Linear Bandits with Partially Observable Features

Wonyoung Kim [*1] **Sungwoo Park** [*2] **Garud Iyengar** [3] **Assaf Zeevi** [3] **Min-hwan Oh** [2]

Abstract

We study the linear bandit problem that accounts for partially observable features. Without proper handling, unobserved features can lead to linear regret in the decision horizon T, as their influence on rewards is unknown. To tackle this challenge, we propose a novel theoretical framework and an algorithm with sublinear regret guarantees. The core of our algorithm consists of (i) feature augmentation, by appending basis vectors that are orthogonal to the row space of the observed features; and (ii) the introduction of a doubly robust estimator. Our approach achieves a regret bound of $\widetilde{O}(\sqrt{(d+d_h)T})$, where d is the dimension of the observed features and d_h depends on the extent to which the unobserved feature space is contained in the observed one, thereby capturing the intrinsic difficulty of the problem. Notably, our algorithm requires no prior knowledge of the unobserved feature space, which may expand as more features become hidden. Numerical experiments confirm that our algorithm outperforms both non-contextual multi-armed bandits and linear bandit algorithms depending solely on observed features.

1. Introduction

We consider a linear bandit problem where the learning agent has access to only a *subset* of the features, while the reward is determined using the *complete set* of features, including both observed and unobserved elements. Conventional linear bandit problems rely on the assumption that the rewards are linear in only observed features, without accounting for the potential presence of unobserved features. However, in many real-world applications, rewards are often affected by the unobserved—hence *latent*—features that are not observable by the agent. For example, in online advertising without personalization—i.e., when serving general users—each advertisement is associated not only with observable features such as its category, format, or display time, but also with unobservable factors such as emotional appeal, creative design quality, and brand familiarity. These latent factors influence users' click-through rates, yet they are not directly quantifiable. Similarly, in clinical trials, treatment outcomes depend not only on observable features like dosage and formulation, but also on unobserved factors such as manufacturing variability and potential side effect risks. Hence, accurately incorporating latent features is essential for providing precise recommendations.

To address latent features, Park & Faradonbeh (2022; 2024); Kim et al. (2023a); Zeng et al. (2025) typically assume that the true features follow a specific distribution, such as a Gaussian distribution. Establishing a sublinear regret bound in the decision horizon without such structural assumptions on the latent features remains a significant challenge and has not been accomplished yet. Key challenges in the bandit problem with partially observable features arise from the complete lack of information about the latent features—indeed, the agent does not even know whether the features are partially observed. We tackle these challenges by proposing a novel linear bandit algorithm that is agnostic to partial observability. Despite having no knowledge of the unobserved features, our algorithm achieves a tighter regret bound than both linear bandit algorithms that consider only observed features and multi-armed bandit (MAB) algorithms that ignore features entirely. For brevity, we will refer to linear bandit algorithms relying solely on observed features as "OLB algorithms" henceforth. Specifically, our proposed algorithm achieves a $\sqrt{T}$-rate regret bound, without requiring any prior knowledge of the unobserved features, where T is the decision horizon.

The key idea of our proposed algorithm can be summarized in the following two procedures: (i) reconstructing feature vectors via feature augmentation to capture the influence of unobserved features on rewards, and (ii) introducing a novel doubly robust (DR) estimator to mitigate information loss due to partial observability. For (i), we decompose the reward into two additive components: one projected onto the row space spanned by the observed features, and the other projected onto its orthogonal complement. The former term maximally captures the effect of observed features, while

[*]Equal contribution [1]Chung-Ang University, Seoul, Korea [2]Seoul National University, Seoul, Korea [3]Columbia University, New York, USA. Correspondence to: Min-hwan Oh <minoh@snu.ac.kr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

the latter minimizes the impact of unobserved features. We then *augment* observed features with an orthonormal basis from the complementary space, which is orthogonal to the row space of observed features. This formulation enables us to reformulate the problem within a conventional linear bandit framework, where the reward function is defined as a dot product of the augmented features and an unknown parameter, without any additional additive term.

However, since these augmented features are not identical to the unobserved features, estimation errors may arise from information loss. To mitigate such errors, we leverage (ii) a DR estimator, which is widely used in the statistical literature for its robustness to errors caused by missing data. These two approaches allow our algorithm to compensate for missing information due to partial observability, improving both estimation accuracy and adaptability to the environment.

Our main contributions are summarized as follows:

- We propose a linear bandit problem with partially observable features. Our problem setting is more general and challenging than those in the existing literature on linear bandits with latent features, which often rely on specific structural assumptions governing the relationship between observed and latent features. In contrast, our approach assumes no additional structure for the unobserved features beyond the linearity of the reward function, which is commonly adopted in the linear bandit literature (Section 3).
- We introduce a novel estimation strategy by (i) augmenting the features that maximally capture the effect of reward projected onto the observed features, while minimizing the impact of unobserved features (Section 4), and (ii) constructing a DR estimator that mitigates errors from unobserved features. By integrating augmented features with the DR estimator, we guarantee a $\sqrt{t}$-convergence rate on the rewards for *all* arms in each round t (Theorem 2).
- We propose an algorithm named *Robust to Latent Features* (RoLF) for a general linear bandit framework with latent features (Algorithm 1). The algorithm achieves a regret bound of $\widetilde{O}(\sqrt{(d+d_h)T})$[1] (Theorem 3), where d_h is the number of nonzero coefficients associated with the component of the reward projected onto the orthogonal complement of the row space of observed features (Section 4.2). RoLF requires no prior knowledge or modeling of unobserved features, yet achieves, to the best of our knowledge, a sharper regret bound than OLB and MAB algorithms, as well as existing methods accounting for partial observability or model misspecification within the linear bandit framework.
- Our numerical experiments confirm that our algorithm consistently outperforms OLB (Li et al., 2010; Agrawal & Goyal, 2013; Kim & Paik, 2019) and MAB (Lattimore & Szepesvári, 2020) algorithms, validating both its practicality and theoretical guarantees (Section 6).

[1] $\widetilde{O}(\cdot)$ is the Big-O notation omitting logarithmic factors.

2. Related Works

While our setting appears similar to prior works on bandit problems with (i) model misspecification and (ii) partial observability, it differs from both lines of research in several aspects, including the nature of the unobserved features and the strategies used to address the problem.

First, our problem setting is more general and challenging than misspecified linear bandit problems, where the reward model deviates from the true reward due to nonlinearity (Lattimore & Szepesvári, 2020) or additive deviation terms (Ghosh et al., 2017; Bogunovic et al., 2021; He et al., 2022). While prior works incorporate cumulative misspecification error into the regret bound, we obtain a regret bound that remains unaffected by such deviations. In particular, Ghosh et al. (2017) employ hypothesis testing for model selection and obtain a regret bound of $O(K\sqrt{T}\log T)$ under high misspecification. In contrast, our algorithm achieves a regret bound of $O(\sqrt{(d+d_h)T\log T})$ without such tests, while handling partial observability in a unified framework.

Second, in bandit problems with partially observable features, prior works often rely on structural assumptions. For example, Park & Faradonbeh (2022; 2024); Kim et al. (2023a) assume the true features are drawn from a Gaussian distribution, while Zeng et al. (2025) model them as evolving according to a linear dynamical system with additive Gaussian noise, where observed features are generated via a known linear mapping, also corrupted by Gaussian noise. These methods typically aim to recover the true features—via decoders (Park & Faradonbeh, 2022; 2024), Bayesian oracles (Kim et al., 2023a), or Kalman filtering (Zeng et al., 2025). In contrast, our approach makes no structural assumptions about the relationship between observed and unobserved features and does not attempt to recover the latter. Instead, we mitigate the information loss from partial observability by projecting the latent part of the reward using only the observed features. A more detailed and comprehensive literature review is provided in Appendix A.

3. Preliminaries

3.1. Notation

For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, 2, \ldots, n\}$. For a vector $\mathbf{v}$, we denote its L_1, L_2 and supremum norms by $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty$, respectively. The L_2-norm weighted by a positive definite matrix $\mathbf{D}$ is denoted by

$\|\mathbf{v}\|_{\mathbf{D}}$. For two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, the inner product is defined as the dot product, i.e., $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle := \mathbf{v}_1^\top \mathbf{v}_2$, and we use both notations interchangeably. For a matrix $\mathbf{M}$, its minimum and maximum eigenvalues are denoted by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$, respectively. We let $\mathrm{R}(\mathbf{M})$ denote the row space of $\mathbf{M}$, i.e., the subspace spanned by the rows of $\mathbf{M}$.

3.2. Problem Formulation

In this section, we outline our problem setting and introduce several key assumptions. Each arm $a \in [K]$ is associated with a true feature vector $\mathbf{z}_a \in \mathbb{R}^{d_z}$ that determines its rewards. However, the agent can observe only a subset of its elements, with the others remaining unobserved. Specifically, $\mathbf{z}_a$ is defined as follows:

$$\mathbf{z}_a := \left[x_a^{(1)}, \cdots, x_a^{(d)}, u_a^{(1)}, \cdots, u_a^{(d_u)} \right]^\top. \tag{1}$$

For clarity, we highlight the observed components in blue and the unobserved components in red. Note that the dimensions of the latent feature vector, $d_u = d_z - d$, and the true feature vector, d_z, are both *unknown* to the agent. As a result, the agent is unaware of how many features, if any, are unobserved, or even whether partial observability exists at all. This lack of structural information presents a fundamental challenge, as it prevents the agent from explicitly modeling or compensating for the unobserved components when selecting an appropriate learning strategy.

It is worth noting that the setting with fixed observed features[2] includes linear bandits with misspecification error (Ghosh et al., 2017; Bogunovic et al., 2021; He et al., 2022) as special cases. In Appendix D, we present a setting with varying observed features and an algorithm that achieves a $\sqrt{T}$-rate regret bound. Moreover, if latent features were allowed to change arbitrarily over time, the problem would become non-learnable and thus ill-posed. Consequently, assuming fixed features is both natural and well-justified (see Table 1 for comparisons).

The reward associated with arm a is defined as the dot product between its true feature vector $\mathbf{z}_a$ and an unknown parameter $\boldsymbol{\theta}_\star \in \mathbb{R}^{d_z}$, namely

$$y_{a,t} = \langle \mathbf{z}_a, \boldsymbol{\theta}_\star \rangle + \epsilon_t \qquad \forall\, a \in [K], \tag{2}$$

where $a_t \in [K]$ denotes the action selected by the agent at round t. Here, ϵ_t represents a random noise term that captures the inherent stochasticity in the reward generation process. We make the following assumption on ϵ_t, which is commonly adopted in bandit problems.

Assumption 1 (Sub-Gaussian noise)**.** *Let $\mathcal{F}_t$ denote the history at round t, represented by a filtration of σ-algebras, e.g., $\sigma(a_1, y_{a_1,1}, \ldots, a_{t-1}, y_{a_{t-1},t-1}, a_t)$. The reward noise ϵ_t is conditionally σ-sub-Gaussian, i.e., for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda \epsilon_t) \,|\, \mathcal{F}_{t-1}] \le \exp\left(\frac{\lambda^2 \sigma^2}{2} \right).$$

Since ϵ_t is sampled after a_t is observed, ϵ_t is $\mathcal{F}_t$-measurable. Under Assumption 1, it follows that $\mathbb{E}[\epsilon_t \,|\, \mathcal{F}_{t-1}] = 0$, thus $\mathbb{E}[y_{a,t} \,|\, \mathcal{F}_{t-1}] = \langle \mathbf{z}_a, \boldsymbol{\theta}_\star \rangle$. For brevity, we use $\mathbb{E}_{t-1}[\cdot]$ to denote $\mathbb{E}[\cdot \,|\, \mathcal{F}_{t-1}]$ henceforth. To eliminate issues of scale for analysis, we assume that the expected reward $|\langle \mathbf{z}_a, \boldsymbol{\theta}_\star \rangle| \le 1$ for all $a \in [K]$, and we do not assume any bound on the norm of $\mathbf{z}_a$ as well.

Table 1. Summary of problem settings covered in this paper and the corresponding results. Note that if latent features arbitrarily change over time, the problem itself would become non-learnable, making the problem ill-posed (see Appendix D for details).

Observed Features	Unobserved Features	Learnable?	Results
Fixed	Fixed	Yes	Theorem 3
Varying	Fixed	Yes	Theorem 6
Varying	Varying	No	-

Let $a_\star := \operatorname{argmax}_{a \in [K]} \langle \mathbf{z}_a, \boldsymbol{\theta}_\star \rangle$ denote the optimal action, considering both observed and latent features. Moreover, we denote by $\mathbf{z}_\star$ and $y_{\star,t}$ the true feature vector and the realized reward associated with the optimal action $a_\star$, respectively. We evaluate the theoretical performance of our algorithm using cumulative regret, which is defined as the expected sum of the differences between the reward of $a_\star$ and a_t:

$$\begin{aligned} \mathrm{Reg}(T) &= \sum_{t=1}^{T} \langle \boldsymbol{\theta}_\star, \mathbf{z}_\star - \mathbf{z}_{a_t} \rangle \\ &= \sum_{t=1}^{T} \left(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}] \right). \end{aligned}$$

Considering the composition of $\mathbf{z}_a$ defined in Eq. (1), we decompose the parameter vector as $\boldsymbol{\theta}_\star = [(\boldsymbol{\theta}_\star^{(o)})^\top, (\boldsymbol{\theta}_\star^{(u)})^\top]^\top$, where $\boldsymbol{\theta}_\star^{(o)} \in \mathbb{R}^d$ and $\boldsymbol{\theta}_\star^{(u)} \in \mathbb{R}^{d_u}$ correspond to the parameters associated with the observed and latent features, respectively. Adopting this decomposition of $\boldsymbol{\theta}_\star$, the reward $y_{a_t,t}$ defined in Eq. (2) can be decomposed into the following three terms:

$$y_{a_t,t} = \langle \mathbf{x}_{a_t}, \boldsymbol{\theta}_\star^{(o)} \rangle + \epsilon_t + \langle \mathbf{u}_{a_t}, \boldsymbol{\theta}_\star^{(u)} \rangle. \tag{3}$$

The last term of Eq. (3) corresponds to the inaccessible portion of the reward. This reward model is equivalent to that imposed in the linear bandits with misspecification error (Ghosh et al., 2017; Lattimore et al., 2020). While the regret bound in Lattimore et al. (2020) includes misspecification error that grows linearly with decision horizon, our proposed method (Section 4) addresses this misspecification error and achieves a regret bound that is sublinear in T.

[2] This assumption is standard in linear bandits with model misspecification (Ghosh et al., 2017; Lattimore et al., 2020), which is a special case of our partially observable feature setting.

Before presenting our method and algorithm, we first establish a lower bound on the regret of linear bandit algorithms that disregard the unobserved portion of the rewards. In particular, the following theorem provides lower bounds for two OLB algorithms: `OFUL` (Abbasi-yadkori et al., 2011) and `LinTS` (Agrawal & Goyal, 2013).

Theorem 1 (Regret lower bound of `OFUL` and `LinTS` ignoring latent features)**.** *Under partial observability, there exists a problem instance where both `OFUL` and `LinTS` suffer from cumulative expected regret that grows linearly with T due to their disregard for the unobserved components.*

Sketch of proof. Consider a linear bandit problem instance with an action set $\mathcal{A} := \{1, 2\}$, where $d = d_u = 1$, implying that $d_z = 2$. The true feature vectors are given by $\mathcal{Z} := \{[1, 3]^\top, [2, 19/4]^\top\} \subset \mathbb{R}^2$, with the agent observing only the first element of each vector, while the second element remains unobserved. In this setting, we take arm 2 to be the optimal action. However, an estimator relying solely on the observed features is inconsistent, leading `OFUL` and `LinTS` to select the suboptimal arm with probability $\Theta(1)$, and thereby incur regret that grows linearly in T.

Theorem 1 shows that neglecting the latent portion of the reward in decision-making may cause the agent to fail in identifying the optimal action. The comprehensive proof is deferred to Appendix E.1. While Theorem 1 specifically considers `OFUL` and `LinTS`—which are known to achieve the most efficient regret bounds for UCB- and Thompson sampling-based policies in the linear bandit framework—we additionally present an algorithm-agnostic lower bound based on a different analysis (see Appendix F for details).

4. Robust Estimation for Partially Observable Features

4.1. Feature Vector Augmentation with Orthogonal Projection

In the linear bandit framework, accurate estimation of the unknown reward parameter $\boldsymbol{\theta}_\star$ contributes to optimal decision-making. However, in our problem setting, the learner lacks access to the latent reward component—namely, the last term in Eq. (3). Consequently, without an appropriate mechanism to compensate for the absence of $\boldsymbol{\theta}_\star^{(u)}$, the agent fails to accurately estimate $\boldsymbol{\theta}_\star$, resulting in ineffective learning, as demonstrated in Theorem 1.

That said, minimizing regret—that is, selecting the optimal arm—does not require recovery of the true reward parameter $\boldsymbol{\theta}_\star$ entirely; rather, it suffices to estimate the K expected rewards $\{\mathbf{z}_a^\top \boldsymbol{\theta}_\star : a \in [K]\}$. A straightforward approach is to ignore the features altogether and apply MAB algorithms such as `UCB1` (Auer et al., 2002), which are known to achieve a regret bound of $\widetilde{O}(\sqrt{KT})$. However, these algorithms tend to suffer higher regret than feature-based algorithms, particularly when the number of arms is significantly larger than the dimension of the feature vectors, i.e., $K \gg d_z$.

To tackle this dilemma, we propose a unified approach that handles partial observability and ensures efficient estimation. Let us define $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_K) \in \mathbb{R}^{d \times K}$ as the matrix formed by concatenating the observed part of the true features, and $\mathbf{U} := (\mathbf{u}_1^{(u)}, \ldots, \mathbf{u}_K^{(u)}) \in \mathbb{R}^{d_u \times K}$ as the matrix formed by concatenating their latent complements for each arm. Without loss of generality, we assume a set of K vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ spans $\mathbb{R}^d$.[3] We denote by $\mathbf{P}_\mathbf{X} := \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}$ the projection matrix onto the row space of $\mathbf{X}$, denoted by $\mathrm{R}(\mathbf{X})$. Then the vector of rewards for all arms, $\mathbf{Y}_t = (y_{1,t}, \ldots, y_{K,t})^\top$, is decomposed as:

$$\begin{aligned}
\mathbf{Y}_t &= \left(\mathbf{X}^\top \boldsymbol{\theta}_\star^{(o)} + \mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)}\right) + \epsilon_t \mathbf{1}_K \\
&= \mathbf{P}_\mathbf{X}\left(\mathbf{X}^\top \boldsymbol{\theta}_\star^{(o)} + \mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)}\right) \\
&\quad + (\mathbf{I}_K - \mathbf{P}_\mathbf{X})\left(\mathbf{X}^\top \boldsymbol{\theta}_\star^{(o)} + \mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)}\right) + \epsilon_t \mathbf{1}_K \\
&= \mathbf{X}^\top\left(\boldsymbol{\theta}_\star^{(o)} + (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)}\right) \\
&\quad + (\mathbf{I}_K - \mathbf{P}_\mathbf{X})\mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)} + \epsilon_t \mathbf{1}_K, \qquad (4)
\end{aligned}$$

where the first term corresponds to the reward projected onto $\mathrm{R}(\mathbf{X})$, whereas the second term is the reward projected onto $\mathrm{R}(\mathbf{X})^\perp$, the subspace of $\mathbb{R}^K$ orthogonal to $\mathrm{R}(\mathbf{X})$. We reparametrize $\boldsymbol{\theta}_\star^{(o)} + (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)}$ as $\boldsymbol{\mu}_\star^{(o)}$, representing the parameter associated with the observed features.

