Contextual Linear Bandits under Noisy Features: Towards Bayesian Oracles

Jung-hun Kim   **Se-Young Yun**[†]   **Minchan Jeong**   **Junhyun Nam**   **Jinwoo Shin**

Korea Advanced Institute of Science & Technology (KAIST), South Korea

Richard Combes

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, France

Abstract

We study contextual linear bandit problems under feature uncertainty; they are noisy with missing entries. To address the challenges of the noise, we analyze Bayesian oracles given observed noisy features. Our Bayesian analysis finds that the optimal hypothesis can be far from the underlying realizability function, depending on the noise characteristics, which are highly non-intuitive and do not occur for classical noiseless setups. This implies that classical approaches cannot guarantee a non-trivial regret bound. Therefore, we propose an algorithm that aims at the Bayesian oracle from observed information under this model, achieving $\tilde{O}(d\sqrt{T})$ regret bound when there is a large number of arms. We demonstrate the proposed algorithm using synthetic and real-world datasets.

1   INTRODUCTION

The bandit problem (Lai and Robbins, 1985) is a fundamental sequential decision-making problem when dealing with the exploration-exploitation trade-off. It has received considerable attention due to its applicability to a wide range of real-world problems, such as clinical trials (Thompson, 1933), economics (Schlag, 1998), routing (Awerbuch and Kleinberg, 2004), and ranking (Radlinski et al., 2008). In a basic multi-armed-bandit (MAB) problem, there is a finite number of actions or "arms" and in each round, an agent selects an arm and observes a random reward. The goal is to minimize regret, which is the difference in expected cumulative reward between the agent and an oracle policy that knows latent parameters.

A natural extension for the basic MAB is to provide the agent with contextual information (Langford and Zhang, 2008) that is present in many real-life problems such as personalized recommendations (Bouneffouf et al., 2012), web server defense (Jung et al., 2012), and information retrieval (Hofmann et al., 2011). In each round, the environment draws a context, and the agent observes it. Then the agent chooses an arm based on the contextual information and receives a random reward. contextual linear bandit problems include feature maps between context and arms, so each arm has a feature vector in $\mathbb{R}^d$. There is also a latent parameter in $\mathbb{R}^d$, and the mean reward for each arm follows a linear model between the latent parameter and feature vector. For contextual linear bandit problems, Auer (2002); Chu et al. (2011), and Abbasi-Yadkori et al. (2011) proposed algorithms based on the principle of optimism in the face of uncertainty.

The uncertainty of features is an important issue for many domains, including computational biology, clinical studies, and economics (Sterne et al., 2009; Troyanskaya et al., 2001; Wooldridge, 2007). Therefore, the estimation of latent parameters or learning models under noisy observations has been widely studied by Loh and Wainwright (2011); Lounici (2014); Pavez and Ortega (2020), and You et al. (2020). Recommendation systems (Li et al., 2010; Balakrishnan et al., 2018) can construct feature vectors of items by pre-processing item information such as text descriptions, categories, or figures; it is natural to have some feature noise from the pre-processing. Noisy features can even have missing entries for several reasons, including communication failure and human error. Noise can also be intentionally added to features to preserve privacy. For example, features often represent user profile information, and a recent trend in providing services that respect privacy (called differential privacy (Dwork, 2008)) is to add noise to user profile information. Therefore, noisy feature information seems natural and essential for real-world applications.

In this paper, we consider a variant of the contextual lin-

ear bandit problem in which random noise exists in the feature vectors. Here, we briefly describe these noisy features. At each time t, the true feature $z_{a,t} \in \mathbb{R}^d$ for an arm a in the set of arms is generated randomly encoding context, and the mean reward is $z_{a,t}^\top \theta^\star$ where $\theta^\star \in \mathbb{R}^d$ is a latent parameter. An agent can only observe a noise feature vector $x_{a,t} \in \mathbb{R}^d$, rather than $z_{a,t}$, which is defined as $x_{a,t} = (z_{a,t} + \varepsilon_{a,t}) \circ m_{a,t}$ where $\circ$ is the element-wise product, $\varepsilon_{a,t} \in \mathbb{R}^d$ is randomly generated from a Gaussian noise vector, and $m_{a,t} \in \{0,1\}^d$ is randomly generated from a Bernoulli distribution for missing entries, which follows the same framework as in Loh and Wainwright (2011). The missing data can be said to be missing completely at random (MCAR). To handle the noisy features, we first define an oracle policy from a Bayesian perspective given the observed noisy features. Taking insights from this oracle policy, we propose an algorithm that can achieve regret bound $\tilde{O}(d\sqrt{T})$ with respect to the feature dimension d and horizon time T.

