymbol{\mu}_{a_*} - \boldsymbol{\mu}) \geq 0$ for all $\boldsymbol{\mu} \in \mathbf{Conv}(\mathcal{M})$. Hence, we have

$$a_* = \operatorname*{argmax}_{a \in \mathcal{A}} \boldsymbol{w}^\top \boldsymbol{\mu}_a$$

Conversly, Suppose $a^*_{\boldsymbol{w}} = \arg\max_{a \in \mathcal{A}} \boldsymbol{w}^\top \boldsymbol{\mu}_a$ is unique arm. By Lemma 6, the existence of effective Pareto optimal arm implies $a^*_{\boldsymbol{w}} \in \mathcal{C}^*$. ■

B Analysis of `MOLB-TS`

In this section, we provide the analysis of the worst-case regret of algorithm `MOLB-TS`.

We begin with the proof of Lemma 1.

B.1 Proof of Lemma 1

Proof. The parameters $(\tilde{\theta}^{(\ell)}_{t,m})_{m \in [M]}$ are sampled from Gaussian distribution $\mathcal{N}(\hat{\theta}^{(\ell)}_t, c^2_{1,t} V_t^{-1})$. Then for any given d-dimensional vector $x_{t,a}$, we can rewrite this probability distribution as

$$x^\top_{t,a} \tilde{\theta}^{(\ell)}_{t,m} \sim \mathcal{N}(x^\top_{t,a} \hat{\theta}^{(\ell)}_t, c^2_{1,t} \|x_{t,a}\|^2_{V_t^{-1}}).$$

Also, we can see the probability of the event $\dot{\mathcal{E}}_{t,a}^{(\ell)}$ as

$$\begin{aligned}\mathbb{P}(\dot{\mathcal{E}}_{t,a}^{(\ell)}) &= \mathbb{P}\{\exists m \in [M] : x_{t,a}^\top(\tilde{\theta}_{t,m}^{(\ell)} - \hat{\theta}_t^{(\ell)}) \geq c_{1,t}\|x_{t,a}\|_{V_t^{-1}}\} \\ &= \mathbb{P}\{\exists m \in [M] : \eta_m \geq 1\}\end{aligned}$$

where $(\eta_m)_{m\in[M]}$ is sampled from standard normal distribution $\mathcal{N}(0,1)$. For $M=1$, the probability of optimism $\mathbb{P}(\dot{\mathcal{E}}_{t,a}^{(\ell)})$ is bounded below by p. The probability of optimism, satisfying at least one sampled parameter, is bounded by

$$\mathbb{P}(\dot{\mathcal{E}}_{t,a}^{(\ell)}) \geq 1 - (1-p)^M.$$

These optimism events between different objectives are independent. Let the event $\dot{\mathcal{E}}_{t,a}$ be defined as

$$\dot{\mathcal{E}}_{t,a} = \bigcap_{\ell\in[L]} \dot{\mathcal{E}}_{t,a}^{(\ell)}$$

The event $\dot{\mathcal{E}}_{t,a}$ is the optimism satisfying for all objectives $\ell \in [L]$. Hence, we have

$$\mathbb{P}(\dot{\mathcal{E}}_{t,a}) \geq (1-(1-p)^M)^L.$$

To have $\mathbb{P}(\dot{\mathcal{E}}_{t,a}) \geq p$, we need to choose large enough M so that

$$(1-(1-p)^M)^L \geq p.$$

Rearrange the equation, we get

$$\begin{aligned}M &\geq \frac{\log(1-p^{1/L})}{\log(1-p)} \\ &= \frac{\log(1-p) - \log(1+p^{1/L}+p^{2/L}...+p^{(L-1)/L})}{\log(1-p)} \\ &= 1 + \frac{\log(1+p^{1/L}+p^{2/L}...+p^{(L-1)/L})}{\log\frac{1}{1-p}}\end{aligned}$$

With sampling by Gaussian distribution, we have $p = 0.15$. However, the probability p can be different by choosing different sampling distribution. But, as long as $p \in [0,1)$, we have

$$1 + \frac{\log L}{\log\frac{1}{1-p}} \geq 1 + \frac{\log(1+p^{1/L}+p^{2/L}...+p^{(L-1)/L})}{\log\frac{1}{1-p}}$$

Hence, by choosing M as

$$M \geq \frac{\log L}{\log\frac{1}{1-p}},$$

we get $\mathbb{P}(\dot{\mathcal{E}}_{t,a}) \geq p$ ■

Lemma 1 shows the minimum number of M for the inequality $\mathbb{P}(\dot{\mathcal{E}}_{t,a}) \geq p$ to hold.

B.2 Proof of Theorem 2

Before proving Theorem 2, we prove the following lemma, which bounds the conditional expectation of the regret of `MOLB-TS` at round t given the historical information up to that point.

Lemma 7 *For any filtration $\mathcal{F}_{t-1}$, on event $\hat{\mathcal{E}}$, we have*

$$\mathbb{E}_t[\Delta_{a_t}^{EPR}] \leq \left(1 + \frac{2}{0.15 - \frac{\delta}{T}}\right)(c_{1,t}(\delta) + c_{2,t}(\delta))\,\mathbb{E}_t[\|x_{t,a_t}\|_{V_t^{-1}}] + \frac{\delta}{T}\Delta_{max}$$

Proof. Let $(\beta_{t,a_*})_{a_*\in\mathcal{C}_t^*}$ be one that maximizes $\Delta_{a_t}^{EPR}$.

$$
\begin{aligned}
\Delta_{a_t}^{EPR} &= \max_{\beta\in\mathcal{S}^{|\mathcal{C}_t^*|}} \min_{\ell\in[L]} \left\{ \left(\sum_{a_*\in\mathcal{C}_t^*} \beta_{a_*} x_{t,a_*}^\top \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \theta_*^{(\ell)} \right\} \\
&= \min_{\ell\in[L]} \left\{ \left(\sum_{a_*\in\mathcal{C}_t^*} \beta_{t,a_*} x_{t,a_*}^\top \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \theta_*^{(\ell)} \right\}
\end{aligned}
$$

Let $\boldsymbol{w}_t$ be the weight vector in unit L-simplex sampled, described in Section 5.2. Then,

$$
\begin{aligned}
\Delta_{a_t}^{EPR} &= \min_{\ell\in[L]} \left\{ \left(\sum_{a_*\in\mathcal{C}_t^*} \beta_{t,a_*} x_{t,a_*}^\top \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \theta_*^{(\ell)} \right\} \\
&\le \sum_{\ell\in[L]} w_t^{(\ell)} \left(\left(\sum_{a_*\in\mathcal{C}_t^*} \beta_{t,a_*} x_{t,a_*}^\top \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \theta_*^{(\ell)} \right) \\
&= \sum_{a_*\in\mathcal{C}_t^*} \beta_{t,a_*} x_{t,a_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)} \right).
\end{aligned}
$$