To handle the second term in Eq. (4), we consider a set of row vector bases $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\} \in \mathrm{R}(\mathbf{X})^\perp$, where $\mathbf{b}_i \in \mathbb{R}^K$ for $i \in [K - d]$. Given the set, there exist coefficients $\mu_{\star,1}^{(u)}, \ldots, \mu_{\star,K-d}^{(u)} \in \mathbb{R}$ such that the term can be expressed as a linear combination:

$$(\mathbf{I}_K - \mathbf{P}_\mathbf{X})\mathbf{U}^\top \boldsymbol{\theta}_\star^{(u)} = \sum_{i=1}^{K-d} \mu_{\star,i}^{(u)} \mathbf{b}_i. \qquad (5)$$

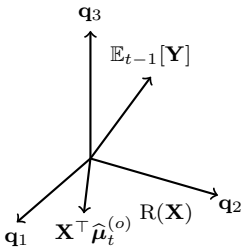
We define the number of nonzero coefficients as:

$$d_h(\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top) := |\{i \in [K - d] : \mu_{\star,i}^{(u)} \neq 0\}|. \qquad (6)$$

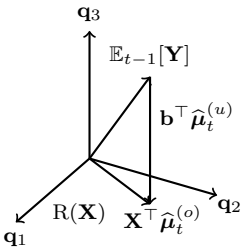
Note that $d_h = 0$ for any basis set $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\}$ when the latent feature space is completely included in the observed feature space, i.e., $\mathrm{R}(\mathbf{U}) \subseteq \mathrm{R}(\mathbf{X})$. In this case, $(\mathbf{I}_K - \mathbf{P}_\mathbf{X})\mathbf{U}^\top = \mathbf{0}_{K \times d_u}$, which means that the projected rewards onto $\mathrm{R}(\mathbf{X})^\perp$ can be linearly expressed by the observed features. If $\mathrm{R}(\mathbf{U}) \supseteq \mathrm{R}(\mathbf{X})$, on the other hand, then $d_h = K - d$ for any basis set $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\}$.

[3] When $d > K$, we can apply singular value decomposition (SVD) on $\mathbf{X}$ to reduce the feature dimension to $\bar{d} \leq K$ with $\mathrm{R}(\mathbf{X}) = \bar{d}$.

Figure 1. Illustration comparing OLB algorithms and our approach in estimating rewards of $K = 3$ arms. OLB algorithms find estimates within $\mathrm{R}(\mathbf{X})$ thus accumulating errors from unobserved features. However, our approach utilizes the projection of the latent reward, $\mathbf{b}^\top \widehat{\boldsymbol{\mu}}_t^{(u)}$, onto the orthogonal complement of $\mathrm{R}(\mathbf{X})$, thereby enabling reward estimation in $\mathbb{R}^K$.



(a) OLB algorithms: Estimation confined to $\mathrm{R}(\mathbf{X})$



(b) Ours: Projection onto orthogonal complement

In other cases, the quantity d_h depends on the choice of the basis $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\}$, and tends to be smaller when a larger portion of $\mathrm{R}(\mathbf{U})$ is included within $\mathrm{R}(\mathbf{X})$. For any choice of the basis, our algorithm achieves $\widetilde{O}(\sqrt{(d+d_h)T})$ regret without prior knowledge of d_h, which does not exceed the $\widetilde{O}(\sqrt{KT})$ regret bound achieved by MAB algorithms ignoring features. Further details are provided in Section 5.2.

Similar to the reparametrization of $\boldsymbol{\mu}_\star^{(o)}$, we denote $\boldsymbol{\mu}_\star^{(u)} := [\mu_{\star,1}^{(u)}, \ldots, \mu_{\star,K-d}^{(u)}]^\top$. By concatenating $\boldsymbol{\mu}_\star^{(o)}$ and $\boldsymbol{\mu}_\star^{(u)}$, we define $\boldsymbol{\mu}_\star := [(\boldsymbol{\mu}_\star^{(o)})^\top, (\boldsymbol{\mu}_\star^{(u)})^\top]^\top \in \mathbb{R}^K$, so that the reward for each $a \in [K]$ can be rewritten as follows:

$$\begin{aligned} y_{a,t} &= \mathbf{e}_a^\top \mathbf{Y} \\ &= \mathbf{e}_a^\top [\mathbf{X}^\top \ \mathbf{b}_1 \cdots \mathbf{b}_{K-d}] \boldsymbol{\mu}_\star + \epsilon_t \\ &= [\mathbf{x}_a \ \mathbf{e}_a^\top \mathbf{b}_1 \cdots \mathbf{e}_a^\top \mathbf{b}_{K-d}] \boldsymbol{\mu}_\star + \epsilon_t. \end{aligned} \tag{7}$$

The decomposition in Eq. (7) implies that Eq. (4) takes the form $\mathbf{Y}_t = [\mathbf{X}^\top \ \mathbf{b}_1 \cdots \mathbf{b}_{K-d}]\boldsymbol{\mu}_\star + \epsilon_t \mathbf{1}_K$. Note that $\mathbf{e}_a \in \mathbb{R}^K$ denotes the a-th standard basis vector. With this modification, the rewards are now represented as a linear function of the augmented feature vectors: $\widetilde{\mathbf{x}}_a := [\mathbf{x}_a^\top \ \mathbf{e}_a^\top \mathbf{b}_1 \cdots \mathbf{e}_a^\top \mathbf{b}_{K-d}]^\top \in \mathbb{R}^K$, *without any misspecification error*. A toy example illustrating our strategy is shown in Figure 1.

In terms of regret minimization, while `SupLinUCB` (Chu et al., 2011) achieves a regret bound of $\widetilde{O}(\sqrt{dT})$, it is often considered impractical as it computes $\log T$ distinct batches and estimators, requiring knowledge of T and, more critically, discards a significant portion of samples at each parameter update. To retain the theoretical efficiency while improving practicality, we adopt the doubly robust (DR) estimation framework, which is known to achieve $\widetilde{O}(\sqrt{dT})$ (Kim et al., 2021; 2023c). Given that $\widetilde{\mathbf{x}}_a \in \mathbb{R}^K$, we propose an efficient algorithm that employs a DR estimator with ridge regularization and obtains a regret bound of $\widetilde{O}(\sqrt{KT})$ (see Appendix C). However, when $K > d$ and $d_u = 0$, the regret is the regret is higher than that of OLB algorithms. To address this challenge, we propose a novel estimation strategy in the following section that eliminates the dependence on K in the regret bound.

4.2. Doubly Robust Lasso Estimator

In Eq. (7), the parameter $\boldsymbol{\mu}_\star$ is sparse, with its sparsity depending on the number of nonzero elements required to represent the inaccessible portion of the reward projected onto $\mathrm{R}(\mathbf{X})^\perp$, namely d_h defined in Eq. (6). Recall from Eq. (5) that this projection can be expressed using at most d_h basis vectors, implying that $\boldsymbol{\mu}_\star^{(u)}$ has at most d_h nonzero entries.

Let $\check{\boldsymbol{\mu}}_t^L$ denote the Lasso estimator of $\boldsymbol{\mu}_\star$ using the augmented feature vectors, defined as follows:

$$\begin{aligned} \check{\boldsymbol{\mu}}_t^L := \operatorname*{argmin}_{\boldsymbol{\mu}} &\sum_{\tau=1}^{t} (y_{a_\tau,\tau} - \widetilde{\mathbf{x}}_{a_\tau}^\top \boldsymbol{\mu})^2 \\ &+ 2\widetilde{\sigma}_{\max}\sigma \sqrt{2pt \log \frac{2Kt^2}{\delta}} \|\boldsymbol{\mu}\|_1, \end{aligned} \tag{8}$$

where p is the coupling probability used to define the multinomial distribution for pseudo-action sampling (as defined in Eq. (9)), and δ is the confidence parameter of the algorithm. Note that $\widetilde{\sigma}_{\max}^2 := \max_{a\in[K]} \mathbf{e}_a^\top (\sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top) \mathbf{e}_a$, i.e., the largest diagonal entry of the Gram matrix constructed from the augmented feature vectors $\widetilde{\mathbf{x}}_a$ over $\mathcal{A}$.

For the estimator in Eq. (8) to correctly identify the zero entries in $\boldsymbol{\mu}_\star$, the compatibility condition needs to be satisfied (van de Geer & Bühlmann, 2009). Although the compatibility condition does not, in general, require a positive minimum eigenvalue for the Gram matrix, in our setting, $\lambda_{\min}(t^{-1} \sum_{s=1}^t \widetilde{\mathbf{x}}_{a_s} \widetilde{\mathbf{x}}_{a_s}^\top) > 0$. Therefore, the compatibility condition is implied without any additional assumption. However, ensuring a sufficiently large minimum eigenvalue typically requires collecting a large number of exploration samples, which in turn increases regret. Achieving this with fewer exploration samples remains a key challenge in the bandit literature, as the minimum eigenvalue affects the convergence rate of the estimator and, consequently, the regret bound (Soare et al., 2014; Kim et al., 2021).

We introduce a doubly robust (DR) estimator that employs the Gram matrix constructed from the augmented feature vectors over the entire action space, $\sum_{s=1}^t \sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top$, rather than only from the chosen actions, $\sum_{s=1}^t \widetilde{\mathbf{x}}_{a_s} \widetilde{\mathbf{x}}_{a_s}^\top$. Originating in the missing-data literature (Bang & Robins, 2005), DR estimation yields consistent estimators if either the imputation or the observation probability model is correct (Kim et al., 2021). In the bandit setting, only the reward of the chosen arm is observed for each decision round, leaving the other $K-1$ as *missing*. Hence, DR estimation imputes these $K-1$ missing rewards and incorporates all the

associated feature vectors into the estimation process. Since the observation probabilities are determined by the policy—which is known to the learner—the DR estimator remains robust to errors in the reward estimation. While Kim & Paik (2019) proposed a DR Lasso estimator under IID features satisfying the compatibility condition, we propose a novel DR Lasso estimator that does not rely on such assumptions.

We improve the DR estimation by incorporating resampling and coupling methods. For each t, let $\mathcal{E}_t \subseteq [t]$ denote the set of exploration rounds such that for any $\tau \in \mathcal{E}_t$, the action a_τ is sampled uniformly over $[K]$. The set $\mathcal{E}_t$ is constructed recursively: starting with $\mathcal{E}_0 = \emptyset$, we define

$$\mathcal{E}_t = \begin{cases} \mathcal{E}_{t-1} \cup \{t\} & \text{if} \quad |\mathcal{E}_t| \le C_e \log \dfrac{2Kt^2}{\delta}, \\ \mathcal{E}_{t-1} & \text{otherwise.} \end{cases}$$

Here, C_e is defined as $(8K)^3 \widetilde{\sigma}_{\min}^{-2} \widetilde{\sigma}_{\max}^2 (1-p)^{-2}$, where $\widetilde{\sigma}_{\min}^2 := \lambda_{\min}(\sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top)$ denotes the minimum eigenvalue of the Gram matrix constructed from the augmented feature vectors over the *entire* action space.

At each round $t \notin \mathcal{E}_t$, the algorithm selects a_t according to an ϵ_t-greedy strategy: with probability $\epsilon_t = t^{-1/2}$, it selects $\widehat{a}_t = \operatorname{argmax}_{a\in[K]} \widetilde{\mathbf{x}}_a^\top \widehat{\boldsymbol{\mu}}_{t-1}^L$; with $1-\epsilon_t$, it selects an action uniformly at random from $[K] \setminus \{\widehat{a}_t\}$. Given a_t, a *pseudo-action* $\widetilde{a}_t$ is sampled from a multinomial distribution:

$$\begin{aligned} \phi_{a_t,t} &:= \mathbb{P}(\widetilde{a}_t = a_t \mid a_t) = p, \\ \phi_{k,t} &:= \mathbb{P}(\widetilde{a}_t = k \mid a_t) = \frac{1-p}{K-1}, \end{aligned} \tag{9}$$

for all $k \in [K] \setminus \{a_t\}$, where $p \in (1/2, 1)$ is the coupling probability specified by the algorithm. To couple the policies for the actual action a_t and the pseudo-action $\widetilde{a}_t$, we repeatedly resample both of them until they match.

With the pseudo-actions coupled with the actual actions, we construct unbiased pseudo-rewards for all $a \in [K]$ as:

$$\widetilde{y}_{a,t} := \widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t^L + \frac{\mathbb{I}(\widetilde{a}_t = a)}{\phi_{a,t}} \big(y_{a,t} - \widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t^L\big), \tag{10}$$

where $\check{\boldsymbol{\mu}}_t^L$ is the imputation estimator that fills in the missing rewards of unselected arms at round t, as defined in Eq. (8).

As shown in Eq. (10), the pseudo-reward includes an inverse probability term. Hence, when DR estimation is applied under the ϵ_t-greedy policy with $\epsilon_t = t^{-1/2}$, its variance can grow unbounded over time. To mitigate this issue, we propose a resampling and coupling strategy: by coupling the ϵ_t-greedy policy with the multinomial distribution (Eq. (9)), we ensure that each inverse probability weight $\phi_{a,t}^{-1}$, for $a \in [K]$, remains bounded by $O(K)$. This strategy effectively stabilizes the variance of the pseudo-rewards.

Moreover, one can show that the resampling succeeds with high probability for each round. Let $\mathcal{M}_t$ denote the event that the pseudo-action $\widetilde{a}_t$ matches the chosen action a_t within a specified number of resamples. For a given $\delta' \in (0,1)$, we set the maximum number of resamples to $\rho_t := \log((t+1)^2/\delta')/\log(1/(1-p))$; then $\mathcal{M}_t$ occurs with probability at least $1 - \delta'/(t+1)^2$. Resampling allows the algorithm to further explore the action space to balance regret minimization with accurate reward estimation.

For $a \neq \widetilde{a}_t$, i.e., an arm that is *not* selected in the round t, we impute the missing rewards using $\widetilde{\mathbf{x}}_k^\top \check{\boldsymbol{\mu}}_t^L$. For $a = \widetilde{a}_t$, in contrast, the term $\mathbb{I}(\widetilde{a}_t = a) y_{a,t}/\phi_{a,t}$ calibrates the predicted reward to ensure the unbiasedness of the pseudo-rewards for all arms. Given that $\mathbb{E}_{\widetilde{a}_t}[\mathbb{I}(\widetilde{a}_t = a)] = \mathbb{P}(\widetilde{a}_t = a) = \phi_{a,t}$, taking the expectation over $\widetilde{a}_t$ on both sides of Eq. (10) gives $\mathbb{E}_{\widetilde{a}_t}[\widetilde{y}_{a,t}] = \mathbb{E}_{t-1}[y_{a,t}] = \widetilde{\mathbf{x}}_a^\top \boldsymbol{\mu}_\star$ for all $a \in [K]$. Although the estimate $\widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t$ may have a large error, it is multiplied by the mean-zero random variable $(1 - \mathbb{I}(\widetilde{a}_t = a)/\phi_{a,t})$, which makes the resulting pseudo-rewards defined in Eq. (10) robust to errors of $\widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t$. The pseudo-rewards can only be computed—and thus can only be used—when the pseudo-action $\widetilde{a}_t$ matches the actual action a_t, that is, when the event $\mathcal{M}_t$ occurs. Since $\mathcal{M}_t$ occurs with high probability, we are able to compute pseudo-rewards for almost all rounds.

Incorporating the pseudo-rewards, $\widetilde{y}_{a,t}$, into estimation, we define our DR Lasso estimator as follows:

$$\begin{aligned} \widehat{\boldsymbol{\mu}}_t^L := \operatorname*{argmin}_{\boldsymbol{\mu}} & \sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a\in[K]} \big(\widetilde{y}_{a,\tau} - \widetilde{\mathbf{x}}_a^\top \boldsymbol{\mu}\big)^2 \\ & + \frac{4\sigma\widetilde{\sigma}_{\max}}{p} \sqrt{2t \log \frac{2Kt^2}{\delta}} \|\boldsymbol{\mu}\|_1, \end{aligned} \tag{11}$$

and note that our estimator uses the Gram matrix aggregated over all actions at each round, $\sum_{\tau=1}^t \sum_{a=1}^K \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top$, instead of the one built only from the chosen actions, $\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau} \widetilde{\mathbf{x}}_{a_\tau}^\top$. The following theorem guarantees convergence of the estimator, across all arms, to the true parameter, after a sufficient number of exploration rounds.

Theorem 2 (Consistency of the DR Lasso estimator)**.** *Let d_h denote the dimension of the projected latent rewards defined in Eq. (6). Then with probability at least $1 - 3\delta$,*

$$\max_{a\in[K]} |\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star)| \le \frac{8\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}} \sqrt{\frac{2(d+d_h)}{t} \log \frac{2Kt^2}{\delta}},$$

for all rounds t such that $t \ge |\mathcal{E}_t|$.

Although the DR Lasso estimator leverages K-dimensional feature vectors, its error bound depends only logarithmically on K. Such fast convergence is typically achieved under classical regularity conditions, including the compatibility condition and the restrictive minimum eigenvalue condition (van de Geer & Bühlmann, 2009; Bühlmann & van de Geer, 2011). Existing Lasso-based bandit approaches (Kim & Paik, 2019; Bastani & Bayati, 2020; Oh

Algorithm 1 Robust to Latent Feature (RoLF)

INPUT: Observed features $\{\mathbf{x}_a : a \in [K]\}$, coupling probability $p \in (1/2, 1)$, confidence parameter $\delta > 0$.
Initialize $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}_K$, exploration phase $\mathcal{E}_0 = \emptyset$, and exploration factor $C_{\mathrm{e}} := (8K)^3 \widetilde{\sigma}_{\min}^{-2} \widetilde{\sigma}_{\max}^2 (1-p)^{-2}$.
Find orthogonal basis $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\} \subseteq \mathrm{R}(\mathbf{X})^\perp$ and construct $\{\widetilde{\mathbf{x}}_a : a \in [K]\}$.
for $t = 1, \ldots, T$ **do**
  if $|\mathcal{E}_t| \leq C_{\mathrm{e}} \log(2Kt^2/\delta)$ **then**
    Randomly sample $\widehat{a}_t$ uniformly over $[K]$ and $\mathcal{E}_t = \mathcal{E}_{t-1} \cup \{t\}$.
  else
    Compute $\widehat{a}_t := \arg\max_{a \in [K]} \widetilde{\mathbf{x}}_a^\top \widehat{\boldsymbol{\mu}}_{t-1}^L$.
  end if
  while $\widetilde{a}_t \neq a_t$ and count $\leq \rho_t$ **do**
    Sample a_t with $\mathbb{P}(a_t = \widehat{a}_t) = 1 - (t^{-1/2})$ and $\mathbb{P}(a_t = k) = t^{-1/2}/(K-1)$, $\forall k \neq \widehat{a}_t$.
    Sample $\widetilde{a}_t$ according to Eq. (9).
    count = count + 1.
  end while
  Play a_t and observe $y_{a_t,t}$.
  if $\widetilde{a}_t \neq a_t$ **then**
    Set $\widehat{\boldsymbol{\mu}}_t^L := \widehat{\boldsymbol{\mu}}_{t-1}^L$.
  else
    Update $\widehat{\boldsymbol{\mu}}_t^L$ following Eq. (11) with $\widetilde{y}_{a,t}$ and update $\check{\boldsymbol{\mu}}_t^L$ following Eq. (8).
  end if
end for

et al., 2021; Ariu et al., 2022; Chakraborty et al., 2023; Lee et al., 2025) generally impose these conditions directly on the feature vectors. Our approach, on the other hand, does not require such assumptions, as the augmented features are orthogonal vectors lying in $\mathrm{R}(\mathbf{X})^\perp$. Moreover, their average Gram matrix satisfies $\lambda_{\min}(\sum_{a \in [K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top) \geq \min\{1, \lambda_{\min}(\sum_{a \in [K]} \mathbf{x}_a \mathbf{x}_a^\top)\}$. Thus, the convergence rate scales only as $\sqrt{\log K}$ with respect to K.