Related work. Auer (2002) first analyzed the linear payoff model in the bandit problem. Algorithms LinRel (Auer, 2002) and LinUCB (Chu et al., 2011) compute the expected rewards and the corresponding confidence intervals to control the exploration-exploitation trade-off and achieve an $\tilde{O}(\sqrt{dT})$ regret bound. Abbasi-Yadkori et al. (2011) considered the linear bandits allowed to have infinitely many arms and proposed OFUL, which has an $\tilde{O}(d\sqrt{T})$ regret bound. However, the previous studies assumed that feature vectors are noiseless, so they cannot be directly applied to our noisy settings.

In our noisy settings, there exists a gap between true mean reward $z_{a,t}^\top \theta^\star$ and contaminated mean reward $x_{a,t}^\top \theta^\star$. We discuss some variants of contextual linear bandits for contaminated reward functions. The semi-parametric contextual bandits, where the mean reward for an arm is modeled as a linear function with a bounded confounding term that is equal for all arms, were studied by Greenewald et al. (2017); Krishnamurthy et al. (2018), and Kim and Paik (2019). The algorithm proposed by Krishnamurthy et al. (2018) achieved $\tilde{O}(d\sqrt{T})$ regret in this setting. Another variant is the misspecified setting, where the mean rewards are allowed to have at most $\epsilon \geq 0$ distance from the best-fit linear model. In this setting, Lattimore et al. (2020) achieved an $\tilde{O}(\sqrt{dT} + \epsilon T\sqrt{d})$ regret bound. However, proposed methods for the semi-parametric or misspecified contextual bandits show trivial regret bounds in our setting. This is because the contaminated terms in our setting differ for each arm at each time, and they are unbounded stochastic values.

Some recent studies have considered noise in feature vectors for contextual linear bandits. Lamprier et al. (2018) considered bounded zero-mean feature noise under the assumption that true features for each arm, which are not given to the agent, are fixed over time; they achieved an $\tilde{O}(d\sqrt{T})$ regret bound. Kirschner and Krause (2019) did not fix the true

features and considered different noise settings such that the distributions of the contexts were given to the agent each time, but the sampled contexts were hidden. They achieved an $\tilde{O}(d\sqrt{T})$ regret bound. However, we consider that true features are randomly sampled at each time without providing the true feature distribution to the agent, and only randomly sampled noisy features are observed. The noisy features are even allowed to have missing entries, making the problem more challenging.

Several previous studies have considered how to intentionally add noise to contextual linear bandits. Differential private bandit learning (Shariff and Sheffet, 2018; Zheng et al., 2020) added noise to protect privacy. The authors considered adding some noise to the matrices and vectors that contain feature information. Noise vectors are generated from a Gaussian distribution with zero mean and the identity covariance matrix. Importantly, the agent knows the noise distribution to compute an upper confidence bound, which is the main difference from our setting. The algorithms suggested in Shariff and Sheffet (2018) and Zheng et al. (2020) with knowledge of the noise distribution parameters, achieved $\tilde{O}((d + d^{3/4}\epsilon^{-1/2})\sqrt{T})$ for (ϵ, δ)-differential privacy and $\tilde{O}(T^{3/4}/\epsilon)$ for (ϵ, δ)-locally differential privacy, respectively. In our noisy feature setting, we consider more complex noise generated from a multivariate Gaussian distribution that allows non-identity covariance matrices with missing entries. Moreover, we consider that the parameters of the noise distribution are not known a priori. Adversarial attacks for linear bandits were also studied by Garcelon et al. (2020), in which an adversary attempts to confuse the agent by adding noise to rewards or features. The authors suggested a way of attacking feature vectors to prevent any algorithm from achieving a sub-linear regret bound, which is the opposite of our purpose.

This paper focuses on i.i.d. feature noise randomly generated from a multivariate Gaussian distribution with missing entries for each arm. This noise cannot be handled by previous approaches, such as the semi-parametric, misspecified, or differential private bandit models, as discussed above. Interestingly, this is the first work to handle missing entries in features for contextual linear bandits as far as we know, even though missing data has been commonly studied for various learning problems such as regression for prediction and PCA (Shang et al., 2014; Ramoni and Sebastiani, 2001; Han, 2014; Bang and Robins, 2005).