By Theorem 1, there exists $\bar{a}_* \in \mathcal{C}^*$ satisfying

$$
\bar{a}_* = \operatorname*{argmax}_{a\in\mathcal{A}} x_{t,a}^\top \sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}.
$$

Hence, we have

$$
\begin{aligned}
\Delta_{a_t}^{EPR} &\le \sum_{a_*\in\mathcal{C}_t^*} \beta_{t,a_*} x_{t,a_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)} \right) && (5) \\
&\le x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)} \right) - x_{t,a_t}^\top \left(\sum_{l\in[L]} w_t^{(\ell)} \theta_*^{(\ell)} \right). && (6)
\end{aligned}
$$

From M multiple sampled parameters, we define $\tilde{\theta}_{a,t}^{(\ell)}$ as optimal sampled parameter with arm a, i.e.,

$$
\tilde{\theta}_{a,t}^{(\ell)} = \operatorname*{argmax}_{\tilde{\theta}_{t,m}} \{ x_{t,a}^\top \tilde{\theta}_{t,1}^{(\ell)}, x_{t,a}^\top \tilde{\theta}_{t,2}^{(\ell)}, ..., x_{t,a}^\top \tilde{\theta}_{t,M}^{(\ell)} \}.
$$

At round t, the arm a is evaluated with the sampled parameters $\tilde{\theta}_{a,t}^{(\ell)}$ for all $\ell \in [L]$. As we described in Section 5.2, we can write $\bar{a}_*, a_t$ as

$$
\begin{aligned}
\bar{a}_* &= \operatorname*{argmax}_{a\in\mathcal{A}} x_{t,a}^\top \sum_{l\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}, \\
a_t &= \operatorname*{argmax}_{a\in\mathcal{A}} x_{t,a}^\top \sum_{l\in[L]} w_t^{(\ell)} \tilde{\theta}_{a,t}^{(\ell)}.
\end{aligned}
$$

Define the events $\hat{\mathcal{E}}_t, \tilde{\mathcal{E}}_t$ such that the true parameters $\theta_*^{(\ell)}$ and all sampled parameters $(\tilde{\theta}_{t,m}^{(\ell)})_{m\in[M]}$ are close enough to the RLS estimate parameters $\hat{\theta}_t^{(\ell)}$ for all objectives, respectively.

$$
\begin{aligned}
\hat{\mathcal{E}}_t &:= \{ \forall \ell \in [L] : \|\theta_*^{(\ell)} - \hat{\theta}_t^{(\ell)}\|_{V_t} \le c_{1,t}(\delta) \}, \\
\tilde{\mathcal{E}}_t &:= \{ \forall m \in [M], \forall \ell \in [L] : \|\tilde{\theta}_{t,m}^{(\ell)} - \hat{\theta}_t^{(\ell)}\|_{V_t} \le c_{2,t}(\delta) \},
\end{aligned}
$$

where $c_{1,t}(\delta)$ and $c_{2,t}(\delta)$ are defined as

$$c_{1,t}(\delta) := R\sqrt{d\log\left(\frac{1+(t-1)/(\lambda d)}{\delta/L}\right)} + \lambda^{1/2},$$
$$c_{2,t}(\delta) := c_{1,t}(\delta)\sqrt{2d\log\frac{2LMdT}{\delta}}.$$

Let $\hat{\mathcal{E}} = \bigcap_{t\geq 0}\hat{\mathcal{E}}_t$. By Lemma 8, we have $\mathbb{P}(\hat{\mathcal{E}}) \geq 1-\delta$, and by Lemma 9, we have $\mathbb{P}_t(\tilde{\mathcal{E}}_t) = \mathbb{P}(\tilde{\mathcal{E}}_t \mid \mathcal{F}_t) \geq 1 - \delta/T$. Let $c_t(\delta) = c_{1,t}(\delta) + c_{2,t}(\delta)$. We separated arms into two sets with given weight vector, saturated and unsaturated [4].

- $\mathcal{B}_t$: set of saturated arms, that is, for all $a \in \mathcal{B}_t$, we have

$$x_{t,\bar{a}_*}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a}^\top\left(\sum_{l\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) > c_t(\delta)\|x_{t,a}\|_{V_t^{-1}}.$$

- $\overline{\mathcal{B}}_t$: set of unsaturated arms, that is, for all $a \in \overline{\mathcal{B}}_t$, we have

$$x_{t,\bar{a}_*}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a}^\top\left(\sum_{l\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) \leq c_t(\delta)\|x_{t,a}\|_{V_t^{-1}}.$$

Note that $\boldsymbol{w}_t$ is random variable, since the arm a_t is uniform randomly selected from $\tilde{\mathcal{C}}_t$. Hence, those sets of saturated and unsaturated arms $(\mathcal{B}_t, \bar{\mathcal{B}}_t)$ are not fixed. Let $\bar{a}_t = \operatorname{argmin}_{a\in\bar{\mathcal{B}}_t}\|x_{t,a}\|_{V_t^{-1}}$ be arm in $\bar{\mathcal{B}}_t$ with smallest matrix norm. From Equation (6), bounding the sub-optimality gap $\Delta_{a_t}^{EPR}$ on event $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}_t$, we have

$$\begin{aligned}
\Delta_{a_t}^{EPR} &\leq x_{t,\bar{a}_*}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a_t}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) \\
&= x_{t,\bar{a}_*}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) + x_{t,\bar{a}_t}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) \\
&\quad - x_{t,\bar{a}_t}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a_t}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right)
\end{aligned}$$

For any arm a, by the Cauchy-Schwarz inequality of matrix norm, we have

$$\left|x_{t,a}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a,t}^{(\ell)}\right) - x_{t,a}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right)\right| \leq \|x_{t,a}\|_{V_t^{-1}}\left(\sum_{\ell\in[L]} w_t^{(\ell)}\|\tilde{\theta}_{a,t}^{(\ell)} - \theta_*^{(\ell)}\|_{V_t}\right).$$

And by triangle inequality of norm, we have

$$\sum_{\ell\in[L]} w_t^{(\ell)}\|\tilde{\theta}_{a,t}^{(\ell)} - \theta_*^{(\ell)}\|_{V_t} \leq \sum_{\ell\in[L]} w_t^{(\ell)}\left(\|\tilde{\theta}_{a,t}^{(\ell)} - \hat{\theta}_t^{(\ell)}\|_{V_t} + \|\hat{\theta}_t^{(\ell)} - \theta_*^{(\ell)}\|_{V_t}\right).$$