The consistency is proved by bounding the two components of the error in the pseudo-rewards (Eq. (10)): (i) the noise of the reward and (ii) the error of the imputation estimator $\check{\boldsymbol{\mu}}_t$. Since (i) is sub-Gaussian, it can be bounded using martingale concentration inequalities. For (ii), the imputation error $\widetilde{\mathbf{x}}_a^\top (\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star)$ is multiplied by the mean-zero random variable $(1 - \mathbb{I}(\widetilde{a}_t = a)/\phi_{a,t})$ and thus the magnitude of the error can be bounded by $\|\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\|_1 / \sqrt{t}$. The detailed proof is deferred to Appendix E.2.

5. Proposed Algorithm and Regret Analysis

5.1. Robust to Latent Features (RoLF) Algorithm

In this section, we present our algorithm RoLF in Algorithm 1. In the initialization step, given the observed features, our algorithm constructs a set of orthogonal basis vectors $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\} \subseteq \mathrm{R}(\mathbf{X})^\perp$ to augment each observed feature vector. The algorithm first selects a candidate action $\widehat{a}_t$—either uniformly at random or greedily based on the estimated reward—and then resamples a_t and $\widetilde{a}_t$ until they match or the maximum number of trials allowed for each round, $\rho_t := \log((t+1)^2/\delta')/\log(1/(1-p))$, is reached. Once the resampling phase ends, the algorithm selects a_t and the corresponding reward $y_{a_t,t}$ is observed. If a_t and $\widetilde{a}_t$ match within ρ_t, then both the imputation and the main estimators are updated; otherwise, neither is updated.

Table 2. An overview of regret bound range of our algorithm, RoLF, depending on $d_h \in [0, K-d]$, the number of nonzero elements to effectively capture the unobserved part of reward projected onto $R(\mathbf{X})^\perp$. Note that $\widetilde{O}$ denotes the big-O notation omitting logarithm factors.

Feature space	Regret bound
span(observed) $\supseteq$ span(latent)	$\widetilde{O}(\sqrt{dT})$
span(observed) $\subseteq$ span(latent)	$\widetilde{O}(\sqrt{KT})$
otherwise	$\widetilde{O}(\sqrt{(d+d_h)T})$

Our proposed algorithm does not require knowledge of the dimension of the unobserved features d_u, nor the dimension of the reward component projected onto $\mathrm{R}(\mathbf{X})^\perp$. Although we present the algorithm for fixed feature vectors, the algorithm is also applicable to time-varying feature vectors. In such cases, we estimate the bias caused by unobserved features by augmenting the standard basis. For further details, refer to Appendix D.

5.2. Regret Analysis

Theorem 3 (Regret bound for Lasso RoLF)**.** *Let d_h denote the number of nonzero coefficients in the representation of the projected latent reward as defined in Eq. (6). Then for any $\delta \in (0,1)$ and $p \in (1/2, 1)$, with probability at least $1 - 6\delta$, the cumulative regret of RoLF is bounded by*

$$\begin{aligned}\mathrm{Reg}(T) \leq\ & 2 \cdot 8^3 K^3 (1-p)^{-2} \log \frac{2KT^2}{\delta} \\ & + \frac{4\sqrt{T}}{K-1} + 2\sqrt{2T \log \frac{2}{\delta}} + 4\delta \\ & + \frac{32\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}} \sqrt{2(d+d_h)T \log \frac{2KT^2}{\delta}}.\end{aligned}$$

To the best of our knowledge, Theorem 3 provides the first regret bound that converges faster than $\widetilde{O}(\sqrt{KT})$, specifically for algorithms that account for unobserved features without relying on any structural assumptions. Assuming $\|\widetilde{\mathbf{x}}_a\|_\infty \leq 1$ (instead of $\|\widetilde{\mathbf{x}}_a\|_2 \leq 1$), $\widetilde{\sigma}_{\min}^2$ and $\widetilde{\sigma}_{\max}^2$ are constant independent of d or K. Note that the number of rounds required for the exploration phase is $O(K^3 \log KT)$, which grows only logarithmically with the time horizon T. The factor K^3 is irreducible, as the algorithm needs information

about all K biases from the missing features. By employing the Gram matrix of the *augmented* feature vectors over the action space $\mathcal{A}$, $\sum_{a=1}^{K} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top$, when combined with DR estimation, we reduce the required exploration phase time from $O(K^4 \log KT)$ to $O(K^3 \log KT)$, thereby improving the sample complexity by a factor of K.

As demonstrated in Theorem 2, the last term on the right-hand side of the regret bound is proportional to $\sqrt{d + d_h}$ rather than $\sqrt{K}$, resulting in an overall regret bound of $O(\sqrt{(d + d_h)T \log KT})$. The specific value of d_h is determined by the relationship between $\mathrm{R}(\mathbf{X})$ and $\mathrm{R}(\mathbf{U})$, as discussed in Section 4.1. Table 2 summarizes how the regret bound varies under different relationships between these subspaces. The formal proof of Theorem 3 is deferred to Appendix E.3.

6. Numerical Experiments

6.1. Experimental Setup

In this experiment, we simulate and compare our algorithms, Algorithm 1 and Algorithm 2 (Appendix C), with baseline OLB algorithms: `LinUCB` (Li et al., 2010) and `LinTS` (Agrawal & Goyal, 2013). These algorithms adopt UCB and Thompson sampling strategies, respectively, assuming a linear reward model based on observed features. We also include `DRLasso` (Kim & Paik, 2019) since our algorithm employs DR estimation with a Lasso estimator, and `UCB(`δ`)` (Lattimore & Szepesvári, 2020), an MAB algorithm that ignores features, to assess the effectiveness of using feature information, even under partial observability.

To provide comprehensive results, we consider two scenarios: one with partial observability and one without. For each scenario, we conduct experiments under three cases, classified by the relationship between the row spaces of the observed and unobserved features, $\mathrm{R}(\mathbf{X})$ and $\mathrm{R}(\mathbf{U})$, which determines the value of d_h: (i) Case 1, the general case where neither $\mathrm{R}(\mathbf{X})$ nor $\mathrm{R}(\mathbf{U})$ fully contains the other; (ii) Case 2, $\mathrm{R}(\mathbf{U}) \subseteq \mathrm{R}(\mathbf{X})$, where the row space spanned by the unobserved features lies entirely within that spanned by the observed features, thus $d_h = 0$; (iii) Case 3, $\mathrm{R}(\mathbf{X}) \subseteq \mathrm{R}(\mathbf{U})$, where the row space spanned by the observed feature space is fully contained within that spanned by the unobserved features, implying $d_h = K - d$. Note that in Scenario 2, Case 3 is excluded because $\mathrm{R}(\mathbf{U}) = \emptyset$ implies $\mathrm{R}(\mathbf{X}) = \emptyset$, which violates our problem setup.

In the simulation, after constructing the true features $\mathbf{z}_a \in \mathbb{R}^{d_z}$ and observed features $\mathbf{x}_a \in \mathbb{R}^d$, we apply singular value decomposition (SVD) to the observed feature matrix $\mathbf{X}$ to derive orthogonal basis vectors $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\}$ orthogonal to $\mathrm{R}(\mathbf{X})$. These vectors are then concatenated with $\mathbf{X}$ to form the augmented feature matrix. The reward parameter $\boldsymbol{\theta}_\star \in \mathbb{R}^{d_z}$ is sampled from $\mathrm{Unif}(-1/2, 1/2)$, and the rewards are generated following Eq. (2). For the hyperparamters in our algorithms, the coupling probability p and the confidence parameter δ, are set to 0.6 and 10^{-4}, respectively. The total decision horizon is $T = 1200$. Throughout the experiments, we fix the number of arms at $K = 30$ and the dimension of the true features at $d_z = 35$, ensuring $d_z \geq d$ to cover both scenarios. Further details on the experimental setup are provided in Appendix B.

6.2. Experimental Results

6.2.1. Scenario 1

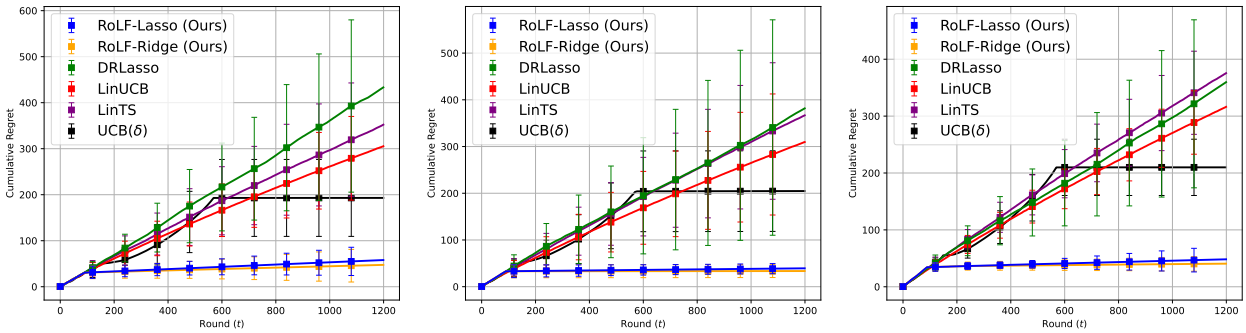
In this scenario, d is set to $\lfloor d_z/2 \rfloor$, so that the agent observes only about half of the full feature dimension. From Figure 2, we observe that the baseline OLB algorithms—`LinUCB`, `LinTS`, and `DRLasso`—perform worse than our algorithms in terms of both the level and robustness of cumulative regret. This pattern is consistently observed across all three cases, implying that linear bandit algorithms fail to identify the optimal arm in the presence of unobserved components, as they cannot capture the portion of the reward associated with the unobserved features. However, our algorithms show almost the same performance regardless of the relationship between $\mathrm{R}(\mathbf{X})$ and $\mathrm{R}(\mathbf{U})$, indicating the robustness to variations in feature observability structure.

Note that for all cases our algorithms—`RoLF-Lasso` (Algorithm 1) and `RoLF-Ridge` (Algorithm 2)—exhibit a sharp decline in the growth rate of cumulative regret after a certain number of rounds. This behavior is primarily due to the forced-exploration phase built into the algorithms, which ensures sufficient coverage of the action space in the early stages. Following this phase, the resampling and coupling strategies further enhance the efficiency of the DR estimation, leading to slower regret accumulation over time. This pattern is commonly shown from experimental results of other bandit algorithms employing the forced-exploration strategy (Goldenshluger & Zeevi, 2013; Hao et al., 2020; Chakraborty et al., 2023).

Furthermore, in Case 2 (Figure 2(b)), the cumulative regret grows more slowly than in other cases. This behavior is explained by the relationship $\mathrm{R}(\mathbf{U}) \subseteq \mathrm{R}(\mathbf{X})$, implying the reward components projected onto $\mathrm{R}(\mathbf{X})^\perp$ can be fully expressed by the observed features. As a result, the augmented features fully capture the underlying reward structure, enabling faster convergence relative to the other scenarios.

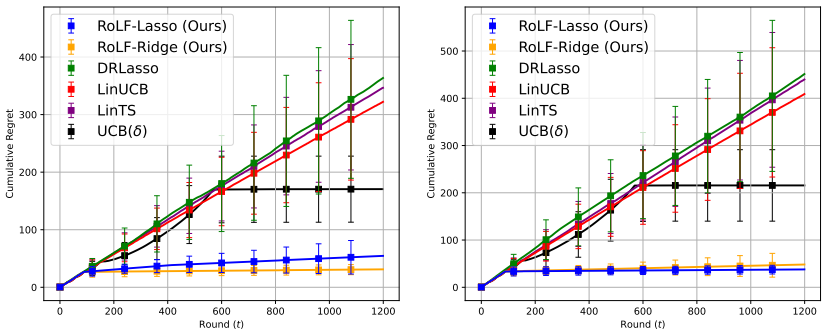
6.2.2. Scenario 2

In Scenario 2, we set $d = d_z = 2K$, implying that no latent features remain and allowing us to evaluate our algorithms under the condition $d > K$. As discussed in Section 4.1, we apply SVD to reduce the dimensionality of the observed features before constructing the augmented feature set. Figure 3 shows that our algorithms perform well even without



(a) Case 1 (General Case) (b) Case 2 ($\mathrm{R}(\mathbf{U}) \subseteq \mathrm{R}(\mathbf{X})$) (c) Case 3 ($\mathrm{R}(\mathbf{X}) \subseteq \mathrm{R}(\mathbf{U})$)

Figure 2. Cumulative regrets of the algorithms with partial observability (Scenario 1).



(a) Case 1 (General Case) (b) Case 2 ($\mathrm{R}(\mathbf{U}) \subseteq \mathrm{R}(\mathbf{X})$)

Figure 3. Cumulative regrets of the algorithms without partial observability (Scenario 2).

unobserved features, both in terms of the level and robustness of the cumulative regret. Moreover, by incorporating dimensionality reduction, our algorithms remain effective even when the feature matrix is not full rank. Lastly, similar to Figure 2, cumulative regret in Case 2 converges slightly more slowly than in Case 1.

Meanwhile, the baseline OLB algorithms continue to struggle in identifying the optimal arm throughout the horizon. For `LinUCB` and `LinTS`, this behavior may be attributed to the curse of dimensionality, where regret scales linearly with the feature dimension d—a well-known limitation of linear bandit algorithms (Oswal et al., 2020; Tran et al., 2024). Additionally, since the true rewards are bounded by 1, the resulting small reward gaps between arms may further hinder the identification of the optimal arm. In the case of `DRLasso`, similar behavior arises from the fixed feature setting: the algorithm uses an "averaged" context vector across the action space at each round for parameter estimation. Consequently, under the fixed-feature setting, this strategy becomes effectively equivalent to using a single fixed vector throughout the entire learning horizon, thereby limiting the expressiveness of the estimation.

7. Conclusion

In this work, we addressed the problem of partially observable features within the linear bandit framework. We showed that conventional algorithms that ignore unobserved features may suffer linear regret due to information loss, and introduced `RoLF`, a novel algorithm that accounts for latent features using only observed data without requiring prior knowledge. Our algorithm achieves a tighter regret bound than existing methods, and this improvement is supported by our numerical experiments.

For future work, from the perspective that our feature augmentation strategy reformulates the problem as another linear bandit problem without model misspecification, extending this approach to other reward models, e.g., generalized linear models, would be an interesting direction. Additionally, although we assumed the latent reward component to be linear in the unobserved features, we can relax this assumption by viewing the latent reward component as an exogenous factor that interferes with the learning process. This perspective allows for modeling the latent reward using a general function class without structural assumptions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2022-NR071853; RS-2023-00222663; RS-2025-16070886), by a grant of Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2024-00512375), by AI-Bio Research Grant through Seoul National University, and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025-02263754; RS-2021-II211341, Artificial Intelligence Graduate School Program, Chung-Ang University). Garud Iyengar's research was partially supported by ONR grant N000142312374 and NSF grant EFMA-2132142.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 3–11. Morgan Kaufmann Publishers Inc., 1999.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 127–135. PMLR, 2013.

Ariu, K., Abe, K., and Proutiere, A. Thresholded lasso bandit. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 878–928. PMLR, 2022.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

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

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Bastani, H. and Bayati, M. Online decision making with high-dimensional covariates. *Operations Research*, 68 (1):276–294, 2020.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic linear bandits robust to adversarial attacks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 991–999. PMLR, 2021.

Bühlmann, P. and van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Berlin Heidelberg, 2011.

Chakraborty, S., Roy, S., and Tewari, A. Thompson sampling for high-dimensional sparse linear contextual bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3979–4008. PMLR, 2023.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214. PMLR, 2011.

Dani, V., 9, ., Hayes, T., and Kakade, S. M. Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland*, pp. 355–366, 2008.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. Balanced linear contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3445–3453, 2019.

Ghosh, A., Ray Chowdhury, S., and Gopalan, A. Misspecified linear bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.

Goldenshluger, A. and Zeevi, A. A linear response bandit problem. *Stochastic Systems*, 3(1):230 – 261, 2013.

Hao, B., Lattimore, T., and Wang, M. High-dimensional sparse linear bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10753–10763. Curran Associates, Inc., 2020.

He, J., Zhou, D., Zhang, T., and Gu, Q. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34614–34625. Curran Associates, Inc., 2022.

Kim, G.-S. and Paik, M. C. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kim, J.-H., Yun, S.-Y., Jeong, M., Nam, J., Shin, J., and Combes, R. Contextual linear bandits under noisy features: Towards bayesian oracles. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1624–1645. PMLR, 2023a.

Kim, W., Kim, G.-S., and Paik, M. C. Doubly robust thompson sampling with linear payoffs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15830–15840. Curran Associates, Inc., 2021.

Kim, W., Lee, K., and Paik, M. C. Double doubly robust thompson sampling for generalized linear contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7):8300–8307, 2023b.

Kim, W., Paik, M. C., and Oh, M.-H. Squeeze all: Novel estimator and self-normalized bound for linear contextual bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3098–3124. PMLR, 2023c.

Kim, W., Iyengar, G., and Zeevi, A. A doubly robust approach to sparse reinforcement learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2305–2313. PMLR, 2024.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, pp. 4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in RL with a generative model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5662–5670. PMLR, 2020.

Lee, H., Hwang, T., and Oh, M.-h. Lasso bandit with compatibility condition on optimal arm. In *International Conference on Learning Representations*, 2025.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670. Association for Computing Machinery, 2010.

Oh, M.-H., Iyengar, G., and Zeevi, A. Sparsity-agnostic lasso bandit. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8271–8280. PMLR, 2021.

Oswal, U., Bhargava, A., and Nowak, R. Linear bandits with feature feedback. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 5331–5338. AAAI Press, 2020.

Park, H. and Faradonbeh, M. K. S. A regret bound for greedy partially observed stochastic contextual bandits. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*, 2022.

Park, H. and Faradonbeh, M. K. S. Thompson sampling in partially observable contextual bandits. *arXiv preprint arXiv:2402.10289*, 2024.

Robbins, H. E. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535, 1952.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Tennenholtz, G., Shalit, U., Mannor, S., and Efroni, Y. Bandits with partially observable confounded data. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 430–439. PMLR, 2021.

Tran, N. P., Ta, T. A., Mandal, D., and Tran-Thanh, L. Symmetric linear bandits with hidden symmetry. In *Advances in Neural Information Processing Systems*, volume 37, pp. 128699–128733. Curran Associates, Inc., 2024.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

Tropp, J. A. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230, 2015.

van de Geer, S. A. and Bühlmann, P. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360 – 1392, 2009.