Notation. For any $A \in \mathbb{R}^{m \times m}$, we denote by A^{-1} the inverse of A when A is invertible; otherwise, denote by A^{-1} the Moore-Penrose inverse. The i-th singular value of A is denoted by $\sigma_i(A)$. For any non-negative integer m, denote by I_m the $m \times m$ identity matrix and by $\mathbf{0}_{m \times 1}$ the m-dimensional vector with all one entries. For any sets $B, C \subset \{1, ..., m\}$, we denote by $A_{B,C}$ the submatrix of A with row and column indexes in B, C, respectively. Then $A_{B,B}^{-1}$ denotes the inverse matrix of $A_{B,B}$. For any $x \in \mathbb{R}^m$,

x_B is a subvector of x with indexes in B. For any $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, denote by $[x; y] \in \mathbb{R}^{m+n}$ the concatenated vector. Lastly, we define $\|x\|_A = \sqrt{x^\top A x}$.

2 PROBLEM STATEMENT

Here we describe the contextual linear bandit models with noisy features. Let $\mathcal{A}$ be a set of arms with $|\mathcal{A}| = K$. At each time $t \in [T]$, for all $a \in \mathcal{A}$, the true feature vector $z_{a,t}$ is assumed to be i.i.d. and generated to follow the Gaussian distribution $\mathcal{N}(\nu_f, \Sigma_f)$ where $\nu_f \in \mathbb{R}^d$ and $\Sigma_f \in \mathbb{R}^{d \times d}$. The noise vector $\varepsilon_{a,t}$ is also assumed to be i.i.d. and follows the Gaussian distribution $\mathcal{N}(\mathbf{0}_{d \times 1}, \Sigma_n)$ where $\Sigma_n \in \mathbb{R}^{d \times d}$. The noise feature vector is then defined as $x_{a,t} = (z_{a,t} + \varepsilon_{a,t}) \circ m_{a,t}$ where $\circ$ is the element-wise product, $m_{a,t}$ is a masking vector such that $m_{a,t} \in \{0,1\}^d$, and each entry in $m_{a,t}$ follows a Bernoulli distribution with parameter p. We define an active set of arms $\mathcal{A}_t \subseteq \mathcal{A}$ for each time t by removing some outliers in $\mathcal{A}$. We will describe the active set in more detail soon. Then, at each time, an agent observes $\mathcal{A}$ with noisy features $x_{a,t}$ for $a \in \mathcal{A}$ and selects an arm $a_t \in \mathcal{A}_t$ and receives reward feedback $y_t = z_{a_t,t}^\top \theta^\star + \eta_t$ with a latent parameter $\theta^\star \in \mathbb{R}^d$ where noise η_t is i.i.d. and follows a σ- sub-Gaussian distribution with mean zero. We note that the agent has no a priori knowledge of the noise distribution parameters $\nu_f, \Sigma_f, \Sigma_n$, and p, true feature vectors $z_{a,t}$, and latent model parameter $\theta^\star$. The agent can observe noisy features $x_{a,t}$ for all $a \in \mathcal{A}$. For simplicity, we assume that $\|\nu_f\|_2 \le 1$, $\|\theta^\star\| \le 1$, and $\sigma = 1$. In addition, we assume that $\sigma_1(\Sigma_f + \Sigma_n)$ and $\sigma_d(\Sigma_f + \Sigma_n)$ are strictly positive constants.

The active action set $\mathcal{A}_t$ is a subset of $\mathcal{A}$ in which all outliers are removed based on observed features such that

$$\mathcal{A}_t = \left\{ a \in \mathcal{A}; \|x_{a,t}\|_2 = O\left(\sqrt{\|m_{a,t}\|_2^2 \log(KT)} \right) \right\}.$$

Since the features follow a Gaussian distribution, the mean reward for an arm may become large, which can induce large regret. Therefore, we devise the active action set $\mathcal{A}_t$ to restrict the available regret scale for our theoretical analysis, and show that $\mathcal{A}_t$ is the same as $\mathcal{A}$ for all t with high probability, to justify the scaling condition of $\mathcal{A}_t$. Since every $x_{a,t}$ is an i.i.d. Gaussian random variable, from Theorem 1 in Hsu et al. (2012) we can show that for all $a \in \mathcal{A}$ and $t \in [T]$, given $\|m_{a,t}\|_2^2$, with probability at least $1 - 1/T$, we have

$$\|x_{a,t}\|_2 = O\left(\sqrt{\|m_{a,t}\|_2^2 \log(KT)} \right),$$

which implies that $\mathcal{A}_t$ equals $\mathcal{A}$ for all $t \in [T]$ with high probability.