And lastly, on event $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}_t$, we get

$$\|\tilde{\theta}_{a,t}^{(\ell)} - \hat{\theta}_t^{(\ell)}\|_{V_t} + \|\hat{\theta}_t^{(\ell)} - \theta_*^{(\ell)}\|_{V_t} \leq (c_{1,t}(\delta) + c_{2,t}(\delta)) = c_t(\delta)$$

In total, we get

$$\left|x_{t,a}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a,t}^{(\ell)}\right) - x_{t,a}^\top\left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right)\right| \leq \left(\sum_{\ell\in[L]} w_t^{(\ell)}\right) c_t(\delta)\|x_{t,a}\|_{V_t^{-1}}.$$

Hence, on event $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}_t$,

$$\begin{aligned}\Delta_{a_t}^{EPR} \leq\ & x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) + x_{t,\bar{a}_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{\bar{a}_t,t}^{(\ell)}\right) + \left(\sum_{\ell\in[L]} w_t^{(\ell)}\right) c_t(\delta)\|x_{t,\bar{a}_t}\|_{V_t^{-1}} \\ & - x_{t,\bar{a}_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a_t,t}^{(\ell)}\right) + \left(\sum_{\ell\in[L]} w_t^{(\ell)}\right) c_t(\delta)\|x_{t,a_t}\|_{V_t^{-1}} \\ =\ & x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) + x_{t,\bar{a}_t}^\top \left(\sum_{l\in[L]} w_t^{(\ell)}\tilde{\theta}_{\bar{a}_t,t}^{(\ell)}\right) + c_t(\delta)\|x_{t,\bar{a}_t}\|_{V_t^{-1}} \\ & - x_{t,\bar{a}_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a_t,t}^{(\ell)}\right) + c_t(\delta)\|x_{t,a_t}\|_{V_t^{-1}}.\end{aligned}$$

Since

$$a_t = \operatorname*{argmax}_{a\in\mathcal{A}} x_{t,a}^\top \sum_{l\in[L]} w_t^{(\ell)}\tilde{\theta}_{a,t}^{(\ell)},$$

we have

$$\begin{aligned}\Delta_{a_t}^{EPR} &\leq x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,\bar{a}_t}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) + c_t(\delta)\|x_{t,\bar{a}_t}\|_{V_t^{-1}} + c_t(\delta)\|x_{t,a_t}\|_{V_t^{-1}} \\ &\leq 2c_t(\delta)\|x_{t,\bar{a}_t}\|_{V_t^{-1}} + c_t(\delta)\|x_{t,a_t}\|_{V_t^{-1}},\end{aligned}$$

where the last inequality holds since $\bar{a}_t$ is unsaturated arm. This inequality holds on event $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}_t$. Define the conditional probability $\mathbb{P}_t = \mathbb{P}(\cdot \mid \mathcal{F}_t)$. Then, on event $\hat{\mathcal{E}}$, we have

$$\begin{aligned}\mathbb{E}_t[\Delta_{a_t}^{CPR}] &\leq \mathbb{E}_t[2c_t(\delta)\|x_{t,\bar{a}_t}\|_{V_t^{-1}} + c_t(\delta)\|x_{t,a_t}\|_{V_t^{-1}}] + (1-\mathbb{P}_t\{\tilde{\mathcal{E}}_t\})\Delta_{max} \\ &\leq \mathbb{E}_t[2c_t(\delta)\|x_{t,\bar{a}_t}\|_{V_t^{-1}} + c_t(\delta)\|x_{t,a_t}\|_{V_t^{-1}}] + \frac{\delta}{T}\Delta_{max}\end{aligned}$$

We bound the term $\mathbb{E}_t[\|x_{t,\bar{a}_t}\|_{V_t^{-1}}]$ with $\|x_{t,a_t}\|_{V_t^{-1}}$. We have

$$\begin{aligned}\mathbb{E}_t[\|x_{t,a_t}\|_{V_t^{-1}}] \geq\ & \mathbb{E}_t[\|x_{t,a_t}\|_{V_t^{-1}} \mid a_t \in \overline{\mathcal{B}}_t]\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\} \\ \geq\ & \|x_{t,\bar{a}_t}\|_{V_t^{-1}}\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\}\end{aligned}$$

The probability $\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\}$ has randomness over algorithm selecting arm from empirical effective Pareto front $\tilde{\mathcal{C}}_t$, where the set $\overline{\mathcal{B}}_t$ varies on this random selection. More precisely, the set $\mathcal{B}_t$ and $\overline{\mathcal{B}}_t$ change as $\boldsymbol{w}_t$ changes. But, for any given $\boldsymbol{w}_t$, the probability $\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\}$ bounds below by the probability that at least one unsaturated arm is evaluated higher compared to all saturated arms, i.e.,

$$\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\} \geq \mathbb{P}_t\left\{\exists a \in \overline{\mathcal{B}}_t : x_{t,a}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a,t}^{(\ell)}\right) > \max_{a'\in\mathcal{B}_t} x_{t,a'}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a',t}^{(\ell)}\right)\right\},$$

where the unsaturated arm exists by $\bar{a}_* \in \overline{\mathcal{B}}_t$. Hence this probability bounds below by the probability that the arm $\bar{a}_*$ is evaluated higher compared to all saturated arms, i.e.,

$$\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\} \geq \mathbb{P}_t\left\{x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{\bar{a}_*,t}^{(\ell)}\right) > \max_{a'\in\mathcal{B}_t} x_{t,a'}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\tilde{\theta}_{a',t}^{(\ell)}\right)\right\}.$$

On event $\tilde{\mathcal{E}}_t$, those saturated arms $a' \in \mathcal{B}_t$ satisfy

$$x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) - x_{t,a'}^\top \left(\sum_{l\in[L]} w_t^{(\ell)}\theta_*^{(\ell)}\right) > c_t(\delta)\|x_{t,a'}\|_{V_t^{-1}} \tag{7}$$

and

$$x_{t,a'}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \tilde{\theta}_{a',t}^{(\ell)}\right) - x_{t,a'}^\top \left(\sum_{l\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}\right) \le c_t(\delta)\|x_{t,a'}\|_{V_t^{-1}}. \tag{8}$$