Zeng, S., Bhatt, S., Koppel, A., and Ganesh, S. Partially observable contextual bandits with linear payoffs. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

A. Related Works

In bandit problems, the learning agent learns only from the outcomes of chosen actions, leaving unchosen alternatives unknown (Robbins, 1952). This constraint requires a balance between exploring new actions and exploiting actions learned to be good, known as the exploration-exploitation tradeoff. Efficiently managing this tradeoff is crucial for guiding the agent towards the optimal policy. To address this, algorithms based on optimism in the face of uncertainty (Lai & Robbins, 1985) employ the Upper Confidence Bound (UCB) strategy, which encourages the learner to select actions with the highest sum of estimated reward and uncertainty. This approach adaptively balances exploration and exploitation, and has been widely and studied in the context of linear bandits (Abe & Long, 1999; Auer, 2002; Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010). Notable examples include `LinUCB` (Li et al., 2010; Chu et al., 2011) and `OFUL` (Abbasi-yadkori et al., 2011), which are known for their practicality and performance guarantees. However, existing approaches differ from ours in two key aspects: (i) they assume that the learning agent can observe the entire feature vector related to the reward, and (ii) their algorithms have regret that scales linearly with the dimension of the observed feature vector, i.e., $\widetilde{O}(d\sqrt{T})$.

In contrast, we develop an algorithm that achieves a sublinear regret bound by employing the doubly robust (DR) technique, thereby avoiding the linear dependence on the dimension of the feature vectors. The DR estimation in the framework of linear contextual bandits was first introduced by Kim & Paik (2019) and Dimakopoulou et al. (2019), and subsequent studies improve the regret bound in this problem setting by a factor of $\sqrt{d}$ (Kim et al., 2021; 2023b). A recent application (Kim et al., 2023c) achieves a regret bound of order $O(\sqrt{dT\log T})$ under IID features over rounds. However, the extension to non-stochastic or non-IID features remains an open question. To address this issue, we develop a novel analysis that applies the DR estimation to non-stochastic features, achieving a regret bound sublinear with respect to the dimension of the augmented feature vectors. Furthermore, we extend DR estimation to handle sparse parameters, thereby further improving the regret bound to be sublinear in the reduced dimension.

Our problem is more general and challenging than misspecified linear bandits, where the assumed reward model fails to accurately reflect the true reward, such as when the true reward function is non-linear (Lattimore & Szepesvári, 2020), or when a deviation term is added to the reward model (Ghosh et al., 2017; Bogunovic et al., 2021; He et al., 2022). While our work assumes that the misspecified (or inaccessible) portion of the reward is linearly related to certain unobserved features, misspecified linear bandit problems can be reformulated as a special case of our framework. While the regret bounds in Lattimore & Szepesvári (2020), Bogunovic et al. (2021) and He et al. (2022) incorporate the sum of misspecification errors that may accumulate over the decision horizon, our work establishes a regret bound that is sublinear in the decision horizon T and is not affected by misspecification errors. Ghosh et al. (2017) proposed a hypothesis test to decide between using linear bandits or MAB, demonstrating an $O(K\sqrt{T}\log T)$ regret bound when the total misspecification error exceeds $\Omega(d\sqrt{T})$. In contrast, our algorithm achieves an $O(\sqrt{(d+d_h)T\log T})$ regret bound without requiring hypothesis tests for misspecification or partial observability.

Last but not least, our problem appears similar to the literature addressing bandit problems with partially observable features (Tennenholtz et al., 2021; Park & Faradonbeh, 2022; 2024; Kim et al., 2023a; Zeng et al., 2025). In particular, Park & Faradonbeh (2022; 2024); Kim et al. (2023a) assume that the true features follow a specific distribution, typically Gaussian. Zeng et al. (2025) further assume that the true features evolve according to a linear dynamical system with additive Gaussian noise. Park & Faradonbeh (2022; 2024) and Zeng et al. (2025) construct the observed features as emissions from the true features via a known linear mapping, also corrupted by additive Gaussian noise, whereas Kim et al. (2023a) first corrupt the true features with Gaussian noise and then generate the observed features by masking elements of the corrupted features following an unknown Bernoulli distribution. In addition, all of these approaches aim to recover the true features: Park & Faradonbeh (2022; 2024) introduce a known decoder mapping from the observed features to the corresponding latent features; Kim et al. (2023a) leverage a Bayesian oracle strategy for estimation; and Zeng et al. (2025) estimate the true features using a Kalman filter. In contrast, our setting imposes no structural assumptions on either the observed or latent features, making the problem more general and challenging than those addressed in the aforementioned works. Furthermore, our approach does not attempt to recover any information related to latent features. Instead, we compensate for the lack of reward information due to unobserved features, in the sense that we project the inaccessible portion of the reward onto the orthogonal complement of the row space spanned by the observed feature vectors.

On the other hand, Tennenholtz et al. (2021) assume that partially observed features are available as an offline dataset, and leverage the dataset to recover up to L dimensions of the d-dimensional true features, where $L \leq d$ is the dimension of the partially observed features. This setting is different from ours in which partial observability arises naturally and no offline access is available. In their framework, the correlation between the observed and unobserved features is used to model the

relationship between the estimator based on the observed features and that based on the true features, which requires further estimation of the unknown correlation. In contrast, our method is agnostic to such correlation, which makes our approach more practical. Moreover, we exploit the observed features via feature augmentation and DR estimation, resulting in a faster convergence rate of regret compared to their UCB-based approach.

B. Detailed Experimental Setup

For both scenarios, the features—including the true features $\mathbf{z}_a$, observed features $\mathbf{x}_a$, and unobserved features $\mathbf{u}_a$—are constructed differently based on the relationship between the row space spanned by the observed features, $\mathrm{R}(\mathbf{X})$, and the row space spanned by the unobserved features, $\mathrm{R}(\mathbf{U})$. In Case 1, the general case, the true features $\mathbf{z}_a$ for each arm $a \in [K]$ are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d_z})$, and the observed features $\mathbf{x}_a$ are obtained by truncating the first d elements of $\mathbf{z}_a$, following the definition given in Eq. (1).

In Cases 2 and 3, on the other hand, the features are generated in a way to explicitly reflect the inclusion relationship between $\mathrm{R}(\mathbf{X})$ and $\mathrm{R}(\mathbf{U})$. Particularly, in Case 2, where $\mathrm{R}(\mathbf{U}) \subseteq \mathrm{R}(\mathbf{X})$, the observed features $\mathbf{x}_a$ are sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for each $a \in [K]$. Then, we generate a coefficient matrix $\mathbf{C}_\mathbf{u} \in \mathbb{R}^{d_u \times d}$, where each element is sampled from $\mathrm{Unif}(-1, 1)$, and construct $\mathbf{u}_a$ by computing $\mathbf{u}_a = \mathbf{C}_\mathbf{u}\mathbf{x}_a$. This construction ensures that $\mathrm{R}(\mathbf{U})$ lies within $\mathrm{R}(\mathbf{X})$. In Case 3, where $\mathrm{R}(\mathbf{X}) \subseteq \mathrm{R}(\mathbf{U})$, we reverse the process by sampling $\mathbf{u}_a \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_u})$, generating $\mathbf{C}_\mathbf{x} \in \mathbb{R}^{d \times d_u}$ from $\mathrm{Unif}(-1, 1)$, and we set $\mathbf{x}_a = \mathbf{C}_\mathbf{x}\mathbf{u}_a$, thereby ensuring $\mathrm{R}(\mathbf{X}) \subseteq \mathrm{R}(\mathbf{U})$. In both Case 2 and Case 3, the true features $\mathbf{z}_a$ are formed by concatenating $\mathbf{x}_a$ and $\mathbf{u}_a$.

After the construction of the features, the orthogonal basis vectors $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\}$ are derived via singular value decomposition (SVD) on the observed feature matrix $\mathbf{X}$, ensuring orthogonality to $\mathrm{R}(\mathbf{X})$. These basis vectors are linearly concatenated to $\mathbf{X}$ to form the augmented feature matrix. The reward parameter $\boldsymbol{\theta}_\star \in \mathbb{R}^{d_z}$ is sampled from the uniform distribution $\mathrm{Unif}(-1/2, 1/2)$, and the rewards are generated via dot products following the definition Eq. (2). The coupling probability p, a hyperparameter used in the sampling distribution of $\widetilde{a}_t$, is set to 0.6 (see Eq. (9)). The confidence parameter δ, which is also a hyperparameter, is set to 10^{-4}, and the total decision horizon is $T = 1200$.

Throughout the experiments, we fix the number of arms at $K = 30$ and the dimensionality of the true features $d_z = 35$. Furthermore, to accommodate both partial and full observability, we set $d_z \geq d$. Specifically, in Scenario 1, d is set to $\lfloor d_z/2 \rfloor = 17$, indicating that only about half of the full feature space is observable to the agent. In Scenario 2, on the other hand, we set $d = 2K = 60$ and $d_z = d$, implying that latent features are absent. The setup of the second scenario also allows us to evaluate our algorithm in the case where $d > K$. For each of the three structural cases (i.e., the relationships between $\mathrm{R}(\mathbf{X})$ and $\mathrm{R}(\mathbf{U})$) considered under both scenarios, we conduct five independent trials using different random seeds. The results are presented in terms of the sample mean and one standard deviation of the cumulative regret.

C. Robust to Latent Feature Algorithm with Ridge Estimator

Our Doubly robust (DR) ridge estimator is defined as follows:

$$\widehat{\boldsymbol{\mu}}_t^R := \left(\sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a \in [K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top + \mathbf{I}_K \right)^{-1} \left(\sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a \in [K]} \widetilde{\mathbf{x}}_a \widetilde{y}_{a,\tau} \right), \tag{12}$$

where $\widetilde{y}_{a,\tau}$ is the DR pseudo reward:

$$\widetilde{y}_{a,t} := \widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t^R + \frac{\mathbb{I}(\widetilde{a}_t = a)}{\phi_{a,t}} \big(y_{a,t} - \widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t^R \big),$$

and the imputation estimator $\check{\boldsymbol{\mu}}_t^R$ is defined as

$$\check{\boldsymbol{\mu}}_t^R := \left(\sum_{\tau=1}^{t} \widetilde{\mathbf{x}}_{a_\tau} \widetilde{\mathbf{x}}_{a_\tau}^\top + p\mathbf{I}_K \right)^{-1} \left(\sum_{\tau=1}^{t} \widetilde{\mathbf{x}}_{a_\tau} y_{a_\tau,\tau} \right). \tag{13}$$

The following theorem shows that this Ridge estimator is consistent, meaning it converges to the true parameter $\boldsymbol{\mu}_\star$ with high probability as the agent interacts with the environment.

Algorithm 2 Robust to Latent Feature with Ridge Estimator (`RoLF-Ridge`)

INPUT: Observed features $\{\mathbf{x}_a : a \in [K]\}$, coupling probability $p \in (1/2, 1)$, confidence parameter $\delta > 0$.
Initialize $\widehat{\boldsymbol{\mu}}_0 = \mathbf{0}_K$, exploration phase $\mathcal{E}_t = \emptyset$ and exploration factor $C_{\mathrm{e}} := 32(1-p)^{-2}K^2$.
Find orthogonal basis $\{\mathbf{b}_1^\top, \ldots, \mathbf{b}_{K-d}^\top\} \subseteq \mathrm{R}(\mathbf{X})^\perp$ and construct $\{\widetilde{\mathbf{x}}_a : a \in [K]\}$.
for $t = 1, \ldots, T$ **do**
  if $|\mathcal{E}_t| \leq C_{\mathrm{e}} \log(2Kt^2/\delta)$ **then**
    Randomly sample a_t uniformly over $[K]$ and $\mathcal{E}_t = \mathcal{E}_{t-1} \cup \{t\}$.
  else
    Compute $\widehat{a}_t := \arg\max_{a \in [K]} \widetilde{\mathbf{x}}_a^\top \widehat{\boldsymbol{\mu}}_{t-1}^R$.
  end if
  while $\widetilde{a}_t \neq a_t$ and count $\leq \rho_t$ **do**
    Sample a_t with $\mathbb{P}(a_t = \widehat{a}_t) = 1 - (t^{-1/2})$ and $\mathbb{P}(a_t = k) = t^{-1/2}/(K-1),\ \forall k \neq \widehat{a}_t$.
    Sample $\widetilde{a}_t$ according to Eq. (9).
    count = count + 1.
  end while
  Play a_t and observe $y_{a_t,t}$.
  if $\widetilde{a}_t \neq a_t$ **then**
    Set $\widehat{\boldsymbol{\mu}}_t^R := \widehat{\boldsymbol{\mu}}_{t-1}^R$.
  else
    Update $\widehat{\boldsymbol{\mu}}_t^R$ following Eq. (12) with $\widetilde{y}_{a,t}$ and update $\check{\boldsymbol{\mu}}_t^R$ following Eq. (13).
  end if
end for

Theorem 4 (Consistency of the DR Ridge estimator)**.** *For each t, let $\mathcal{E}_t \subseteq [t]$ denote an exploration phase such that for each $\tau \in \mathcal{E}_t$ the action a_τ is sampled uniformly over $[K]$. Assume that $\|\boldsymbol{\mu}_\star\|_\infty \leq 1$ and $\|\widetilde{\mathbf{x}}_a\|_\infty \leq 1$ for all $a \in [K]$. Then with probability at least $1 - 3\delta$,*

$$\max_{a \in [K]} |\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star)| \leq \frac{2}{\sqrt{t}} \left(\frac{\sigma}{p} \sqrt{K \log \frac{t+1}{\delta}} + \sqrt{K} \right),$$

for all rounds t such that $|\mathcal{E}_t| \geq 32(1-p)^{-2}K^2 \log(2Kt^2/\delta)$.

With $|\mathcal{E}_t| = O(K^2 \log Kt)$ exploration rounds, the DR Ridge estimator achieves an $O(\sqrt{K/t})$ convergence rate over all K rewards. This is possible because the DR pseudo-rewards defined in Eq. (10) impute the missing rewards for all arms $a \in [K]$ using $\widetilde{\mathbf{x}}_a^\top \check{\boldsymbol{\mu}}_t$, based on the samples collected during the exploration phase, $\mathcal{E}_t$. With this convergence guarantee, we establish a regret bound for `RoLF-Ridge`, which is an adaptation of Algorithm 1 using the Ridge estimator.

Theorem 5 (Regret bound for Ridge `RoLF`)**.** *Suppose $\|\boldsymbol{\mu}_\star\|_\infty \leq 1$ and $\|\mathbf{z}_a\|_\infty \leq 1$ for all $a \in [K]$. For $\delta \in (0, 1)$, with probability at least $1 - 4\delta$, the cumulative regret of the proposed algorithm using the DR Ridge estimator is bounded by*

$$\mathrm{Reg}(T) \leq \frac{32K^2}{(1-p)^2} \log \frac{2dT^2}{\delta} + 2\sqrt{2T \log \frac{2}{\delta}} + \frac{4\sqrt{T}}{K-1} + 4\delta + 8\sqrt{KT} \left(\frac{\sigma}{p} \sqrt{\log \frac{T}{\delta}} + 1 \right),$$

The first and second terms come from the distribution of a_t, which is a combination of the $1 - t^{-1/2}$-greedy policy and resampling up to $\rho_t := \log((t+1)^2/\delta)/\log(1/p)$ trials. The third term is determined by the size of the exploration set, $\mathcal{E}_t$, while the last term arises from the estimation error bounded by the DR estimator as described in Theorem 4. The hyperparameter $p \in (1/2, 1)$ balances the size of the exploration set in the third term and the estimation error in the last term. Overall, the regret is $O(\sqrt{KT \log T})$, which shows a significant improvement compared to the regret lower bound in Theorem 1 for any linear bandit algorithm that does not account for unobserved features and unobserved rewards.

D. A Modified Algorithm for Time-Varying Observed Features

In this section, we define the problem of linear bandits with partially observable features under a setting where the observed features vary over time, describe our proposed method, and provide theoretical guarantees.

D.1. Problem Formulation

Let $\mathbf{x}_{1,t}, \dots, \mathbf{x}_{K,t}$ denote the observed features and $\mathbf{u}_1, \dots, \mathbf{u}_K$ denote the unobserved features. Now the observed features arbitrarily change over t but the unobserved features are fixed over time. When the algorithm selects an arm a_t, the reward is

$$y_{a_t,t} = \langle \mathbf{x}_{a_t,t}, \boldsymbol{\theta}_\star^{(o)} \rangle + \langle \mathbf{u}_{a_t}, \boldsymbol{\theta}_\star^{(u)} \rangle + \epsilon_t,$$

where ϵ_t is Sub-Gaussian noise that follows Assumption 1. The expected reward of each arm is stable over time, where MAB algorithms without using features are applicable for achieving a $\widetilde{O}(\sqrt{KT})$ regret bound. When the observed features vary over time, the expected reward of each arm $\mathbb{E}[y_{a_t,t}] = \langle \mathbf{x}_{a_t,t}, \boldsymbol{\theta}_\star^{(o)} \rangle + \langle \mathbf{u}_{a_t}, \boldsymbol{\theta}_\star^{(u)} \rangle$ also arbitrarily changes over time, and MAB algorithms suffer regret linear in T. To our knowledge, there is no other work that addresses this challenging setting.

D.2. Proposed Method: Orthogonal Basis Augmentation

We address the problem by augmenting standard basis $\mathbf{e}_1, \dots, \mathbf{e}_K$ in $\mathbb{R}^K$ to estimate bias caused by the unobserved features. Let $\widetilde{\mathbf{x}}_{a,t} := \mathbf{e}_a^\top [\mathbf{X}_t \ \mathbf{e}_1 \ \cdots \ \mathbf{e}_K] \in \mathbb{R}^{d+K}$ and let $\Delta_a := \langle \mathbf{u}_a, \theta_\star^{(u)} \rangle$ denote the bias that stems from the latent features. Then,

$$\begin{aligned} y_{a,t} &= \langle \mathbf{x}_{a,t}^\top, \boldsymbol{\theta}_\star^{(o)} \rangle + \langle \mathbf{u}_{a,t}^\top, \boldsymbol{\theta}_\star^{(u)} \rangle + \epsilon_t \\ &= \langle \mathbf{e}_a^\top [\mathbf{X}_t \ \mathbf{e}_1 \ \dots \ \mathbf{e}_K], [\boldsymbol{\theta}_\star^{(o)} \ \Delta_1 \ \dots \ \Delta_K] \rangle + \epsilon_t. \end{aligned}$$

Therefore, applying the `RoLF-Ridge` algorithm to the augmented features $\widetilde{\mathbf{x}}_{a,t} := \mathbf{e}_a^\top [\mathbf{X}_t \ \mathbf{e}_1 \ \dots \ \mathbf{e}_K]$ yields the following regret bound.

Theorem 6 (Regret bound for `Ridge-RoLF-V` with time-varying observed features)**.** *If the observed features are vary over time, then for any $\delta \in (0,1)$, with probability at least $1-4\delta$, the cumulative regret of the proposed algorithm `Ridge-RoLF-V` using DR Ridge estimator is bounded by*

$$\mathrm{Reg}(T) \leq 4\delta + \frac{2\sqrt{T}}{d+K-1} + \frac{32(K+d)^2}{(1-p)^2} \log \frac{2(K+d)T^2}{\delta} + 8\sqrt{(d+K)T} \left(\frac{\sigma}{p} \sqrt{\log \frac{T^2}{\delta}} + 1 \right).$$

The proof follows similar arguments given in Theorem 5 and is therefore omitted. The resulting regret bound achieves a rate of $\widetilde{O}(\sqrt{(d+K)T})$, which, to the best of our knowledge, is the first sublinear regret guarantee for partially observable linear bandits (as well as misspecified linear bandits) with arbitrarily time-varying observed features.