Objective function. The goal of this problem is to design a policy that minimizes regret over a time horizon, which

is defined as the difference between the cumulative reward from the optimal policy and the suggested policy. Therefore, it is essential to find the optimal arm at each time. In this problem, we consider a Bayesian perspective for the oracle defined on the observed information, which can be a general framework. For example, in the standard linear bandit where the true features $z_{a,t}$ are observed such that $x_{a,t} = z_{a,t}$, the standard oracle can be interpreted from a Bayesian perspective so that $a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}[z_{a,t}^\top \theta^\star | x_{b,t}; b \in \mathcal{A}] = \arg\max_{a \in \mathcal{A}_t} \mathbb{E}[z_{a,t}^\top \theta^\star | z_{b,t}; b \in \mathcal{A}] = \arg\max_{a \in \mathcal{A}_t} z_{a,t}^\top \theta^\star$. In our noisy setting, the agent can only observe noisy features for each arm. Therefore, determining the Bayesian oracle from noisy observed features is non-trivial. We assume that the oracle has complete knowledge of the latent parameters $\nu_f, \Sigma_f, \Sigma_n$, and $\theta^\star$. Then, given $x_{a,t}$ for $a \in \mathcal{A}$, the oracle's action at time t from a Bayesian point of view is defined as:

$$a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}\left[z_{a,t}^\top \theta^\star | x_{b,t}; b \in \mathcal{A} \right]$$
$$= \arg\max_{a \in \mathcal{A}_t} \mathbb{E}\left[z_{a,t}^\top \theta^\star | x_{a,t} \right],$$

where the equality comes from the independence of the observed features. We note that $\arg\max_{a \in \mathcal{A}_t} z_{a,t}^\top \theta^\star$ can be too strong for an oracle in our setting, which is not achievable in general to obtain non-trivial regret. This is because $z_{a,t}$ cannot be perfectly recovered from $x_{a,t}$ in general due to unknown independent noise $\varepsilon_{a,t}$ and missing entries for $a \in \mathcal{A}_t$ at each time t.

Adopting the Bayesian optimal arm $a_t^\star$, we define regret for a policy of selecting arm a_t at time step $t \in [T]$ as follows:

$$R(T) = \sum_{t=1}^T \mathbb{E}\left[z_{a_t^\star,t}^\top \theta^\star - z_{a_t,t}^\top \theta^\star \right].$$

Challenges. Now, we explain why it is non-trivial to find the Bayesian oracle in our setting even without missing entries where $x_{a,t} = z_{a,t} + \varepsilon_{a,t}$. The true feature is $z_{a,t} = x_{a,t} - \varepsilon_{a,t}$ so the oracle policy is $a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}\left[z_{a,t}^\top \theta^\star | x_{a,t} \right] = \arg\max_{a \in \mathcal{A}_t} \{x_{a,t}^\top \theta^\star - \mathbb{E}[\varepsilon_{a,t}^\top \theta^\star | x_{a,t}]\}$. Since $\varepsilon_{a,t}$ and $x_{a,t}$ are *not* independent, we can observe $\mathbb{E}[\varepsilon_{a,t}^\top \theta^\star | x_{a,t}] \ne 0$ in general, which means that the optimal arm is highly dependent on the distribution of $\varepsilon_{a,t}$ and $z_{a,t}$. Therefore, the Bayesian oracle $a_t^\star \ne \arg\max_{a \in \mathcal{A}_t} x_{a,t}^\top \theta^\star$ in general, which makes our problem different from standard linear bandits where $x_{a,t} = z_{a,t}$.