Subtracting Equation (8) from Equation (7), we get

$$x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}\right) - x_{t,a'}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \tilde{\theta}_{a',t}^{(\ell)}\right) \ge 0.$$

for all $a' \in \mathcal{B}_t$. Using this inequality, the Probability bounds as

$$\begin{aligned}
\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\} &\ge \mathbb{P}_t\left\{x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \tilde{\theta}_{\bar{a}_*,t}^{(\ell)}\right) > x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}\right), \tilde{\mathcal{E}}_t\right\} \\
&\ge \mathbb{P}_t\left\{x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \tilde{\theta}_{\bar{a}_*,t}^{(\ell)}\right) > x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}\right)\right\} - (1 - \mathbb{P}_t\{\tilde{\mathcal{E}}_t\})
\end{aligned}$$

Since $\tilde{\theta}_{\bar{a}_*,t}^{(\ell)}$ is objective wise independent, the probability bounds by objective wise probability,

$$\mathbb{P}_t\left\{x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \tilde{\theta}_{\bar{a}_*,t}^{(\ell)}\right) > x_{t,\bar{a}_*}^\top \left(\sum_{\ell\in[L]} w_t^{(\ell)} \theta_*^{(\ell)}\right)\right\} \ge \bigcap_{\ell\in[L]} \mathbb{P}_t\{x_{t,\bar{a}_*}^\top \tilde{\theta}_{\bar{a}_*,t}^{(\ell)} > x_{t,\bar{a}_*}^\top \theta_*^{(\ell)}\}$$

Removing the random vector $\boldsymbol{w}_t$, this inequality holds for any $\boldsymbol{w}_t$. In other words, this inequality holds for any random selection of arms from the set $\tilde{\mathcal{C}}_t$. Hence, we get

$$\begin{aligned}
\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\} &\ge \bigcap_{\ell\in[L]} \mathbb{P}_t\{x_{t,\bar{a}_*}^\top \tilde{\theta}_{\bar{a}_*,t}^{(\ell)} > x_{t,\bar{a}_*}^\top \theta_*^{(\ell)}\} - (1 - \mathbb{P}_t\{\tilde{\mathcal{E}}_t\}) \\
&\ge \bigcap_{\ell\in[L]} \mathbb{P}_t\{x_{t,\bar{a}_*}^\top \tilde{\theta}_{\bar{a}_*,t}^{(\ell)} > x_{t,\bar{a}_*}^\top \theta_*^{(\ell)}\} - \frac{\delta}{T} \\
&= \mathbb{P}_t\{x_{t,\bar{a}_*}^\top \tilde{\theta}_{\bar{a}_*,t}^{(1)} > x_{t,\bar{a}_*}^\top \theta_*^{(1)}\}^L - \frac{\delta}{T}
\end{aligned}$$

As we remove $\boldsymbol{w}_t$, the probability $\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\}$ gets exponentially small as the number of objective increases. We remove this by adopting optimistic sampling strategy. With the number multiple samples M, following Lemma 1, we have

$$\mathbb{P}_t\{x_{t,\bar{a}_*}^\top \tilde{\theta}_{\bar{a}_*,t}^{(1)} > x_{t,\bar{a}_*}^\top \theta_*^{(1)}\} \ge 1 - (1-p)^M.$$

Hence, we get

$$\mathbb{P}_t\{a_t \in \overline{\mathcal{B}}_t\} \ge (1 - (1-p)^M)^L - \frac{\delta}{T}$$

This inequality holds for any $\boldsymbol{w}_t$. With $M = \lceil 1 - \frac{\log L}{\log(1-p)} \rceil$, we have $(1-(1-p)^M)^L \ge p$. Finally, we have

$$\mathbb{E}_t[\|x_{a_t}\|_{V_t^{-1}}] \ge \|x_{\bar{a}_t}\|_{V_t^{-1}} \left(p - \frac{\delta}{T}\right)$$

Replacing the term $\|x_{\bar{a}_t}\|_{V_t^{-1}}$ to $\|x_{a_t}\|_{V_t^{-1}}$, we get.

$$\mathbb{E}_t[\Delta_{a_t}^{EPR}] \le \left(1 + \frac{2}{p - \frac{\delta}{T}}\right) c_t(\delta)\mathbb{E}_t[\|x_{a_t}\|_{V_t^{-1}}] + \frac{\delta}{T}\Delta_{max}$$

■

Now we begin the proof of Theorem 2.

Proof. We have

$$\begin{aligned}
\mathbb{E}[EPR(T)] &= \sum_{t=1}^{T} \mathbb{E}[\Delta_{a_t}^{EPR}] \\
&= \mathbb{P}(\hat{\mathcal{E}}) \sum_{t=1}^{T} \mathbb{E}[\Delta_{a_t}^{EPR} \mathbb{1}\{\hat{\mathcal{E}}\}] + (1 - \mathbb{P}(\hat{\mathcal{E}}))\Delta_{\max} \\
&\leq \sum_{t=1}^{T} \mathbb{E}[\Delta_{a_t}^{EPR} \mathbb{1}\{\hat{\mathcal{E}}\}] + \delta \Delta_{\max} \\
&= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}_t[\Delta_{a_t}^{EPR}] \mathbb{1}\{\hat{\mathcal{E}}\}] + \delta \Delta_{\max}
\end{aligned}$$

By Lemma 7, bounding the term $\mathbb{E}_t[\Delta_{a_t}^{EPR}]$, we have

$$\begin{aligned}
\mathbb{E}[EPR(T)] &\leq \sum_{t=1}^{T} \left(1 + \frac{2}{0.15 - \frac{\delta}{T}}\right) c_t(\delta) \mathbb{E}[\mathbb{E}_t[\|x_{t,a_t}\|_{V_t^{-1}}] \mathbb{1}\{\hat{\mathcal{E}}\}] + 2\delta\Delta_{\max} \\
&\leq \left(1 + \frac{2}{0.15 - \frac{\delta}{T}}\right) c_T(\delta) \mathbb{E}[\sum_{t=1}^{T} \mathbb{E}_t[\|x_{t,a_t}\|_{V_t^{-1}}] \mathbb{1}\{\hat{\mathcal{E}}\}] + 2\delta\Delta_{\max} \\
&= \left(1 + \frac{2}{0.15 - \frac{\delta}{T}}\right) c_T(\delta) \mathbb{E}[\sum_{t=1}^{T} \|x_{t,a_t}\|_{V_t^{-1}}] + 2\delta\Delta_{\max} \\
&\leq \left(1 + \frac{2}{0.15 - \frac{\delta}{T}}\right) c_T(\delta) \sqrt{2Td \log\left(1 + \frac{T}{\lambda}\right)} + 2\delta\Delta_{\max},
\end{aligned}$$

where the last inequality follows by Proposition 2, ∎

C Additional Technical Tools

Proposition 1 (Gordan's Theorem, page 31, Mangasarian 12) *For given matrix $M \in \mathbb{R}^{m \times L}$, either one of the following statements holds, but not both.*