E. Missing Proofs

E.1. Proof of Theorem 1

Throughout this paper, we consider a bandit problem where the agent observes only a subset of the reward-generating feature vector and cannot access or estimate the unobserved portion. If the agent uses online decision-making algorithms that rely solely on observed features, as defined in Definition 1, the resulting issue can be interpreted as a model misspecification. Therefore, in this theorem, we present a problem instance where "misspecified" algorithms, considering only observed features, may incur regret that grows linearly in T.

Following the statement of Theorem 1, we assume that $d = d_u = 1$, which means $d_z = 2$. Given the true feature set $\mathcal{Z} = \{[1,3]^\top, [2, 19/4]^\top\}$, let the first element of each vector is observed to the agent; while the second element remains unobserved. This results in $\mathbf{x}_1 = x_1 = 1$, $\mathbf{x}_2 = x_2 = 2$, $\mathbf{u}_1 = u_1 = 3$ and $\mathbf{u}_2 = u_2 = 19/4$. We set the true parameter as $\boldsymbol{\theta}_\star \in \mathbb{R}^2 = [2, -1]^\top$, meaning $\boldsymbol{\theta}_\star^{(o)} = \theta_\star^{(o)} = 2$ and $\boldsymbol{\theta}_\star^{(u)} = \theta_\star^{(u)} = -1$. Using the reward function from Section 3.2 and considering Assumption 1, the expected reward for each arm is given by

$$\gamma_i := \mathbb{E}[y_i] = \mathbf{z}_i^\top \theta_\star = x_i \theta_\star^{(o)} + u_i \theta_\star^{(u)} \quad \forall \, i \in \{1, 2\}.$$

Plugging the values in, the true mean reward for each arm is directly computed as $\gamma_1 = 2 - 3 = -1$ and $\gamma_2 = 4 - 19/4 = -3/4$, which satisfies the assumption that its absolute value does not exceed 1 (Section 3.2), and since $\gamma_1 < \gamma_2$, arm 2 is the

optimal action. We further assume that the total learning horizon $T > 256\sigma^2 \max\{\log 2, \log(1/\delta')\}$ for any $\delta' \in (0, 1/2)$, where σ is the sub-Gaussian parameter of the reward noise.

For brevity, we denote the latent reward components as $g_1 := u_1\theta_\star^{(u)}$ and $g_2 := u_2\theta_\star^{(u)}$, yielding $g_1 = -3$ and $g_2 = -19/4$. Since $\gamma_2 \neq 2\gamma_1$ and $|g_i| \geq 3 > 0$ for all $i \in \{1, 2\}$, our problem setup satisfies the "large deviation" criterion in Definition 1 and Theorem 2 of Ghosh et al. (2017), with $l = 3$ and $\beta = 0$. Applying the theorem, it follows that `OFUL` (Abbasi-yadkori et al., 2011) suffers linear regret in this problem instance, i.e., $\Omega(T)$. Motivated by this result, we further show that `LinTS` (Agrawal & Goyal, 2013) is similarly affected.

For each round $t \in [T]$, `LinTS` estimates the true parameter using the ridge estimator, given by:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_t &= (\mathbf{X}_t^\top \mathbf{X}_t + \lambda \mathbf{I}_d)^{-1}(\mathbf{X}_t^\top \mathbf{Y}_t) \\ &= (\mathbf{X}_t^\top \mathbf{X}_t + \lambda \mathbf{I}_d)^{-1}(\mathbf{X}_t^\top (\mathbf{X}_t \boldsymbol{\theta}_\star^{(o)} + \mathbf{g}_t + \boldsymbol{\epsilon}_t)) \\ &= \boldsymbol{\theta}_\star^{(o)} - \lambda \mathbf{V}_t^{-1}\boldsymbol{\theta}_\star^{(o)} + \mathbf{V}_t^{-1}\mathbf{X}_t^\top \mathbf{g}_t + \mathbf{V}_t^{-1}\mathbf{X}_t^\top \boldsymbol{\epsilon}_t, \end{aligned} \tag{14}$$

where $\mathbf{X}_t := (\mathbf{x}_{a_1}^\top, \ldots, \mathbf{x}_{a_t}^\top) \in \mathbb{R}^{t \times d}$ is a matrix containing features chosen up to round t, $\mathbf{Y}_t := (y_{a_1}, \ldots, y_{a_t}) \in \mathbb{R}^t$ is a vector of observed rewards, and $\boldsymbol{\epsilon}_t := (\epsilon_1, \ldots, \epsilon_t) \in \mathbb{R}^t$ contains noise attached to each reward. Unlike a typical ridge estimator, here the term $\mathbf{g}_t := (g_{a_1}, \ldots, g_{a_t}) \in \mathbb{R}^t$, the vector containing the latent portion of observed rewards is introduced due to model misspecification. Note that $\mathbf{V}_t := (\mathbf{X}_t^\top \mathbf{X}_t + \lambda \mathbf{I}_d) \succ 0$.

For this problem instance, since $d = 1$, Eq. (14) is equivalent to:

$$\widehat{\theta}_t = \theta_\star^{(o)} - \frac{\theta_\star^{(o)}}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} g_{a_\tau}}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} \epsilon_\tau}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1},$$

where we assume $\lambda = 1$. Note that we also denote $\widehat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}_\star^{(o)}$ by $\widehat{\theta}_t$ and $\theta_\star^{(o)}$, respectively, since both are scalars. Hence, the estimation error is computed as:

$$\widehat{\theta}_t - \theta_\star^{(o)} = -\frac{\theta_\star^{(o)}}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} g_{a_\tau}}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} \epsilon_\tau}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1}. \tag{15}$$

Let N_1 and N_2 denote the number of times arms 1 and 2 have been played up to round t, respectively. This implies that $N_1 + N_2 = t$. Then, for the numerator of the second term, since

$$\sum_{\tau=1}^t x_{a_\tau} g_{a_\tau} = (\underbrace{g_1 + \cdots + g_1}_{N_1} + \underbrace{2g_2 + \cdots + 2g_2}_{N_2}) = g_1 N_1 + 2g_2 N_2,$$

we can observe that

$$\begin{aligned}\sum_{\tau=1}^t x_{a_\tau} g_{a_\tau} &= g_1 N_1 + 2g_2 N_2 \\ &\geq \underline{g} N_1 + 2\underline{g} N_2 \\ &= \underline{g} N_1 + 2\underline{g}(t - N_1) \\ &= 2\underline{g} t - \underline{g} N_1 \\ &\geq \underline{g} t \qquad (\because N_1 \leq t) \\ &= -\frac{19}{4} t,\end{aligned}$$

where $\underline{g} = \min\{g_1, g_2\} = -19/4$, which implies that $\sum_{\tau=1}^t x_{a_\tau} g_{a_\tau} = \Theta(t)$. For the denominator, $\sum_{\tau=1}^t x_{a_\tau}^2 + 1$, which grows at a rate of $O(t)$, implying that the second term of the right-hand side in Eq. (15) is $\Theta(1)$, and that $\widehat{\theta}_t$ is not consistent since it does not converge to $\theta_\star^{(o)}$ as $t \to \infty$.

For arm 2, which is optimal, to be selected in round $t + 1$ under `LinTS`, the condition $x_2\widetilde{\theta}_t \geq x_1\widetilde{\theta}_t$ must hold, where $\widetilde{\theta}_t \sim \mathcal{N}\left(\widehat{\theta}_t, v^2(\sum_{\tau=1}^t x_{a_\tau}^2 + 1)^{-1}\right)$. Given the assumptions that $x_1 = 1$ and $x_2 = 2$, arm 2 is selected whenever $\widetilde{\theta}_t \geq 0$.

Thus, for arm 1 to be chosen, we require $\widetilde{\theta}_t < 0$. We will show that the probability of $\widetilde{\theta}_t < 0$ does not diminish sufficiently to be ignored, even when the agent plays for a sufficiently large amount of time. To clarify, let us define the two events $E_{\widetilde{\theta}} := \{\widetilde{\theta}_t \geq 0\}$ and $E_{\widehat{\theta}} := \{\widehat{\theta}_t \geq 0\}$. We revisit Eq. (14) as follows:

$$\begin{aligned}
\widehat{\theta}_t &= \frac{\theta_\star^{(o)} \sum_{\tau=1}^t x_{a_\tau}^2}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} g_{a_\tau}}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} \epsilon_\tau}{\sum_{\tau=1}^t x_{a_\tau}^2 + 1} \\
&\leq \theta_\star^{(o)} + \frac{g_1 N_1 + 2 g_2 N_2}{N_1 + 4N_2 + 1} + \frac{\sum_{\tau=1}^t x_{a_\tau} \epsilon_\tau}{N_1 + 4N_2 + 1} \qquad (\because \theta_\star^{(o)} > 0) \\
&\leq \theta_\star^{(o)} + \frac{g_1 N_1 + 2 g_2 N_2}{N_1 + 4N_2 + 1} + \frac{2}{N_1 + 4N_2 + 1} \sum_{\tau=1}^t \epsilon_\tau, \qquad (16)
\end{aligned}$$

where the last inequality holds since $\sum_{\tau=1}^t x_{a_\tau} \epsilon_\tau \leq \max_{a \in \{1,2\}} x_a \sum_{\tau=1}^t \epsilon_\tau$ and $\max_{a \in \{1,2\}} x_a = 2$. For the second term of Eq. (16), for $t \geq 19$,

$$\begin{aligned}
\frac{g_1 N_1 + 2 g_2 N_2}{N_1 + 4N_2 + 1} &= \frac{-3N_1 - \frac{19}{2} N_2}{N_1 + 4N_2 + 1} \\
&\leq \frac{-\frac{19}{8}(N_1 + 4N_2)}{N_1 + 4N_2 + 1} \\
&= -\frac{19}{8} + \frac{19/8}{t + 3N_2 + 1} \\
&\leq -\frac{19}{8} + \frac{19}{8t},
\end{aligned}$$

which is upper bounded by $-9/4$. This bound is followed by:

$$\begin{aligned}
\widehat{\theta}_t &\leq \theta_\star^{(o)} + \frac{g_1 N_1 + 2 g_2 N_2}{N_1 + 4N_2 + 1} + \frac{2}{N_1 + 4N_2 + 1} \sum_{\tau=1}^t \epsilon_\tau \\
&\leq 2 - \frac{9}{4} + \frac{2}{t + 3N_2 + 1} \sum_{\tau=1}^t \epsilon_\tau \qquad (17) \\
&\leq -\frac{1}{4} + \frac{2}{t} \sum_{\tau=1}^t \epsilon_\tau.
\end{aligned}$$

Thus, we have the following:

$$\mathbb{P}(\widehat{\theta}_t > 0) \leq \mathbb{P}\left(-\frac{1}{4} + \frac{2}{t} \sum_{\tau=1}^t \epsilon_\tau > 0 \right) = \mathbb{P}\left(\frac{2}{t} \sum_{\tau=1}^t \epsilon_\tau > \frac{1}{4} \right).$$

Since ϵ_τ is an IID sub-Gaussian random variable for all $\tau \in [t]$, by applying Hoeffding inequality we obtain: $\mathbb{P}(\widehat{\theta}_t > 0) \leq \exp\left(-t/128\sigma^2\right)$. Given this, we now bound the probability of the event $E_{\widetilde{\theta}}$:

$$\begin{aligned}
\mathbb{P}(E_{\widetilde{\theta}}) &= \mathbb{P}(E_{\widetilde{\theta}} \cap E_{\widehat{\theta}}) + \mathbb{P}(E_{\widetilde{\theta}} \cap E_{\widehat{\theta}}^c) \\
&= \mathbb{P}(E_{\widetilde{\theta}} \mid E_{\widehat{\theta}}) \cdot \mathbb{P}(E_{\widehat{\theta}}) + \mathbb{P}(E_{\widetilde{\theta}} \mid E_{\widehat{\theta}}^c) \cdot \mathbb{P}(E_{\widehat{\theta}}^c) \\
&= \mathbb{P}(\widetilde{\theta}_t \geq 0 \mid \widehat{\theta}_t \geq 0) \cdot \mathbb{P}(\widehat{\theta}_t \geq 0) + \mathbb{P}(\widetilde{\theta}_t \geq 0 \mid \widehat{\theta}_t < 0) \cdot \mathbb{P}(\widehat{\theta}_t < 0) \\
&\leq \exp\left(-\frac{t}{128\sigma^2} \right) + \mathbb{P}(\widetilde{\theta}_t \geq 0 \mid \widehat{\theta}_t < 0). \qquad (18)
\end{aligned}$$

Since the second term of Eq. (18) is calculated under a Gaussian distribution, thus its value does not exceed $1/2$ for all $t \in [T]$, it follows that $\mathbb{P}(E_{\widetilde{\theta}}^c) \geq 1/2 - \exp\left(-t/128\sigma^2\right)$. Note that the total decision horizon $T > 256\sigma^2 \log(1/\delta')$, thus

for any $t > 128\sigma^2 \log(1/\delta')$, we have $\mathbb{P}(E_{\widehat{\theta}}^{\mathrm{c}}) \geq 1/2 - \delta'$. This implies that for rounds beyond $T/2$, the suboptimal arm will be played at least $(1/2 - \delta')T/2$ times for any $\delta' \in (0, 1/2)$, thus incurring

$$\mathbb{E}[\mathrm{Reg}_{\texttt{LinTS}}(T)] \geq (\gamma_2 - \gamma_1)\left(\frac{1}{2} - \delta'\right)\frac{T}{2} = \frac{1}{4}\left(\frac{1}{2} - \delta'\right)\frac{T}{2}.$$

For `OFUL`, we also present another analysis that does not require the assumption that the suboptimal arm is played for initial t rounds, which is taken in Theorem 2 of Ghosh et al. (2017). The optimal arm, arm 2, is selected when

$$x_2\widehat{\theta}_t + \frac{x_2}{\sqrt{1 + \sum_{\tau=1}^{t-1} x_{a_\tau}^2}} > x_1\widehat{\theta}_t + \frac{x_1}{\sqrt{1 + \sum_{\tau=1}^{t-1} x_{a_\tau}^2}}, \tag{19}$$

where $\widehat{\theta}_t$ is the same ridge estimator as in `LinTS`. The inequality Eq. (19) is equivalent to $\widehat{\theta}_t > (1 + \sum_{\tau=1}^{t-1} x_{a_\tau}^2)^{-1/2}$, which implies $\widehat{\theta}_t > 1/\sqrt{2t}$. By Eq. (17),

$$\begin{aligned}
\mathbb{P}\left(\widehat{\theta}_t > \frac{1}{\sqrt{2t}}\right) &\leq \mathbb{P}\left(-\frac{1}{4} + \frac{2}{t}\sum_{\tau=1}^{t}\epsilon_\tau > \frac{1}{\sqrt{2t}}\right) \\
&= \mathbb{P}\left(\frac{2}{t}\sum_{\tau=1}^{t}\epsilon_\tau > \frac{1}{\sqrt{2t}} + \frac{1}{4}\right) \\
&\leq \mathbb{P}\left(\frac{2}{t}\sum_{\tau=1}^{t}\epsilon_\tau > \frac{1}{4}\right) \leq \exp\left(-\frac{t}{128\sigma^2}\right).
\end{aligned}$$

Thus, for $t \geq 128\sigma^2 \log 2$, the probability of selecting arm 2 is less than $1/2$ and for $T > 256\sigma^2 \log 2$,

$$\mathbb{E}[\mathrm{Reg}_{\texttt{OFUL}}(T)] \geq (\gamma_2 - \gamma_1) \cdot \frac{1}{2} \cdot \frac{T}{2} = \frac{T}{16},$$

and the algorithm suffers expected regret linear in T. □

E.2. Proof of Theorem 2

Let $\mathbf{V}_t := \sum_{\tau=1}^{t}\sum_{a\in[K]} \widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top$. Then

$$\max_{a\in[K]} |\widetilde{\mathbf{x}}_a^\top(\widehat{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star)| \leq \sqrt{\sum_{a\in[K]} |\widetilde{\mathbf{x}}_a^\top(\widehat{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star)|^2} = t^{-1/2}\|\widehat{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\|_{\mathbf{V}_t}.$$

To use Lemma 3, we prove a bound for $\|\sum_{\tau=1}^{t}\sum_{a\in[K]}\left(\widetilde{y}_{a,\tau} - \widetilde{\mathbf{x}}_a^\top\boldsymbol{\mu}_t\right)\widetilde{\mathbf{x}}_a\|_\infty$. By definition of $\widetilde{y}_{a,\tau}$,

$$\begin{aligned}
&\left\|\sum_{\tau=1}^{t}\sum_{a\in[K]}\left(\widetilde{y}_{a,\tau} - \widetilde{\mathbf{x}}_a^\top\boldsymbol{\mu}_\star\right)\widetilde{\mathbf{x}}_a\right\|_\infty \\
&= \left\|\sum_{\tau=1}^{t}\sum_{a\in[K]}\left(1 - \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}}\right)\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top\left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right) + \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}}\left(y_{a,\tau} - \widetilde{\mathbf{x}}_a^\top\boldsymbol{\mu}_\star\right)\widetilde{\mathbf{x}}_a\right\|_\infty \\
&\leq \left\|\sum_{\tau=1}^{t}\sum_{a\in[K]}\left(1 - \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}}\right)\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top\left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right)\right\|_\infty + \left\|\sum_{\tau=1}^{t}\sum_{a\in[K]}\frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}}\left(y_{a,\tau} - \widetilde{\mathbf{x}}_a^\top\boldsymbol{\mu}_\star\right)\widetilde{\mathbf{x}}_a\right\|_\infty.
\end{aligned} \tag{20}$$

With probability at least $1-\delta/(\tau+1)^2$, the event $\mathcal{M}_\tau$ happens for all $\tau \geq 1$ and we obtain a pair of matching sample $\widetilde{a}_\tau$ and a_τ. Under $\mathcal{M}_\tau$, the second term in Eq. (20) is equal to,

$$\left\| \sum_{\tau=1}^{t} \sum_{a\in[K]} \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}} \left(y_{a,\tau} - \widetilde{\mathbf{x}}_a^\top \boldsymbol{\mu}_\star\right) \widetilde{\mathbf{x}}_a \right\|_\infty = \left\| \sum_{\tau=1}^{t} \sum_{a\in[K]} \frac{\mathbb{I}(a_\tau = a)}{\phi_{a,\tau}} \left(y_{a,\tau} - \widetilde{\mathbf{x}}_a^\top \boldsymbol{\mu}_\star\right) \widetilde{\mathbf{x}}_a \right\|_\infty$$
$$= \frac{1}{p} \left\| \sum_{\tau=1}^{t} \epsilon_\tau \widetilde{\mathbf{x}}_{a_\tau} \right\|_\infty .$$

Because $\|\mathbf{v}\|_\infty = \max_{i\in[d]} |\mathbf{e}_i^\top \mathbf{v}|$ for any $\mathbf{v} \in \mathbb{R}^d$,

$$\frac{1}{p} \left\| \sum_{\tau=1}^{t} \epsilon_\tau \widetilde{\mathbf{x}}_{a_\tau} \right\|_\infty = \frac{1}{p} \max_{a\in[K]} \left| \sum_{\tau=1}^{t} \epsilon_\tau \mathbf{e}_a^\top \widetilde{\mathbf{x}}_{a_\tau} \right| .$$

Applying Lemma 1, with probability at least $1-\delta/t^2$,

$$\begin{aligned}
\max_{a\in[K]} \left| \sum_{\tau=1}^{t} \epsilon_\tau \mathbf{e}_a^\top \widetilde{\mathbf{x}}_{a_\tau} \right| &\leq \max_{a\in[K]} \sigma \sqrt{2 \sum_{\tau=1}^{t} (\mathbf{e}_a^\top \widetilde{\mathbf{x}}_{a_\tau})^2 \log \frac{2Kt^2}{\delta}} \\
&= \max_{a\in[K]} \sigma \sqrt{2\mathbf{e}_a^\top \left(\sum_{\tau=1}^{t} \widetilde{\mathbf{x}}_{a_\tau} \widetilde{\mathbf{x}}_{a_\tau}^\top \right) \mathbf{e}_a \log \frac{2Kt^2}{\delta}} \\
&\leq \max_{a\in[K]} \sigma \sqrt{2\mathbf{e}_a^\top \mathbf{V}_t \mathbf{e}_a \log \frac{2Kt^2}{\delta}} \\
&= \sigma \widetilde{\sigma}_{\max} \sqrt{2t \log \frac{2Kt^2}{\delta}},
\end{aligned}$$

where the last equality follows by the definition $\widetilde{\sigma}_{\max}^2 = \max_{a\in[K]} \mathbf{e}_a^\top (\sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top) \mathbf{e}_a$. Thus,

$$\frac{1}{p} \left\| \sum_{\tau=1}^{t} \epsilon_\tau \widetilde{\mathbf{x}}_{a_\tau} \right\|_\infty \leq \frac{\sigma \widetilde{\sigma}_{\max}}{p} \sqrt{2t \log \frac{2Kt^2}{\delta}}. \tag{21}$$

Now we turn to the first term in Eq. (20). Define $\mathbf{A}_t := \sum_{\tau=1}^{t} \sum_{a\in[K]} \mathbb{I}(\widetilde{a}_\tau = a) \phi_{a,\tau}^{-1} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top$. Then the first term is rearranged as

$$\left\| \sum_{\tau=1}^{t} \sum_{a\in[K]} \left(1 - \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}}\right) \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top \left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right) \right\|_\infty = \left\| \left(\mathbf{V}_t - \mathbf{A}_t\right) \left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right) \right\|_\infty . \tag{22}$$

Since $\|\mathbf{v}\|_\infty = \max_{i\in[d]} |\mathbf{e}_i^\top \mathbf{v}|$ for any $\mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned}
\left\| \left(\mathbf{V}_t - \mathbf{A}_t\right) \left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right) \right\|_\infty &= \max_{a\in[K]} \left| \mathbf{e}_a^\top \left(\mathbf{V}_t - \mathbf{A}_t\right) \left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right) \right| \\
&\leq \max_{a\in[K]} \left\| \mathbf{e}_a^\top \left(\mathbf{V}_t - \mathbf{A}_t\right) \mathbf{A}_t^{-1/2} \right\|_2 \left\| \check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star \right\|_{\mathbf{A}_t} .
\end{aligned}$$

Because $\check{\boldsymbol{\mu}}_t^L$ is a minimizer of Eq. (8), by Lemma 3 and Eq. (21),

$$\left\| \check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star \right\|_{\mathbf{A}_t} \leq \frac{4\sigma \widetilde{\sigma}_{\max}}{p} \sqrt{\frac{2t(d+d_h) \log(2Kt^2/\delta)}{\lambda_{\min}\left(\mathbf{A}_t\right)}}.$$

By Corollary 1, with $\epsilon \in (0,1)$ to be determined later, for $t \geq 8\epsilon^{-2}(1-p)^{-2}K^2 \log(2Kt^2/\delta)$, with probability at least $1-\delta/t^2$,

$$\left\| \mathbf{I}_K - \mathbf{V}_t^{-1/2} \mathbf{A}_t \mathbf{V}_t^{-1/2} \right\|_2 \leq \epsilon. \tag{23}$$

Eq. (23) implies $(1-\epsilon)\mathbf{I}_K \preceq \mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2}$ and $(1-\epsilon)\mathbf{V}_t \preceq \mathbf{A}_t$. Thus,

$$\left\|\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right\|_{\mathbf{A}_t} \leq \frac{4\sigma\widetilde{\sigma}_{\max}}{p}\sqrt{\frac{2t(d+d_h)\log(2Kt^2/\delta)}{(1-\epsilon)\lambda_{\min}(\mathbf{V}_t)}} = \frac{4\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}}\sqrt{\frac{2(d+d_h)\log(2Kt^2/\delta)}{1-\epsilon}},$$

and Eq. (22) is bounded by,

$$\begin{aligned}\left\|(\mathbf{V}_t - \mathbf{A}_t)\left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right)\right\|_\infty &\leq \max_{i\in[K]}\left\|e_i^\top(\mathbf{V}_t - \mathbf{A}_t)\mathbf{A}_t^{-1/2}\right\|_2 \frac{4\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}}\sqrt{\frac{2(d+d_h)\log(2Kt^2/\delta)}{1-\epsilon}} \\ &\leq \max_{i\in[K]}\left\|e_i^\top(\mathbf{V}_t - \mathbf{A}_t)\mathbf{V}_t^{-1/2}\right\|_2 \frac{4\sigma\widetilde{\sigma}_{\max}}{p(1-\epsilon)\widetilde{\sigma}_{\min}}\sqrt{2(d+d_h)\log\frac{2Kt^2}{\delta}}. \qquad (24)\end{aligned}$$

For the first term in Eq. (24), we have that

$$\begin{aligned}\max_{i\in[K]}\left\|\mathbf{e}_i^\top(\mathbf{V}_t - \mathbf{A}_t)\mathbf{V}_t^{-1/2}\right\|_2 &= \max_{i\in[K]}\left\|\mathbf{e}_i^\top\mathbf{V}_t^{1/2}\mathbf{V}_t^{-1/2}(\mathbf{V}_t - \mathbf{A}_t)\mathbf{V}_t^{-1/2}\right\|_2 \\ &\leq \widetilde{\sigma}_{\max}\sqrt{t}\left\|\mathbf{I}_K - \mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2}\right\|_2 \\ &\leq \widetilde{\sigma}_{\max}\sqrt{t}\epsilon,\end{aligned}$$

where the last inequality holds due to Eq. (23). Combining this result with Eq. (24), we obtain that

$$\left\|(\mathbf{V}_t - \mathbf{A}_t)\left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right)\right\|_\infty \leq \frac{4\epsilon\sigma\widetilde{\sigma}_{\max}^2}{p(1-\epsilon)\widetilde{\sigma}_{\min}}\sqrt{2t(d+d_h)\log\frac{2Kt^2}{\delta}}.$$

Setting $\epsilon = K^{-1/2}\widetilde{\sigma}_{\min}/8\widetilde{\sigma}_{\max}$ yields that $1-\epsilon \geq 1/2$ and that

$$\begin{aligned}\left\|(\mathbf{V}_t - \mathbf{A}_t)\left(\check{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\right)\right\|_\infty &\leq \frac{\sigma\widetilde{\sigma}_{\max}}{p}\sqrt{2t\frac{d+d_h}{K}\log\frac{2Kt^2}{\delta}} \\ &\leq \frac{\sigma\widetilde{\sigma}_{\max}}{p}\sqrt{2t\log\frac{2Kt^2}{\delta}}.\end{aligned}$$

Now we conclude that for $t \geq 8^3K^3\widetilde{\sigma}_{\min}^{-2}\widetilde{\sigma}_{\max}^2(1-p)^{-2}\log(2Kt^2/\delta)$,

$$\left\|\sum_{\tau=1}^t\sum_{a\in[K]}(\widetilde{y}_{a,\tau} - \widetilde{\mathbf{x}}_a^\top\boldsymbol{\mu}_t)\widetilde{\mathbf{x}}_a\right\|_\infty \leq \frac{\sigma\widetilde{\sigma}_{\max}}{p}\sqrt{2t\log\frac{2Kt^2}{\delta}},$$

by Lemma 3 and taking a union bound on the both terms in Eq. (20), with probability at least $1 - 2\delta/t^2$,

$$\|\widehat{\boldsymbol{\mu}}_t^L - \boldsymbol{\mu}_\star\|_{\mathbf{V}_t} \leq \frac{8\sigma\widetilde{\sigma}_{\max}}{p}\sqrt{\frac{2t(d+d_h)\log(2Kt^2/\delta)}{\lambda_{\min}(\mathbf{V}_t)}} = \frac{8\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}}\sqrt{2(d+d_h)\log\frac{2Kt^2}{\delta}},$$

which completes the proof. □

E.3. Proof of Theorem 3

Since we have shown that the reward defined in Eq. (2) is equivalent to its form obtained via the projection and augmentation strategy in Eq. (7), we henceforth calculate the regret bound based on Eq. (7). Under the assumption that the expected reward is bounded by 1 (Section 3.2), the instantaneous regret $\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]$ is bounded above by 2 for any $t \in [T]$, and that the number of rounds for the exploration phase satisfies $|\mathcal{E}_T| \leq 8^3K^3(1-p)^{-2}\log(2KT^2/\delta)$. Given these bounds, the

cumulative regret is bounded as:

$$\begin{aligned}
\mathrm{Reg}(T) &= \sum_{t=1}^{T}(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]) \\
&= \sum_{t\in\mathcal{E}_T}(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]) + \sum_{t\in[T]\setminus\mathcal{E}_T}(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]) \\
&\le 2\cdot 8^3K^3(1-p)^{-2}\log\frac{2KT^2}{\delta} + \sum_{t\in[T]\setminus\mathcal{E}_T}(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]) \\
&= 2\cdot 8^3K^3(1-p)^{-2}\log\frac{2KT^2}{\delta} + \sum_{t\in[T]\setminus\mathcal{E}_T}\{\mathbb{I}(a_t=\widehat{a}_t)(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}])\} \\
&\quad + \sum_{t\in[T]\setminus\mathcal{E}_T}\{\mathbb{I}(a_t\neq\widehat{a}_t)(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}])\}\,.
\end{aligned} \tag{25}$$

We first consider the second term of Eq. (25). By Theorem 2, on the event $\{a_t=\widehat{a}_t\}$,

$$\begin{aligned}
\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}] &= \widetilde{\mathbf{x}}_{a_\star}^\top\boldsymbol{\mu}_\star - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top\boldsymbol{\mu}_\star \\
&= \widetilde{\mathbf{x}}_{a_\star}^\top\boldsymbol{\mu}_\star + \widetilde{\mathbf{x}}_{a_\star}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L - \widetilde{\mathbf{x}}_{a_\star}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L + \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top\boldsymbol{\mu}_\star \\
&\le \left|\widetilde{\mathbf{x}}_{a_\star}^\top(\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^L)\right| + \left|\widetilde{\mathbf{x}}_{\widehat{a}_t}^\top(\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^L)\right| + \widetilde{\mathbf{x}}_{a_\star}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L \\
&\le 2\max_{a\in[K]}\left|\widetilde{\mathbf{x}}_a^\top(\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^L)\right| + \widetilde{\mathbf{x}}_{a_\star}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top\widehat{\boldsymbol{\mu}}_{t-1}^L \\
&\le 2\max_{a\in[K]}\left|\widetilde{\mathbf{x}}_a^\top(\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^L)\right| \\
&\le \frac{16\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}}\sqrt{\frac{2(d+d_h)}{t}\log\frac{2Kt^2}{\delta}},
\end{aligned}$$

with probability at least $1-2\delta/t^2$ for each $t\in[T]\setminus\mathcal{E}_T$. Summing over t and applying a union bound, we obtain that, with probability at least $1-4\delta$,

$$\sum_{t\in[T]\setminus\mathcal{E}_T}\{\mathbb{I}(a_t=\widehat{a}_t)\,(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}])\} \le \frac{32\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}}\sqrt{2(d+d_h)T\log\frac{2KT^2}{\delta}}. \tag{26}$$

For the last term of Eq. (25),

$$\begin{aligned}
&\sum_{t\in[T]\setminus\mathcal{E}_T}\{\mathbb{I}(a_t\neq\widehat{a}_t)\,(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}])\} \\
&\le 2\left[\sum_{t\in[T]}\mathbb{I}(a_t\neq\widehat{a}_t) - \mathbb{P}\,(a_t\neq\widehat{a}_t) + \mathbb{P}\,(a_t\neq\widehat{a}_t)\right] \qquad (\because \mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}] \le 2) \\
&\le 2\sqrt{2T\log\frac{2}{\delta}} + \frac{4\sqrt{T}}{K-1} + 4\delta,
\end{aligned} \tag{27}$$

where the first term in Eq. (27) holds with probability at least $1-\delta$ by Hoeffding's inequality, and the remaining terms follow from Lemma 5. Taking a union bound over Eq. (26), Eq. (27), and $\mathcal{M}_t$, we obtain, with probability at least $1-6\delta$,

$$\begin{aligned}
\mathrm{Reg}(T) &\le 2\cdot 8^3K^3(1-p)^{-2}\log\frac{2KT^2}{\delta} + \frac{4\sqrt{T}}{K-1} \\
&\quad + 2\sqrt{2T\log\frac{2}{\delta}} + 4\delta + \frac{32\sigma\widetilde{\sigma}_{\max}}{p\widetilde{\sigma}_{\min}}\sqrt{2(d+d_h)T\log\frac{2KT^2}{\delta}},
\end{aligned}$$

which concludes the proof. □

E.4. Proof of Theorem 4

Let $\widetilde{\mathbf{V}}_t := \sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top + \mathbf{I}_K$ and $\mathbf{V}_t := \sum_{\tau=1}^{t} \sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top + \mathbf{I}_K$. By definition of $\widehat{\boldsymbol{\mu}}_t^R$ presented in Eq. (12),

$$\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star) = \widetilde{\mathbf{x}}_a^\top \widetilde{\mathbf{V}}_t^{-1} \left\{ \sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a\in[K]} \widetilde{\mathbf{x}}_a \left(\widetilde{y}_{a,\tau} - \widetilde{\mathbf{x}}_a^\top \boldsymbol{\mu}_\star \right) - \boldsymbol{\mu}_\star \right\}.$$

By definition of the pseudo-rewards,

$$\widetilde{y}_{a,\tau} - \widetilde{\mathbf{x}}_a^\top \boldsymbol{\mu}_\star = \left(1 - \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,t}} \right) \widetilde{\mathbf{x}}_a^\top \left(\check{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star \right) + \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}} \epsilon_\tau.$$

Let $\widetilde{\mathbf{A}}_t := \sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a\in[K]} \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,t}} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top + \mathbf{I}_K$ and $\mathbf{A}_t := \sum_{\tau=1}^{t} \sum_{a\in[K]} \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,t}} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top + \mathbf{I}_K$ Then,

$$\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star) = \widetilde{\mathbf{x}}_a^\top \widetilde{\mathbf{V}}_t^{-1} \left\{ \left(\widetilde{\mathbf{V}}_t - \widetilde{\mathbf{A}}_t \right) \left(\check{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star \right) + \sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a\in[K]} \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}} \widetilde{\mathbf{x}}_a \epsilon_\tau - \boldsymbol{\mu}_\star \right\}. \tag{28}$$

By definition of the imputation estimator $\check{\boldsymbol{\mu}}_t$,

$$\begin{aligned}
\check{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star &= \left(\sum_{\tau=1}^{t} \widetilde{\mathbf{x}}_{a_\tau} \widetilde{\mathbf{x}}_{a_\tau}^\top + p\mathbf{I}_K \right)^{-1} \left(\sum_{\tau=1}^{t} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - p\boldsymbol{\mu}_\star \right) \\
&= \left(\sum_{\tau=1}^{t} \frac{1}{\phi_{a_\tau,\tau}} \widetilde{\mathbf{x}}_{a_\tau} \widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K \right)^{-1} \left(\sum_{\tau=1}^{t} \frac{1}{\phi_{a_\tau,\tau}} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right) \\
&= \left(\sum_{\tau=1}^{t} \sum_{a\in[K]} \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}} \widetilde{\mathbf{x}}_{a_\tau} \widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K \right)^{-1} \left(\sum_{\tau=1}^{t} \frac{1}{p} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right),
\end{aligned}$$

where the second equality holds because $\phi_{a_\tau,\tau} = p$ and the coupling event $\cap_{\tau=1}^{t} \mathcal{M}_\tau$. Thus,

$$\check{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star = \mathbf{A}_t^{-1} \left(\sum_{\tau=1}^{t} \frac{1}{p} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right).$$

Plugging in (28),

$$\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star) = \widetilde{\mathbf{x}}_a^\top \widetilde{\mathbf{V}}_t^{-1} \left\{ \left(\widetilde{\mathbf{V}}_t - \widetilde{\mathbf{A}}_t \right) \mathbf{A}_t^{-1} \left(\sum_{\tau=1}^{t} \frac{1}{p} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right) + \sum_{\tau=1}^{t} \mathbb{I}(\mathcal{M}_\tau) \sum_{a\in[K]} \frac{\mathbb{I}(\widetilde{a}_\tau = a)}{\phi_{a,\tau}} \widetilde{\mathbf{x}}_a \epsilon_\tau - \boldsymbol{\mu}_\star \right\}$$

Under the coupling event, $\widetilde{\mathbf{A}}_t = \mathbf{A}_t$, $\widetilde{\mathbf{V}}_t = \mathbf{V}_t$ and $\sum_{a\in[K]} \phi_{a,\tau}^{-1} \mathbb{I}(\widetilde{a}_\tau = a) \widetilde{\mathbf{x}}_a = \widetilde{\mathbf{x}}_{a_\tau} / p$. It follows that

$$\begin{aligned}
\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star) &= \widetilde{\mathbf{x}}_a^\top \mathbf{V}_t^{-1} \left\{ \left(\mathbf{V}_t - \mathbf{A}_t \right) \mathbf{A}_t^{-1} + \mathbf{I}_K \right\} \left(\sum_{\tau=1}^{t} \frac{1}{p} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right) \\
&= \widetilde{\mathbf{x}}_a^\top \mathbf{V}_t^{-1/2} \left(\mathbf{V}_t^{1/2} \mathbf{A}_t^{-1} \mathbf{V}_t^{1/2} \right) \mathbf{V}_t^{-1/2} \left(\sum_{\tau=1}^{t} \frac{1}{p} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right).
\end{aligned}$$

Taking absolute value on both sides, by Cauchy-Schwarz inequality,

$$\max_{a\in[K]} |\widetilde{\mathbf{x}}_a^\top (\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star)| \leq \max_{a\in[K]} \|\widetilde{\mathbf{x}}_a\|_{\mathbf{V}_t^{-1}} \|\mathbf{V}_t^{1/2} \mathbf{A}_t^{-1} \mathbf{V}_t^{1/2}\|_2 \left\| \sum_{\tau=1}^{t} \frac{1}{p} \widetilde{\mathbf{x}}_{a_\tau} \epsilon_\tau - \boldsymbol{\mu}_\star \right\|_{\mathbf{V}_t^{-1}}. \tag{29}$$

Corollary 1 implies that, for $t \geq 8\epsilon^{-2}(1-p)^{-2}K^2\log(2Kt^2/\delta)$, with ϵ to be determined later,

$$\mathbf{I}_K - \mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2} \preceq \epsilon\mathbf{I}_K,$$

with probability at least $1-\delta$ for all $t \geq 1$. Rearranging terms gives

$$\mathbf{V}_t^{1/2}\mathbf{A}_t^{-1}\mathbf{V}_t^{1/2} \preceq (1-\epsilon)^{-1}\mathbf{I}_K.$$