1. *There exists L-dimensional vector $\mathbf{w}$, that $M\mathbf{w}$ has all positive entries.*
2. *$M^\top \beta = 0$, $\beta \succ 0$ has solution $\beta \in \mathbb{R}^m$.*

Proposition 2 (Lemma 11, Abbasi-Yadkori et al. 1) *Let $\lambda \geq 1$. For arbitrary sequence $(x_{t,a_t})_{t\in[T]}$, we have*

$$\sum_{t=1}^{T} \|x_{t,a_t}\|^2_{V_t^{-1}} \leq 2d \log\left(1 + \frac{T}{\lambda}\right).$$

Lemma 8 (Theorem 2, Abbasi-Yadkori et al. 1) *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. Let $(\xi_t^{(\ell)})$ be a real-valued stochastic process such that $\xi_t^{(\ell)}$ is conditionally R-sub-Gaussian, given filtration $\mathcal{F}_t$ for any $\ell \in [L]$. Then with probability at least $1-\delta$, the event*

$$\hat{\mathcal{E}}_t = \left\{ \forall \ell \in [L] : \|\hat{\theta}_t^{(\ell)} - \theta_*^{(\ell)}\|_{V_t} \leq R\sqrt{d \log\left(\frac{1 + (t-1)/(\lambda d)}{\delta / L}\right)} + \lambda^{1/2} \right\}$$

holds for all $t \geq 1$.

Proof. By Theorem 2 in Abbasi-Yadkori et al. [1], and union bound with L. ∎

Lemma 9 (Definition 1, Abeille and Lazaric 2) *On event $\hat{\mathcal{E}}_t$, with probability at least $1-\delta$, all sampled parameters $(\tilde{\theta}_{t,m}^{(\ell)})_{m\in[M],\ell\in[L]}$ follow concentration property, i.e.,*

$$\tilde{\mathcal{E}}_t := \left\{ \|\tilde{\theta}_{t,m}^{(\ell)} - \hat{\theta}_t^{(\ell)}\|_{V_t} \leq \sqrt{2d\log\frac{2LMd}{\delta}} \left(R\sqrt{d\log\left(\frac{1+(t-1)/(\lambda d)}{\delta/L}\right)} + \lambda^{1/2} \right) \right\}.$$

for all $m \in [M], \ell \in [L]$.

Proof. By Definition 1 in Abeille and Lazaric [2], and union bound with M and L. ■

D Discussions

Our proposed algorithm demonstrates strong empirical performance with various settings. But its theoretical worst-case regret bound is not tighter than that of UCB-based algorithms. This gap between UCB-type algorithms and TS algorithms is well-known in the regret analysis of previous TS algorithms [4, 2] bounding the worst-case frequentist regret.

Our work is studied under the standard linear contextual bandit setting. Our framework can be readily extended to generalized linear contextual bandits. Extension to more complex function class such as neural networks requires analysis that is beyond the scope of this work, but is certainly the promising avenue for future work. Yet, our work introduces the first randomized algorithm with Pareto regret guarantees in the multi-objective bandit framework. We hope that our work lays a foundation basis for extending such techniques to follow-up works.

E Computing resources for experiments

All experiments are conducted with INTEL(R) XEON(R) GOLD 6526Y CPU and 4 TB memory. The software environment includes Python 3.12.7, Scipy 1.14.1, and Numpy 1.26.4. The experiments took approximately 4 hours to 1 day, as it takes longer with increasing numbers of arms, dimensions, and objectives.