Combining this with Eq. (29),

$$\begin{aligned}
\max_{a\in[K]} |\widetilde{\mathbf{x}}_a^\top(\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star)| &\leq \frac{\max_{a\in[K]}\|\widetilde{\mathbf{x}}_a\|_{\mathbf{V}_t^{-1}}}{1-\epsilon}\left\|\sum_{\tau=1}^t \frac{1}{p}\widetilde{\mathbf{x}}_{a_\tau}\epsilon_\tau - \boldsymbol{\mu}_\star\right\|_{\mathbf{V}_t^{-1}} \\
&\leq \frac{\max_{a\in[K]}\|\widetilde{\mathbf{x}}_a\|_{\mathbf{V}_t^{-1}}}{1-\epsilon}\left(\frac{1}{p}\left\|\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\epsilon_\tau\right\|_{\mathbf{V}_t^{-1}} + \|\boldsymbol{\mu}_\star\|_{\mathbf{V}_t^{-1}}\right),
\end{aligned}$$

where the last inequality follows from the triangle inequality. Note that $\mathbf{V}_t$ is deterministic, both $\mathbf{V}_t$ and $\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K$ are positive definite, and $\mathbf{V}_t \succeq \sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K$. Then, by Lemma 9 of (Abbasi-yadkori et al., 2011), with probability at least $1-\delta$,

$$\begin{aligned}
\left\|\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\epsilon_\tau\right\|_{\mathbf{V}_t^{-1}} &\leq \left\|\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\epsilon_\tau\right\|_{\left(\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K\right)^{-1}} \\
&\leq \sigma\sqrt{2\log\frac{\det(\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K)^{1/2}}{\delta}} \\
&\leq \sigma\sqrt{\log\frac{\det(\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K)}{\delta}}
\end{aligned}$$

Applying Lemma 10 of Abbasi-yadkori et al. (2011) yields

$$\begin{aligned}
\det\left(\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top + \mathbf{I}_K\right) &\leq \left\{\frac{\mathrm{Tr}\left(\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top\right) + K}{K}\right\}^K \\
&\leq \left\{\frac{t\max_{a\in[K]}\|\widetilde{\mathbf{x}}_{a_\tau}\|_2 + K}{K}\right\}^K \\
&\leq \{t+1\}^K,
\end{aligned}$$

for all $t \geq 1$, where the last inequality holds by $\|\widetilde{\mathbf{x}}_{a_\tau}\|_2 \leq \sqrt{K}\|\widetilde{\mathbf{x}}_{a_\tau}\|_\infty \leq K$. Hence,

$$\left\|\sum_{\tau=1}^t \widetilde{\mathbf{x}}_{a_\tau}\epsilon_\tau\right\|_{\mathbf{V}_t^{-1}} \leq \sigma\sqrt{K\log\frac{t+1}{\delta}},$$

which follows that,

$$\begin{aligned}
\max_{a\in[K]} |\widetilde{\mathbf{x}}_a^\top(\widehat{\boldsymbol{\mu}}_t^R - \boldsymbol{\mu}_\star)| &\leq \frac{\max_{a\in[K]}\|\widetilde{\mathbf{x}}_a\|_{\mathbf{V}_t^{-1}}}{1-\epsilon}\left(\frac{\sigma}{p}\sqrt{K\log\frac{t+1}{\delta}} + \|\boldsymbol{\mu}_\star\|_{\mathbf{V}_t^{-1}}\right) \\
&\leq \frac{1}{\sqrt{t}}\cdot\frac{1}{1-\epsilon}\left(\frac{\sigma}{p}\sqrt{K\log\frac{t+1}{\delta}} + \|\boldsymbol{\mu}_\star\|_{\mathbf{V}_t^{-1}}\right).
\end{aligned}$$

Since $\|\boldsymbol{\mu}_\star\|_{\mathbf{V}_t^{-1}} \leq \|\boldsymbol{\mu}_\star\|_2 \leq \sqrt{K}$, choosing $\epsilon = 1/2$ completes the proof. □

E.5. Proof of Theorem 5

Similar to the proof of Theorem 3 (Appendix E.3), the instantaneous regret is bounded above by 2 for any $t \in [T]$, and the number of rounds for the exploration phase satisfies $|\mathcal{E}_T| \leq 32(1-p)^{-2}K^2 \log(2dT^2/\delta)$. Consequently, the cumulative regret is bounded above as follows:

$$\begin{aligned}
\mathrm{Reg}(T) &\leq 32(1-p)^{-2}K^2 \log \frac{2dT^2}{\delta} + \sum_{t\in[T]\setminus\mathcal{E}_T} \mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}] \\
&= 32(1-p)^{-2}K^2 \log \frac{2dT^2}{\delta} + \sum_{t\in[T]\setminus\mathcal{E}_T} \left\{\mathbb{I}(a_t = \widehat{a}_t)\left(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]\right)\right\} \\
&\quad + \sum_{t\in[T]\setminus\mathcal{E}_T} \left\{\mathbb{I}(a_t \neq \widehat{a}_t)\left(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]\right)\right\}.
\end{aligned} \tag{30}$$

We first consider the second term in Eq. (30). On the event $\{a_t = \widehat{a}_t\}$,

$$\begin{aligned}
\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}] &= \widetilde{\mathbf{x}}_{a_\star}^\top \boldsymbol{\mu}_\star - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top \boldsymbol{\mu}_\star \\
&= \widetilde{\mathbf{x}}_{a_\star}^\top \boldsymbol{\mu}_\star + \widetilde{\mathbf{x}}_{a_\star}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R - \widetilde{\mathbf{x}}_{a_\star}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R + \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top \boldsymbol{\mu}_\star \\
&\leq \left|\widetilde{\mathbf{x}}_{a_\star}^\top (\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^R)\right| + \left|\widetilde{\mathbf{x}}_{\widehat{a}_t}^\top (\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^R)\right| + \widetilde{\mathbf{x}}_{a_\star}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R \\
&\leq 2 \max_{a\in[K]} \left|\widetilde{\mathbf{x}}_a^\top \left(\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^R\right)\right| + \widetilde{\mathbf{x}}_{a_\star}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R - \widetilde{\mathbf{x}}_{\widehat{a}_t}^\top \widehat{\boldsymbol{\mu}}_{t-1}^R \\
&\leq 2 \max_{a\in[K]} \left|\widetilde{\mathbf{x}}_a^\top \left(\boldsymbol{\mu}_\star - \widehat{\boldsymbol{\mu}}_{t-1}^R\right)\right| \\
&\leq \frac{4}{\sqrt{t}} \left(\frac{\sigma}{p}\sqrt{K \log \frac{t}{\delta}} + \sqrt{K}\right),
\end{aligned}$$

with probability at least $1 - 3\delta$, for all t such that $t \geq |\mathcal{E}_t|$ (Theorem 4). Summing over t gives, with probability at least $1 - 3\delta$,

$$\sum_{t\in[T]\setminus\mathcal{E}_T} \left\{\mathbb{I}(a_t = \widehat{a}_t)\left(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]\right)\right\} \leq 8\sqrt{KT}\left(\frac{\sigma}{p}\sqrt{\log \frac{T}{\delta}} + 1\right). \tag{31}$$

For the last term in Eq. (30), we have that

$$\begin{aligned}
&\sum_{t\in[T]\setminus\mathcal{E}_T} \left\{\mathbb{I}(a_t \neq \widehat{a}_t)\left(\mathbb{E}_{t-1}[y_{\star,t}] - \mathbb{E}_{t-1}[y_{a_t,t}]\right)\right\} \\
&\leq 2\left[\sum_{t\in[T]} \mathbb{I}(a_t \neq \widehat{a}_t) - \mathbb{P}\left(a_t \neq \widehat{a}_t\right) + \mathbb{P}\left(a_t \neq \widehat{a}_t\right)\right] \\
&\leq 2\sqrt{2T \log \frac{2}{\delta}} + \frac{4\sqrt{T}}{K-1} + 4\delta,
\end{aligned} \tag{32}$$

where the first term in Eq. (32) holds with probability at least $1 - \delta$ by Hoeffding's inequality, and the remaining terms follow from Lemma 5. Finally, by taking a union bound over Eq. (31) and Eq. (32), we obtain, with probability at least $1 - 4\delta$,

$$\mathrm{Reg}(T) \leq \frac{32K^2}{(1-p)^2} \log \frac{2dT^2}{\delta} + 2\sqrt{2T \log \frac{2}{\delta}} + \frac{4\sqrt{T}}{K-1} + 4\delta + 8\sqrt{KT}\left(\frac{\sigma}{p}\sqrt{\log \frac{T}{\delta}} + 1\right),$$

which completes the proof. □

F. Algorithm-Agnostic Lower Bound of Regret Ignoring Unobserved Features

In this section, extending our argument in Theorem 1, we show that there exists a problem instance such that linear bandit algorithms relying solely on observed features can incur regret that grows linearly in T. We begin by formally defining such algorithms.

Definition 1 (Policy dependent on observed features)**.** *For each $t \in [T]$, let $\pi_t : \mathbb{R}^d \times \mathbb{R}^{t-1} \to [0,1]$ be a policy that maps an observed feature vector $\mathbf{x} \in \{\mathbf{x}_a : a \in [K]\}$, given past reward observations $\{y_{a_s,s} : s \in [t-1]\}$, to a probability of selection. Then the policy π_t is dependent* only *on observed features if, for any $y_{a_1,1}, \ldots, y_{a_{t-1},t-1}$, it holds that $\mathbf{x}_1 = \mathbf{x}_2$ implies $\pi_t(\mathbf{x}_1 | y_{a_1,1}, \ldots, y_{a_{t-1},t-1}) = \pi_t(\mathbf{x}_2 | y_{a_1,1}, \ldots, y_{a_{t-1},t-1})$.*

For instance, the UCB and Thompson sampling-based policies for linear bandits (with observed features), considered in Theorem 1, satisfy Definition 1, as they assign the same selection probability as long as the observed features are the same. In contrast, policies in MAB algorithms (which disregard observed features) may assign different selection probabilities even when the observed features are equal, and thus are not dependent on the observed features. In the theorem below, we particularly provide a lower bound for algorithms that employ policies that are dependent on the observed features.

Theorem 7 (Regret Lower Bound under Policies Dependent on Observed Features)**.** *For any algorithm $\Pi := (\pi_1, \ldots, \pi_T)$ that consists of the policies $\{\pi_t : t \in [T]\}$ that are dependent on observed features, there exists a set of features $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ and a parameter $\boldsymbol{\theta}_\star \in \mathbb{R}^{d_z}$ such that the cumulative regret*

$$\mathrm{Reg}_\Pi(T, \boldsymbol{\theta}_\star, \mathbf{z}_1, \ldots, \mathbf{z}_K) \geq \frac{T}{6}.$$

Proof. We start the proof by providing a detailed account of the scenario described in the theorem. Without loss of generality, we consider the case where $K = 3$. As stated in the theorem, $a_\star$ represents the index of the optimal action when considering the entire reward, including both observed and latent components. In contrast, a_o denotes the index of the optimal action when considering only the observed components. We introduce an additional notation, a', which refers to an action whose observed features are identical to those of $a_\star$, but with a distinct latent component. Specifically, this implies that $a' \neq a_\star$ and $\mathbf{z}_{a'} \neq \mathbf{z}_{a_\star}$, but $\mathbf{x}_{a'} = \mathbf{x}_{a_\star}$. By definition of the policy π_t that depends on the observed features, $\pi_t(\mathbf{x}_{a_\star}) = \pi_t(\mathbf{x}_{a'})$ and the probability of selecting the optimal arm is $\pi_t(\mathbf{x}_{a_\star}) \leq 1/2$.

Taking this scenario into account, the observed part of the features associated with $a_\star$, a', and a_o are defined as follows:

$$\mathbf{x}_{a_\star} := \left[-\frac{1}{2}, \ldots, -\frac{1}{2}\right]^\top, \mathbf{x}_{a'} := \left[-\frac{1}{2}, \ldots, -\frac{1}{2}\right]^\top, \mathbf{x}_{a_o} := \left[\frac{1}{2}, \ldots, \frac{1}{2}\right]^\top.$$

Additionally, we define the unobserved feature vectors for actions $a_\star$, a', and a_o as follows:

$$\mathbf{u}_{a_\star} := [1, \ldots, 1]^\top, \mathbf{u}_{a'} := [-1, \ldots, -1]^\top, \mathbf{u}_{a_o} := [-1, \ldots, -1, 1, \ldots, 1]^\top,$$

where in $\mathbf{u}_{a_o}$, the number of 1's and -1's are equal. This ensures that the scenario aligns with the assumption imposed on the feature vectors throughout this paper. We further define the true parameter as follows:

$$\boldsymbol{\theta}_\star := \left[\frac{1}{3d}, \ldots, \frac{1}{3d}, \frac{2}{3d_u}, \ldots, \frac{2}{3d_u}\right]^\top \in \mathbb{R}^{d_z},$$

thus it follows that $\boldsymbol{\theta}_\star^{(o)} = [1/3d, \ldots, 1/3d]^\top \in \mathbb{R}^d$ and $\boldsymbol{\theta}_\star^{(u)} = [2/3d_u, \ldots, 2/3d_u]^\top \in \mathbb{R}^{d_u}$. Note that it is straightforward to verify that $|\langle \mathbf{z}_a, \boldsymbol{\theta}_\star \rangle| \leq 1$, thereby satisfying the assumption on the mean reward (Section 3.2). With this established, we can also observe that the expected rewards for the three actions are defined as:

$$\begin{aligned}
\langle \mathbf{z}_{a_\star}, \boldsymbol{\theta}_\star \rangle &= \langle \mathbf{x}_{a_\star}, \boldsymbol{\theta}_\star^{(o)} \rangle + \langle \mathbf{u}_{a_\star}, \boldsymbol{\theta}_\star^{(u)} \rangle = -\frac{1}{6} + \frac{2}{3} = \frac{1}{2}, \\
\langle \mathbf{z}_{a'}, \boldsymbol{\theta}_\star \rangle &= \langle \mathbf{x}_{a'}, \boldsymbol{\theta}_\star^{(o)} \rangle + \langle \mathbf{u}_{a'}, \boldsymbol{\theta}_\star^{(u)} \rangle = -\frac{1}{6} - \frac{2}{3} = -\frac{5}{6}, \\
\langle \mathbf{z}_{a_o}, \boldsymbol{\theta}_\star \rangle &= \langle \mathbf{x}_{a_o}, \boldsymbol{\theta}_\star^{(o)} \rangle + \langle \mathbf{u}_{a_o}, \boldsymbol{\theta}_\star^{(u)} \rangle = \frac{1}{6} + 0 = \frac{1}{6},
\end{aligned}$$

respectively, and it is straightforward to verify that $\langle \mathbf{z}_{a_\star}, \boldsymbol{\theta}_\star \rangle - \langle \mathbf{z}_{a_o}, \boldsymbol{\theta}_\star \rangle = 2/3 > 0$ and that $\langle \mathbf{z}_{a_\star}, \boldsymbol{\theta}_\star \rangle - \langle \mathbf{z}_{a'}, \boldsymbol{\theta}_\star \rangle = 4/3 > 0$, which confirms that $a_\star$ is optimal when considering the full feature set.

At each round $t \in [T]$ for any policy π_t satisfying Definition 1, we have

$$\pi_t(\mathbf{x}_{a_\star} | y_{a_1,1}, \ldots, y_{a_{t-1},t-1}) = \pi_t(\mathbf{x}_{a'} | y_{a_1,1}, \ldots, y_{a_{t-1},t-1}),$$

which implies $\mathbb{P}(a_t = a_\star) = \mathbb{P}(a_t = a')$ and

$$\mathbb{P}(a_t = a_\star) = 1 - \mathbb{P}(a_t = a') - \mathbb{P}(a_t = a_o) \leq 1 - \mathbb{P}(a_t = a') = 1 - \mathbb{P}(a_t = a_\star),$$

hence the probability of selecting an optimal arm cannot exceed $1/2$. Thus, the expected regret,

$$\mathrm{Reg}_\Pi(T, \theta_\star, \mathbf{z}_{a_\star}, \mathbf{z}_{a'}, \mathbf{z}_{a_o}) \geq \left(\frac{1}{2} - \frac{1}{6}\right) \sum_{t=1}^{T} \mathbb{P}(a_t \neq a_\star) \geq \frac{T}{6},$$

which completes the proof.

□

G. Technical Lemmas

Lemma 1. *(Exponential martingale inequality) If a martingale $(\mathbf{X}_t; t \geq 0)$, adapted to filtration $\mathcal{F}_t$, satisfies $\mathbb{E}[\exp(\lambda \mathbf{X}_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda^2 \sigma_t^2 / 2)$ for some constant σ_t, for all t, then for any $a \geq 0$,*

$$\mathbb{P}\left(|X_T - X_0| \geq a\right) \leq 2 \exp\left(-\frac{a^2}{2\sum_{t=1}^{T} \sigma_t^2}\right).$$

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

$$|X_T - X_0| \leq \sqrt{2 \sum_{t=1}^{T} \sigma_t^2 \log \frac{2}{\delta}}.$$

G.1. A Hoeffding bound for Matrices

Lemma 2. *Let $\{\mathbf{M}_\tau : \tau \in [t]\}$ be a $\mathbb{R}^{d \times d}$-valued stochastic process adapted to the filtration $\{\mathcal{F}_\tau : \tau \in [t]\}$, i.e., $\mathbf{M}_\tau$ is $\mathcal{F}_\tau$-measurable for $\tau \in [t]$. Suppose that the matrix $\mathbf{M}_\tau$ is symmetric and the eigenvalues of the difference $\mathbf{M}_\tau - \mathbb{E}[\mathbf{M}_\tau | \mathcal{F}_{\tau-1}]$ lie in $[-b, b]$ for some $b > 0$. Then for $x > 0$,*

$$\mathbb{P}\left(\left\|\sum_{\tau=1}^{t} \mathbf{M}_\tau - \mathbb{E}[\mathbf{M}_\tau | \mathcal{F}_{\tau-1}]\right\|_2 \geq x\right) \leq 2d \exp\left(-\frac{x^2}{2tb^2}\right).$$

Proof. The proof adapts Hoeffding's inequality to a matrix stochastic process, following the argument of Tropp (2012). Let $\mathbf{D}_\tau := \mathbf{M}_\tau - \mathbb{E}[\mathbf{M}_\tau | \mathcal{F}_{\tau-1}]$. Then, for $x > 0$,

$$\mathbb{P}\left(\left\|\sum_{\tau=1}^{t} \mathbf{D}_\tau\right\|_2 \geq x\right) \leq \mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} \mathbf{D}_\tau\right) \geq x\right) + \mathbb{P}\left(\lambda_{\max}\left(-\sum_{\tau=1}^{t} \mathbf{D}_\tau\right) \geq x\right). \tag{33}$$

We bound the first term of Eq. (33); a similar argument applies to the second term. For any $v > 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} \mathbf{D}_\tau\right) \geq x\right) \leq \mathbb{P}\left(\exp\left\{v \lambda_{\max}\left(\sum_{\tau=1}^{t} \mathbf{D}_\tau\right)\right\} \geq e^{vx}\right) \leq e^{-vx} \mathbb{E}\left[\exp\left\{v \lambda_{\max}\left(\sum_{\tau=1}^{t} \mathbf{D}_\tau\right)\right\}\right].$$