F Additional experimental results

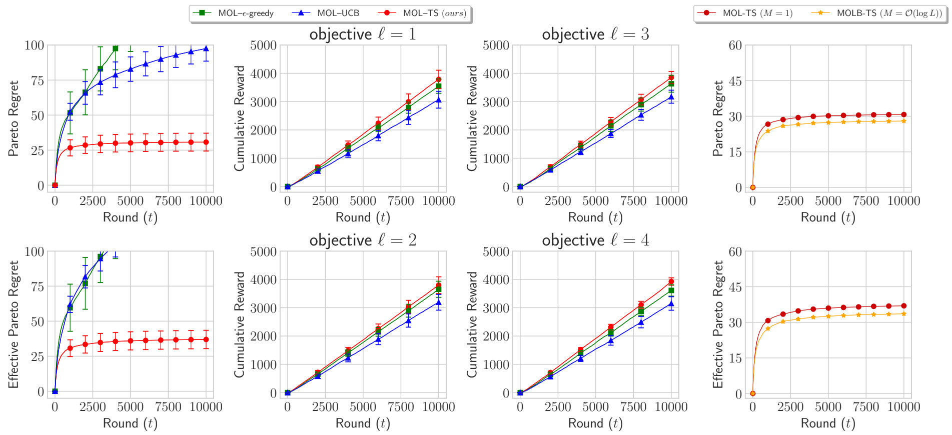

Figure 3: Experimental results with $K = 50,\ d = 10,\ L = 4$

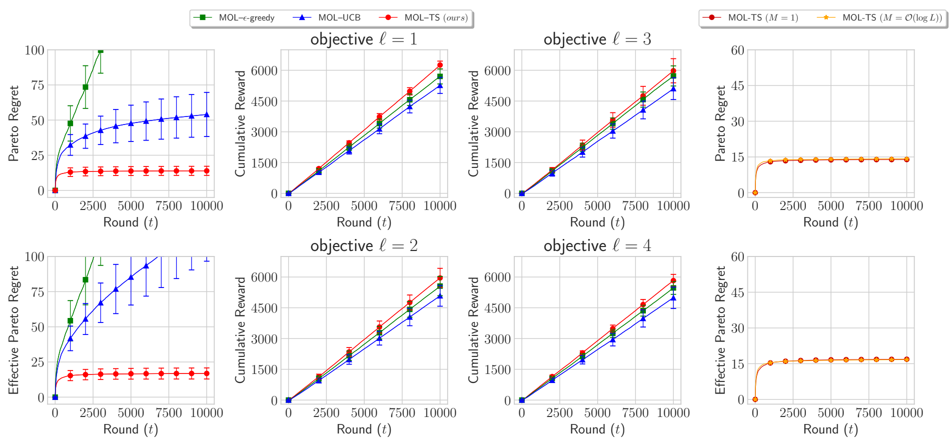

Figure 4: Experimental results with $K = 100,\ d = 5,\ L = 4$

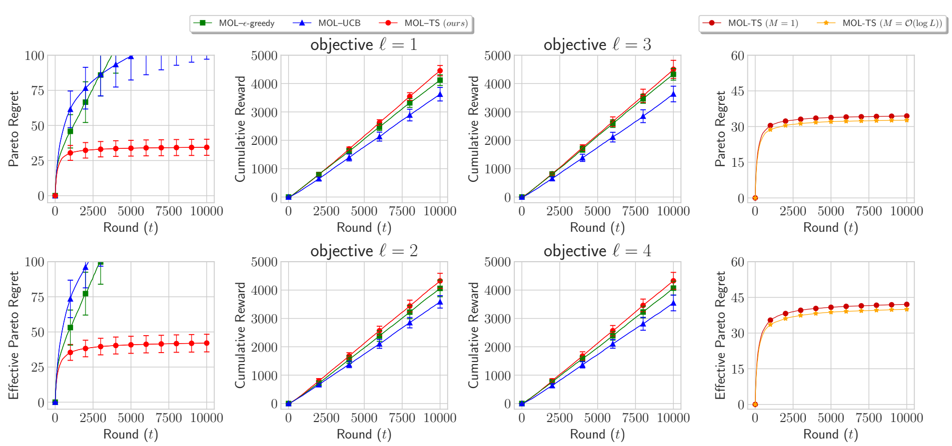

Figure 5: Experimental results with $K = 100,\ d = 10,\ L = 4$

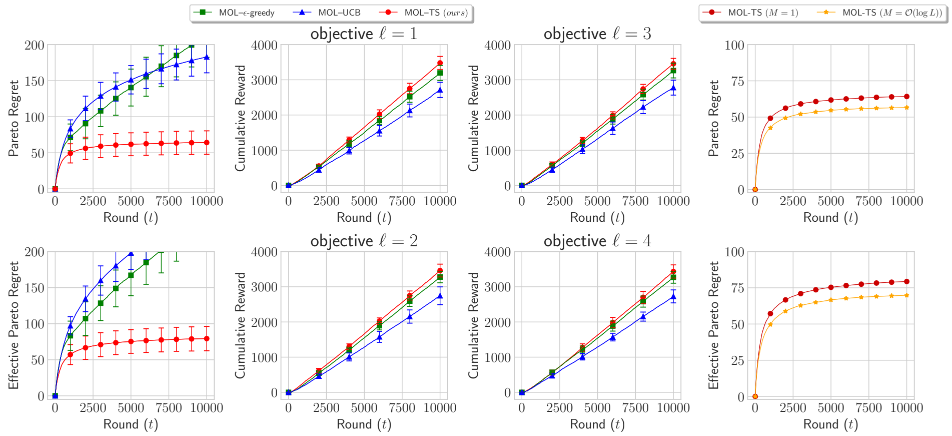

Figure 6: Experimental results with $K = 100,\ d = 15,\ L = 4$

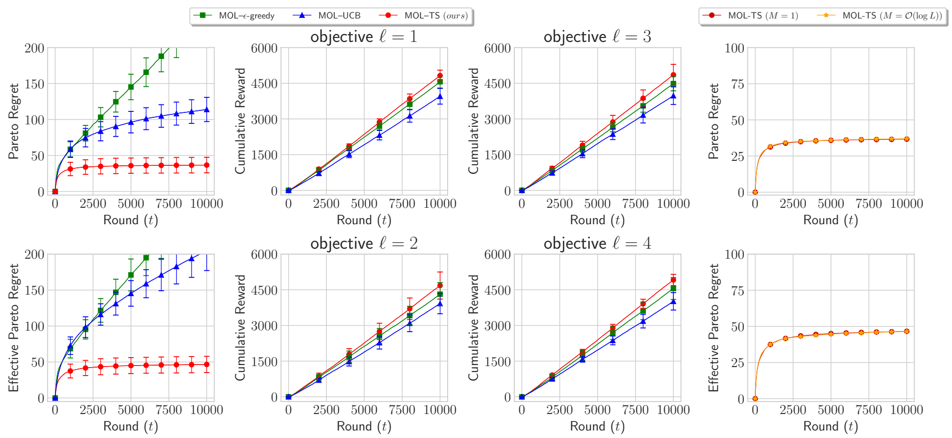

Figure 7: Experimental results with $K = 200,\ d = 10,\ L = 4$

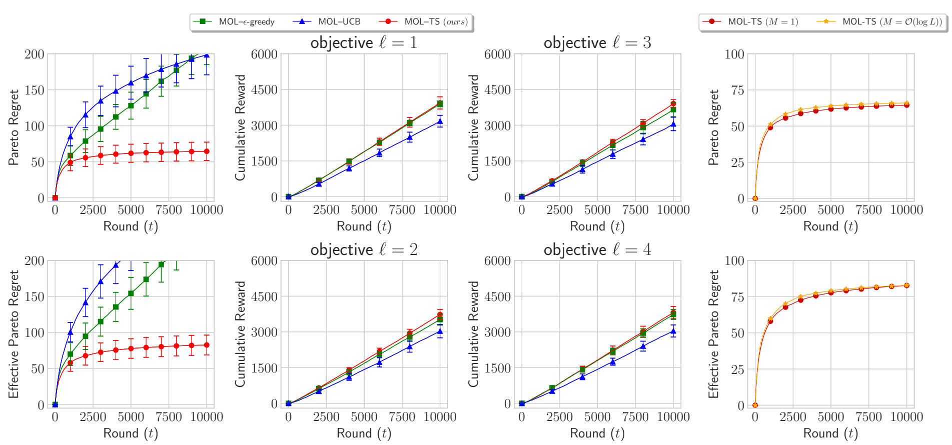

Figure 8: Experimental results with $K = 200,\ d = 15,\ L = 4$

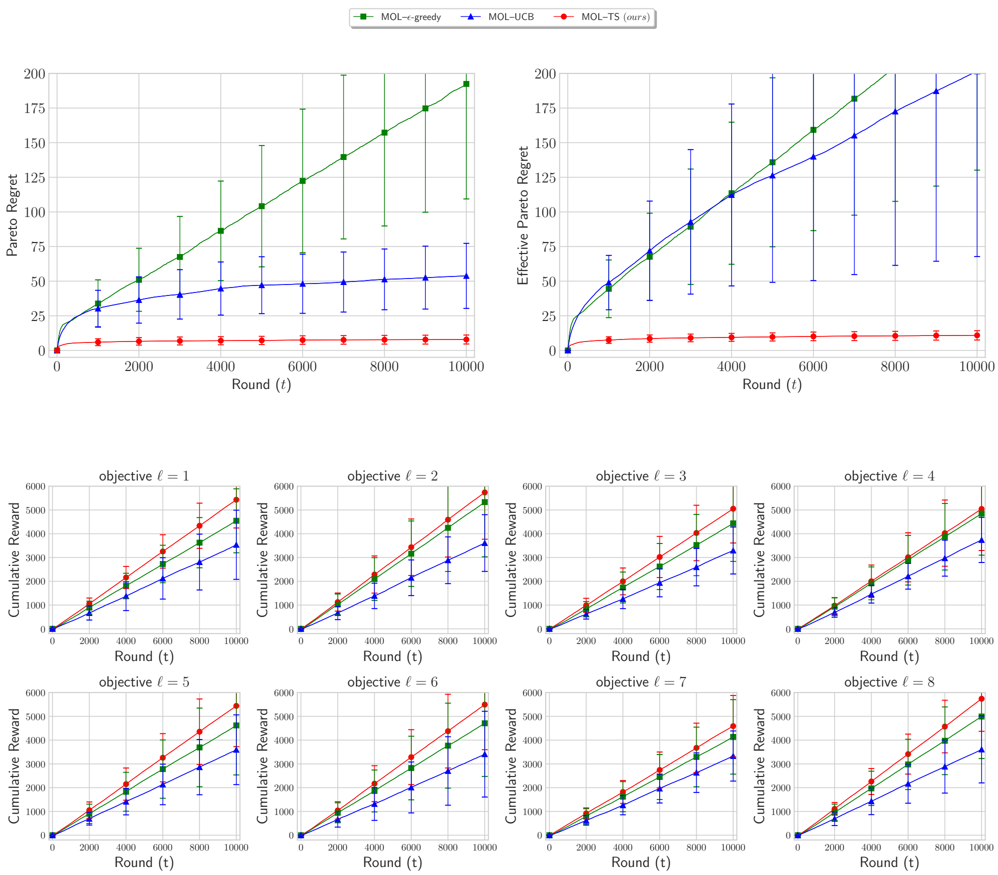

Figure 9: Experimental results with $K = 50,\ d = 5,\ L = 8$

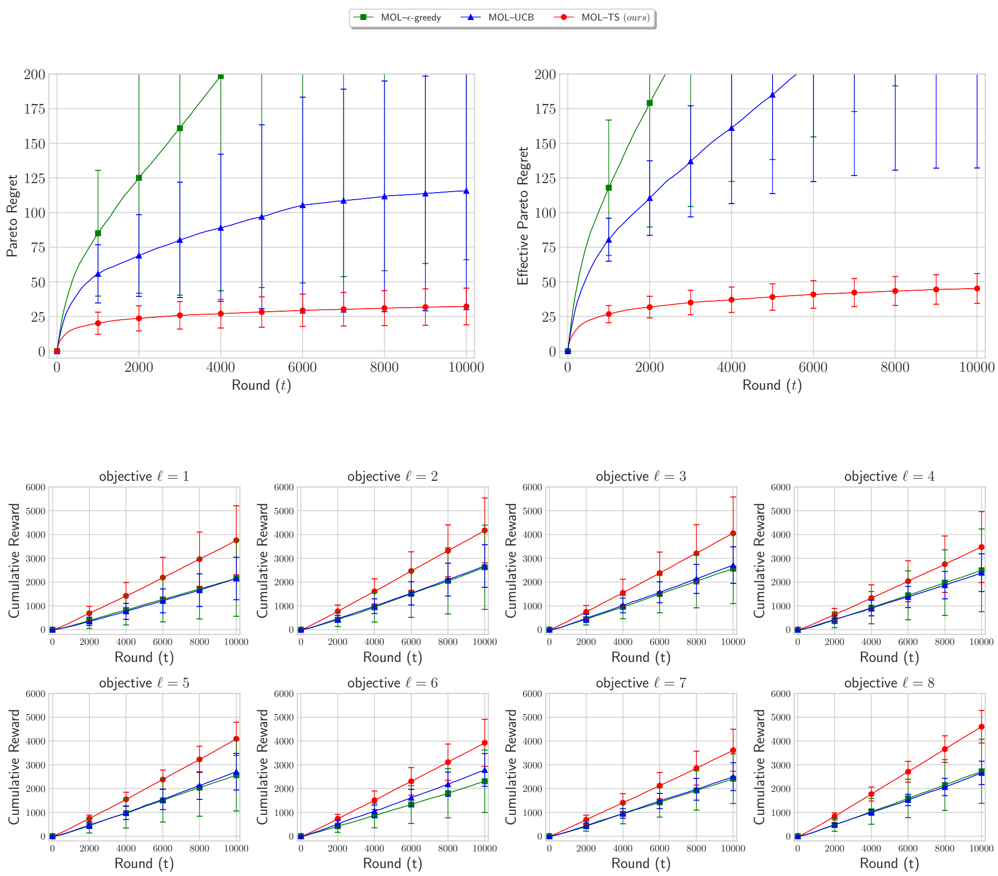