Since $\sum_{\tau=1}^{t} \mathbf{D}_\tau$ is a real symmetric matrix,

$$\exp\left\{v \lambda_{\max}\left(\sum_{\tau=1}^{t} \mathbf{D}_\tau\right)\right\} = \lambda_{\max}\left\{\exp\left(v \sum_{\tau=1}^{t} \mathbf{D}_\tau\right)\right\} \leq \mathrm{Tr}\left\{\exp\left(v \sum_{\tau=1}^{t} \mathbf{D}_\tau\right)\right\},$$

where the last inequality holds since $\exp(v\sum_{\tau=1}^{t}\mathbf{D}_\tau)$ has nonnegative eigenvalues. Taking expectations on both sides yields

$$\begin{aligned}\mathbb{E}\left[\exp\left\{v\lambda_{\max}\left(\sum_{\tau=1}^{t}\mathbf{D}_\tau\right)\right\}\right] &\le \mathbb{E}\left[\mathrm{Tr}\left\{\exp\left(v\sum_{\tau=1}^{t}\mathbf{D}_\tau\right)\right\}\right]\\ &= \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t}\mathbf{D}_\tau\right)\right]\\ &= \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1}\mathbf{D}_\tau + \log\exp(v\mathbf{D}_t)\right)\right].\end{aligned}$$

By Lieb's theorem (Tropp, 2015), the map $\mathbf{D}\mapsto \mathrm{Tr}\exp(\mathbf{H}+\log\mathbf{D})$ is concave over the set of symmetric positive definite matrices for any symmetric positive definite $\mathbf{H}$. Applying Jensen's inequality, we obtain

$$\mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1}\mathbf{D}_\tau + \log\exp(v\mathbf{D}_t)\right)\right] \le \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1}\mathbf{D}_\tau + \log\mathbb{E}\left[\exp(v\mathbf{D}_t)\middle|\mathcal{F}_{t-1}\right]\right)\right]. \tag{34}$$

By convexity of e^{vx} and Hoeffding's lemma, for all $x\in[-b,b]$,

$$e^{vx}\le\frac{b-x}{2b}e^{-vb}+\frac{x+b}{2b}e^{vb}.$$

Since the eigenvalues of $\mathbf{D}_t$ lie within $[-b,b]$, it follows that

$$\begin{aligned}\mathbb{E}\left[\exp(v\mathbf{D}_t)\middle|\mathcal{F}_{t-1}\right] &\preceq \mathbb{E}\left[\frac{e^{-vb}}{2b}(b\mathbf{I}_d-\mathbf{D}_t)+\frac{e^{vb}}{2b}(\mathbf{D}_t+b\mathbf{I}_d)\middle|\mathcal{F}_{t-1}\right]\\ &= \frac{e^{-vb}+e^{vb}}{2}\mathbf{I}_d\\ &\preceq \exp(\frac{v^2b^2}{2})\mathbf{I}_d.\end{aligned}$$

Now we recursively upper bound Eq. (34) as follows:

$$\begin{aligned}\mathbb{E}\left[\exp\left\{v\lambda_{\max}\left(\sum_{\tau=1}^{t}\mathbf{D}_\tau\right)\right\}\right] &\le \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1}\mathbf{D}_\tau+\log\mathbb{E}\left[\exp(v\mathbf{D}_t)\middle|\mathcal{F}_{t-1}\right]\right)\right]\\ &\le \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-1}\mathbf{D}_\tau+(\frac{v^2b^2}{2})\mathbf{I}_d\right)\right]\\ &\le \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-2}\mathbf{D}_\tau+(\frac{v^2b^2}{2})\mathbf{I}_d+\log\mathbb{E}\left[\exp(v\mathbf{D}_{t-1})\middle|\mathcal{F}_{t-2}\right]\right)\right]\\ &\le \mathrm{Tr}\mathbb{E}\left[\exp\left(v\sum_{\tau=1}^{t-2}\mathbf{D}_\tau+(\frac{2v^2b^2}{2})\mathbf{I}_d\right)\right]\\ &\vdots\\ &\le \mathrm{Tr}\exp\left((\frac{tv^2b^2}{2})\mathbf{I}_d\right)\\ &= \exp\left(\frac{tv^2b^2}{2}\right)\mathrm{Tr}\left(\mathbf{I}_d\right)\\ &= d\exp\left(\frac{tv^2b^2}{2}\right).\end{aligned}$$

Thus we have

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t}\mathbf{D}_\tau\right)\geq x\right)\leq d\exp\left(-vx+\frac{tv^2b^2}{2}\right).$$

Minimizing over $v>0$ gives $v=x/(tb^2)$ and

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t}\mathbf{D}_\tau\right)\geq x\right)\leq d\exp\left(-\frac{x^2}{2tb^2}\right),$$

which proves the lemma. □

G.2. A Bound for the Gram Matrix

The Hoeffding bound for matrices (Lemma 2) implies the following bound for the two Gram matrices $\mathbf{A}_t := \sum_{\tau=1}^{t}\widetilde{\mathbf{x}}_{a_\tau}\widetilde{\mathbf{x}}_{a_\tau}^\top$ and $\mathbf{V}_t := \sum_{\tau=1}^{t}\sum_{a\in[K]}\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top$

Corollary 1. *For any $\epsilon\in(0,1)$ and $t\geq 8\epsilon^{-2}(1-p)^{-2}K^2\log(2Kt^2/\delta)$, with probability at least $1-\delta/t^2$,*

$$\left\|\mathbf{I}_K-\mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2}\right\|_2\leq\epsilon.$$

Proof. Note that

$$\mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2}-\mathbf{I}_K=\mathbf{V}_t^{-1/2}\left\{\sum_{\tau=1}^{t}\sum_{a\in[K]}\left(\frac{\mathbb{I}(\widetilde{a}_\tau=a)}{\phi_{a,\tau}}-1\right)\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top\right\}\mathbf{V}_t^{-1/2},$$

and the martingale difference matrix for each $\tau\in[t]$,

$$\begin{aligned}\left\|\sum_{a\in[K]}\left(\frac{\mathbb{I}(\widetilde{a}_\tau=a)}{\phi_{a,\tau}}-1\right)\mathbf{V}_t^{-1/2}\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top\mathbf{V}_t^{-1/2}\right\|_2&\leq\left(\sum_{a\in[K]}\left|\frac{\mathbb{I}(\widetilde{a}_\tau=a)}{\phi_{a,\tau}}-1\right|\right)\max_{a\in[K]}\left\|\mathbf{V}_t^{-1/2}\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top\mathbf{V}_t^{-1/2}\right\|_2\\&\leq\left(\frac{K-1}{1-p}+K-2\right)\max_{a\in[K]}\left\|\mathbf{V}_t^{-1/2}\widetilde{\mathbf{x}}_a\widetilde{\mathbf{x}}_a^\top\mathbf{V}_t^{-1/2}\right\|_2\\&\leq\frac{2K}{1-p}\max_{a\in[K]}\|\widetilde{\mathbf{x}}_a\|^2_{\mathbf{V}_t^{-1}}\\&\leq\frac{2K}{1-p}\cdot\frac{1}{t}.\end{aligned}$$

Note that the second inequality holds under the assumption $p\in(1/2,1)$, which implies $\phi_{a,\tau}^{-1}\leq(K-1)/(1-p)$, and the last inequality is obtained via the Sherman–Morrison formula. By Lemma 2, for $x>0$, we have

$$\mathbb{P}\left(\left\|\mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2}-\mathbf{I}_K\right\|_2>x\right)\leq 2K\exp\left(-\frac{(1-p)^2tx^2}{8K^2}\right).$$

Setting $x=\epsilon\in(0,1)$, for $t\geq 8\epsilon^{-2}(1-p)^{-2}K^2\log(2Kt^2/\delta)$ with probability at least $1-\delta/t^2$,

$$\left\|\mathbf{I}_K-\mathbf{V}_t^{-1/2}\mathbf{A}_t\mathbf{V}_t^{-1/2}\right\|_2\leq\epsilon.$$

□

G.3. An error bound for the Lasso estimator

Lemma 3 (An error bound for the Lasso estimator with unrestricted minimum eigenvalue)**.** *Let $\mathbf{x}_\tau\tau\in[t]\subset[-1,1]^d$ denote the covariates, and let $y\tau=\mathbf{x}_\tau^\top\bar{\mathbf{w}}+e_\tau$ for some $\bar{\mathbf{w}}\in\mathbb{R}^d$ and $e_\tau\in\mathbb{R}$. For $\lambda>0$, consider the estimator*

$$\widehat{\mathbf{w}}_t=\operatorname*{argmin}_{\mathbf{w}}\sum_{\tau=1}^{t}\left(y_\tau-\mathbf{x}_\tau^\top\mathbf{w}\right)^2+\lambda\|\mathbf{w}\|_1.$$

Define $\bar{\mathcal{S}} := i \in [d] : \bar{\mathbf{w}}(i) \neq 0$ *and* $\mathbf{\Sigma}_t := \sum_{\tau=1}^{t} \mathbf{x}_\tau \mathbf{x}_\tau^\top$. *Suppose that* $\mathbf{\Sigma}_t$ *has a strictly positive minimum eigenvalue and that* $\left|\sum_{\tau=1}^{t} e_\tau \mathbf{x}_\tau\right|_\infty \leq \lambda/2$. *Then,*

$$\|\widehat{\mathbf{w}}_t - \bar{\mathbf{w}}\|_{\mathbf{\Sigma}_t} \leq \frac{2\lambda\sqrt{|\bar{\mathcal{S}}|}}{\sqrt{\lambda_{\min}(\mathbf{\Sigma}_t)}}.$$

Proof. The proof follows a similar structure to Lemma B.4 in Kim et al. (2024), but we provide a new argument for the *unrestricted* minimum eigenvalue condition. Let $\mathbf{X}_t^\top := (\mathbf{x}_1, \ldots, \mathbf{x}_t) \in [-1, 1]^{d \times t}$ and $\mathbf{e}_t^\top := (e_1, \ldots, e_t) \in \mathbb{R}^t$. We denote by $\mathbf{X}_t(j)$ the j-th column of $\mathbf{X}_t$ and by $\widehat{\mathbf{w}}_t(j)$ the j-th entry of $\widehat{\mathbf{w}}_t$. By definition of $\widehat{\mathbf{w}}_t$,

$$\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t) + \mathbf{e}_t\|_2^2 + \lambda\|\widehat{\mathbf{w}}_t\|_1 \leq \|\mathbf{e}_t\|_2^2 + \lambda\|\bar{\mathbf{w}}\|_1,$$

which implies

$$\begin{aligned}
\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2^2 + \lambda\|\widehat{\mathbf{w}}_t\|_1 &\leq 2(\widehat{\mathbf{w}}_t - \bar{\mathbf{w}})^\top \mathbf{X}_t^\top \mathbf{e}_t + \lambda\|\bar{\mathbf{w}}\|_1 \\
&\leq 2\|\widehat{\mathbf{w}}_t - \bar{\mathbf{w}}\|_1 \|\mathbf{X}_t^\top \mathbf{e}_t\|_\infty + \lambda\|\bar{\mathbf{w}}\|_1 \\
&\leq \lambda\|\widehat{\mathbf{w}}_t - \bar{\mathbf{w}}\|_1 + \lambda\|\bar{\mathbf{w}}\|_1,
\end{aligned}$$

where the last inequality uses the bound on λ. On the left hand side, by triangle inequality,

$$\begin{aligned}
\|\widehat{\mathbf{w}}_t\|_1 &= \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i)| + \sum_{i \in [d] \setminus \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i)| \\
&\geq \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i)| - \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)| + \sum_{i \in [d] \setminus \bar{\mathcal{S}}} |\bar{\mathbf{w}}(i)| \\
&= \|\bar{\mathbf{w}}\|_1 - \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)| + \sum_{i \in [d] \setminus \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i)|,
\end{aligned}$$

and for the right-hand side,

$$\|\widehat{\mathbf{w}}_t - \bar{\mathbf{w}}\|_1 = \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)| + \sum_{i \in [d] \setminus \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i)|.$$

Plugging in both sides and rearranging the terms,

$$\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2^2 \leq 2\lambda \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)|. \tag{35}$$

Because $\mathbf{X}_t^\top \mathbf{X}_t$ is positive definite,

$$\begin{aligned}
\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2^2 &\geq \lambda_{\min}(\mathbf{X}_t^\top \mathbf{X}_t) \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)|^2 \\
&\geq \frac{\lambda_{\min}(\mathbf{X}_t^\top \mathbf{X}_t)}{|\bar{\mathcal{S}}|} \left(\sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)| \right)^2,
\end{aligned}$$

where the last inequality holds by Cauchy-Schwarz inequality. Plugging in Eq. (35) gives,

$$\begin{aligned}
\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2^2 &\leq 2\lambda \sum_{i \in \bar{\mathcal{S}}} |\widehat{\mathbf{w}}_t(i) - \bar{\mathbf{w}}(i)| \\
&\leq 2\lambda \sqrt{\frac{|\bar{\mathcal{S}}|}{\lambda_{\min}(\mathbf{\Sigma}_t)}} \|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2 \\
&\leq \frac{2\lambda^2 |\bar{\mathcal{S}}|}{\lambda_{\min}(\mathbf{\Sigma}_t)} + \frac{1}{2}\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2^2,
\end{aligned}$$

where the last inequality uses $ab \leq a^2/2 + b^2/2$. Rearranging the terms,

$$\|\mathbf{X}_t(\bar{\mathbf{w}} - \widehat{\mathbf{w}}_t)\|_2^2 \leq \frac{4\lambda^2 |\bar{\mathcal{S}}|}{\lambda_{\min}(\mathbf{\Sigma}_t)},$$

which proves the result. □

G.4. Eigenvalue bounds for the Gram matrix.

Lemma 4. *For $a \in [K]$, let $\widetilde{\mathbf{x}}_a := [\mathbf{x}_a^\top, \mathbf{e}_a^\top \mathbf{p}_1, \cdots, \mathbf{e}_a^\top \mathbf{p}_{K-d}]^\top \in \mathbb{R}^d$ denote augmented features. Then, an eigenvalue of $\sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top$ is in the following interval*

$$\left[\min\left\{ \lambda_{\min}\left(\sum_{a\in[k]} \mathbf{x}_a \mathbf{x}_a^\top \right), 1 \right\}, \max\left\{ \lambda_{\max}\left(\sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top \right), 1 \right\} \right].$$

Proof. Let $\mathbf{P} := (\mathbf{p}_1, \ldots, \mathbf{p}_{K-d}) \in \mathbb{R}^{K\times(K-d)}$. Because the columns in $\mathbf{P}$ are orthogonal each other and to $\mathbf{x}_1, \ldots, \mathbf{x}_K$,

$$\begin{aligned}
\sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top &= \begin{bmatrix} \sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top & \sum_{a\in[K]} \mathbf{x}_a \mathbf{e}_a^\top \mathbf{P} \\ \sum_{a\in[K]} \mathbf{P}^\top \mathbf{e}_a \mathbf{x}_a^\top & \mathbf{P}^\top \mathbf{P} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top & \sum_{a\in[K]} \mathbf{x}_a \mathbf{e}_a^\top \mathbf{P} \\ \sum_{a\in[K]} \mathbf{P}^\top \mathbf{e}_a \mathbf{x}_a^\top & I_{K-d} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top & \sum_{a\in[K]} \mathbf{X} \mathbf{e}_a \mathbf{e}_a^\top \mathbf{P} \\ \sum_{a\in[K]} \mathbf{P}^\top \mathbf{e}_a \mathbf{e}_a^\top \mathbf{X} & I_{K-d} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top & \mathbf{X}\mathbf{P} \\ \mathbf{P}^\top \mathbf{X}^\top & I_{K-d} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top & \mathbf{O} \\ \mathbf{O} & I_{K-d} \end{bmatrix}.
\end{aligned}$$

Thus, for any $\lambda \in \mathbb{R}$, $\det(\sum_{a\in[K]} \widetilde{\mathbf{x}}_a \widetilde{\mathbf{x}}_a^\top - \lambda \mathbf{I}_K) = \det(\sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top - \lambda \mathbf{I}_d)(1-\lambda)^{K-d}$. Solving $\det(\sum_{a\in[K]} \mathbf{x}_a \mathbf{x}_a^\top - \lambda \mathbf{I}_d)(1-\lambda)^{K-d} = 0$ gives the eigenvalues and the lemma is proved. □

G.5. A Bound for the Probability of Exploration

Lemma 5. *For each t, let $\widehat{a}_t = \operatorname{argmax}_{a\in[K]} \widetilde{\mathbf{x}}_a^\top \widehat{\boldsymbol{\mu}}_t$ denote the maximizing action based on the estimator $\widehat{\boldsymbol{\mu}}_t$. Then the action a_t chosen by the resampling scheme in Algorithm 1 and Algorithm 2 satisfies,*

$$\sum_{t=1}^{T} \mathbb{P}(a_t \neq \widehat{a}_t) \leq \frac{2\sqrt{T}}{(K-1)} + 2\delta.$$

Proof. Since the algorithm resamples at most ρ_t times, for a fixed $t \in [T]$,

$$\begin{aligned}
\mathbb{P}(a_t \neq \hat{a}_t) &= \mathbb{P}\left(\{\widetilde{a}_t = a_t\} \cap \{a_t \neq \hat{a}_t\}\right) + \mathbb{P}\left(\{\widetilde{a}_t \neq a_t\} \cap \{a_t \neq \hat{a}_t\}\right) \\
&\leq \mathbb{P}\left(\{\widetilde{a}_t = a_t\} \cap \{a_t \neq \hat{a}_t\}\right) + \mathbb{P}(\widetilde{a}_t \neq a_t) \\
&= \mathbb{P}\left(\{\widetilde{a}_t = a_t\} \cap \{a_t = k\}\right) + \underbrace{\mathbb{P}(\widetilde{a}_t \neq a_t)}_{\text{Failure of resampling}} \qquad \text{for } k \neq \hat{a}_t, \qquad (36)
\end{aligned}$$

where $\mathbb{P}(a_t = k) = t^{-1/2}/(K-1)$ and $\mathbb{P}(\widetilde{a}_t = a_t) = p$, defined by Algorithm 1 and Eq. (9), respectively. For the first term in Eq. (36), by applying the union bound, we obtain

$$\begin{aligned}
\mathbb{P}\left(\{\widetilde{a}_t = a_t\} \cap \{a_t = k\}\right) &= \mathbb{P}\left(\bigcup_{m=1}^{\rho_t} \{\text{Resampling success at trial } m\} \cap \{a_t = k\} \right) \\
&\leq \sum_{m=1}^{\rho_t} \mathbb{P}\left(\{\text{Resampling success at trial } m\} \cap \{a_t = k\}\right) \\
&\leq \mathbb{P}(a_t = k) \\
&= \frac{1}{\sqrt{t}(K-1)},
\end{aligned}$$

which, combined with Eq. (36), gives

$$\mathbb{P}(a_t \neq \hat{a}_t) \leq \frac{1}{\sqrt{t}(K-1)} + \mathbb{P}(\text{Resampling failure})$$

By the definition of ρ_t, the probability that the resampling fails is bounded by $\delta/(t+1)^2$. Thus, the probability of the event $\{a_t \neq \widehat{a}_t\}$ is

$$\mathbb{P}(a_t \neq \hat{a}_t) \leq \frac{1}{\sqrt{t}(K-1)} + \frac{\delta}{(t+1)^2}$$

Summing up over $t \in [T]$ completes the proof. □