Figure 10: Experimental results with $K = 100,\ d = 10,\ L = 8$

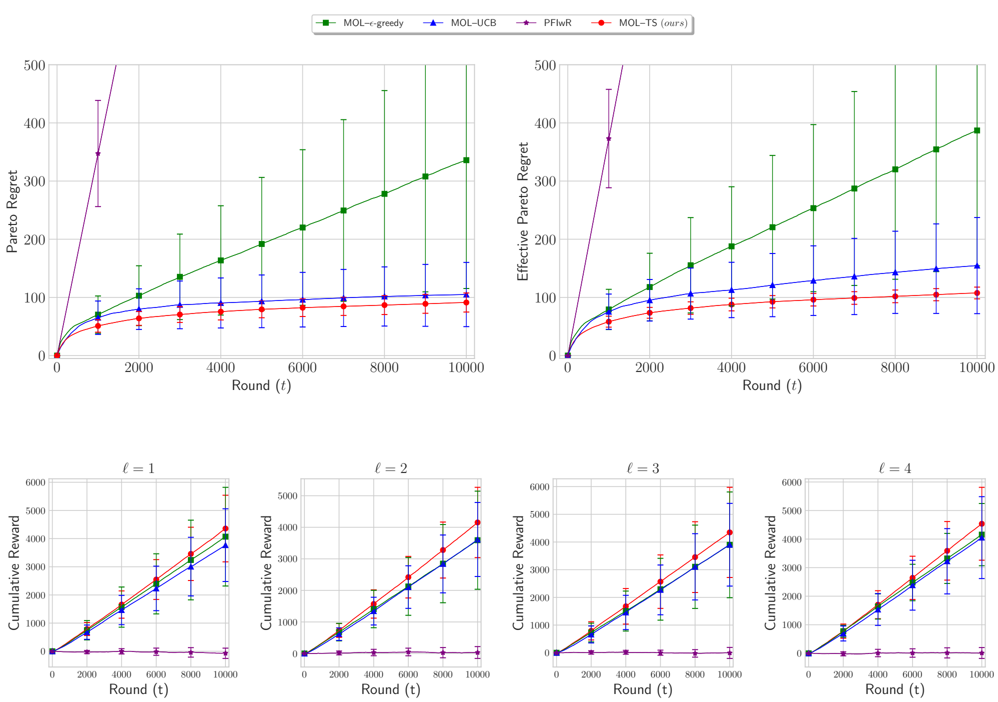

Figure 11: Experimental results with $K = 100,\ d = 10,\ L = 4$, linear (non-contextual) setting.

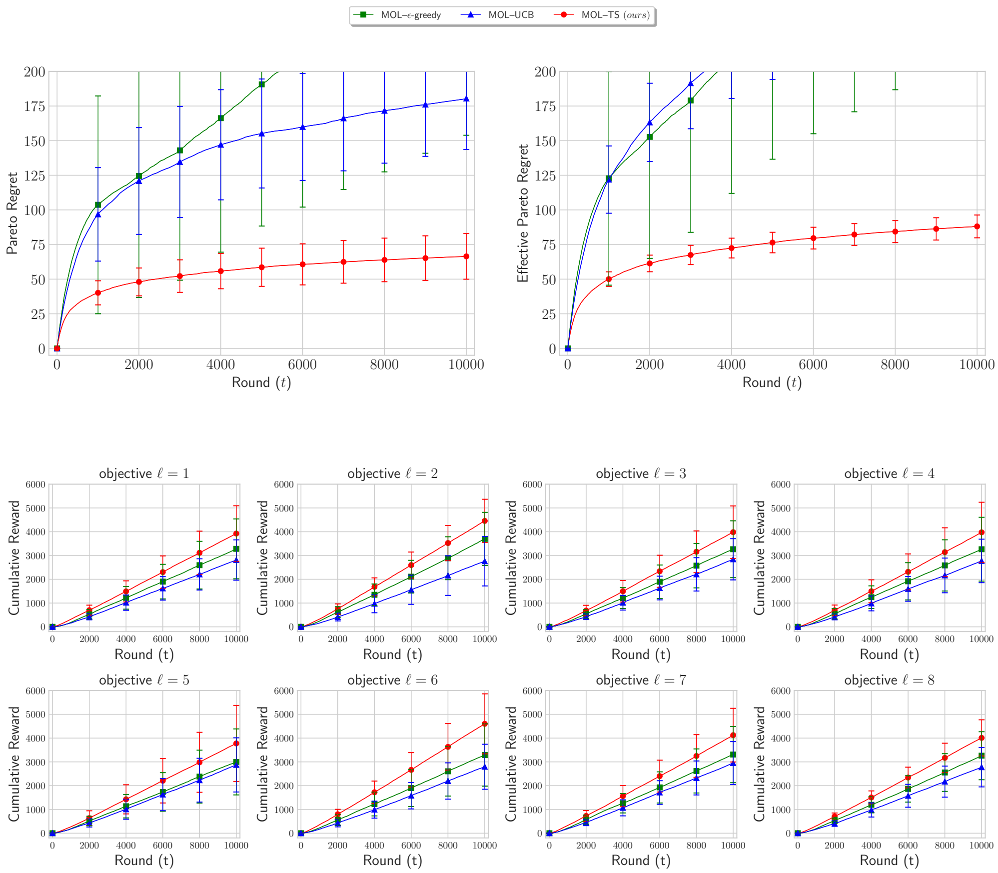

Figure 12: Experimental results with $K = 200,\ d = 15,\ L = 8$

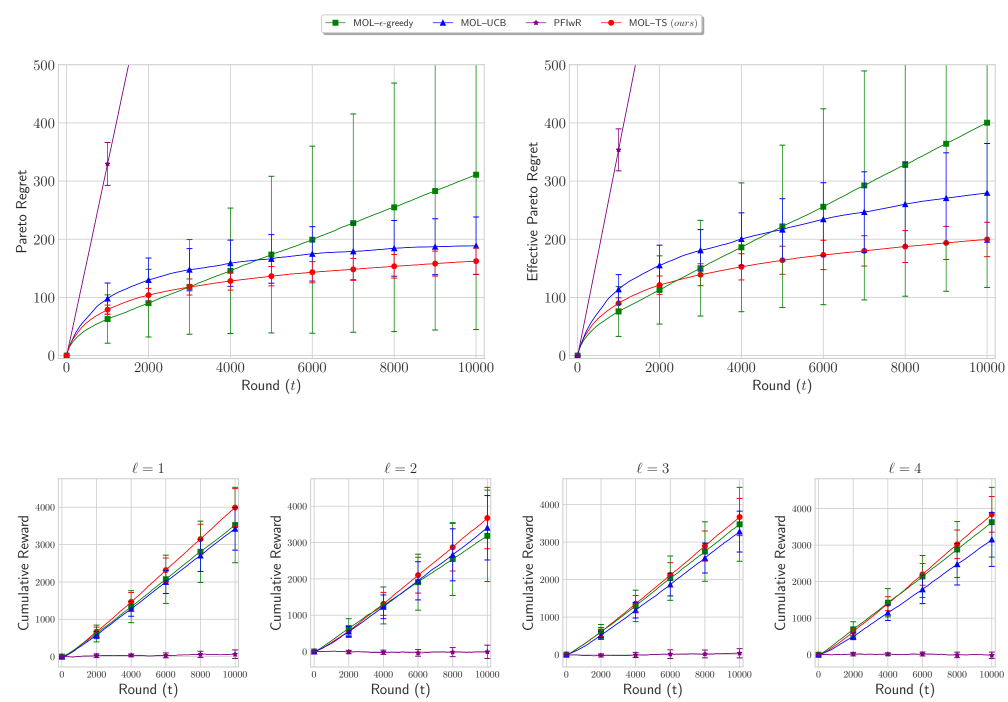

Figure 13: Experimental results with $K = 200,\ d = 15,\ L = 4$, linear (non-contextual) setting.