To illustrate this, we provide a toy example in Figure 1. We consider Gaussian distributions for $z_{a,t}$ and $\varepsilon_{a,t}$ in two-dimensional space such that $z_{a,t}$ follows a Gaussian distribution with $v_f = [0,0]$ and $\Sigma_f = [1,0;0,0]$ and $\varepsilon_{a,t}$ follows a Gaussian distribution with mean $[0,0]$ and $\Sigma_n = [0,0;0,1]$. We note that $z_{a,t}$ and $\varepsilon_{a,t}$ are independent and orthogonal to each other because $z_{a,t}$ is on the x-axis
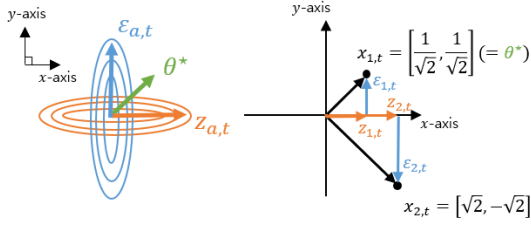
Figure 1: A toy example

and $\varepsilon_{a,t}$ is on the y-axis with probability 1. We assume that there are two arms, $\{1, 2\}$, and the observed features are $x_{1,t} = [1/\sqrt{2}, 1/\sqrt{2}]$ and $x_{2,t} = [\sqrt{2}, -\sqrt{2}]$, respectively. We consider $\theta^\star = [1/\sqrt{2}, 1/\sqrt{2}]$, which points in the same direction as $x_{1,t}$. Then in the standard linear bandits where the observed feature is the true feature ($x_{a,t} = z_{a,t}$), arm 1 is the optimal arm from the fact that $\mathbb{E}[z_{1,t}^\top \theta^\star | x_{1,t}] = x_{1,t}^\top \theta^\star = 1$ and $\mathbb{E}[z_{2,t}^\top \theta^\star | x_{2,t}] = x_{2,t}^\top \theta^\star = 0$. However, this is not the same in our noisy setting where $x_{a,t} = z_{a,t} + \varepsilon_{a,t}$. Since $z_{a,t}$ is on the x-axis and $\varepsilon_{a,t}$ is on the y-axis, we can estimate $z_{a,t}$ by projecting $x_{a,t}$ on the x-axis. Therefore, we obtain $\mathbb{E}[z_{1,t}|x_{1,t}] = [1/\sqrt{2}, 0]$ and $\mathbb{E}[z_{2,t}|x_{2,t}] = [\sqrt{2}, 0]$. Then the Bayesian optimal arm is arm 2 because $\mathbb{E}[z_{1,t}^\top \theta^\star | x_{1,t}] = 1/2$ and $\mathbb{E}[z_{2,t}^\top \theta^\star | x_{2,t}] = 1$. From this example, we can observe that our noisy problem can be totally different from the standard linear bandits where the optimal action is directly determined by observed features. In addition, considering missing entries increases the difficulty of the problem.

As shown in the toy example, we can observe that the optimal arm is highly dependent not only on $x_{a,t}$ and $\theta^\star$, but also on the distribution of $z_{a,t}$ and $\varepsilon_{a,t}$. Therefore, the first difficulty is deriving the Bayesian oracle policy. Then, the second difficulty is learning this oracle in a bandit setting where the latent parameters are unknown.

Contribution. First, we analyze the simple case with $p = 1$ where there are no missing entries in the observed noisy features. Analyzing the Bayesian oracle strategy, we show that slightly modified `OFUL` (Abbasi-Yadkori et al., 2011) can achieve an $\tilde{O}(d\sqrt{T})$ regret bound. Then we dive into the more complicated case with missing entries. We show that the Bayesian oracle strategy requires solving non-linear programming in this case. Based on the insight, we design an algorithm that can achieve $\tilde{O}(d\sqrt{T})$ when K is large. Lastly, we examine our algorithm using synthetic and real-world datasets and observe that the results are consistent with our theoretical analysis.

3 WARM-UP: NOISY FEATURES WITHOUT MISSING ENTRIES

In this section, we analyze the simple case of $p = 1$ where the observed features have no missing entries but only Gaus-

sian noise. First, we analyze the oracle strategy and then provide an algorithm with regret analysis.

Bayesian oracle strategy. We examine the strategy of a Bayesian oracle that has complete knowledge of latent parameters $\nu_f, \Sigma_f, \Sigma_n$, and $\theta^\star$ in the following proposition.

Proposition 1. *Define:*

$$\theta' := (\Sigma_f + \Sigma_n)^{-1}\Sigma_f \theta^\star \text{ and } \overline{\theta} := [\nu_f^\top \theta^\star - \nu_f^\top \theta'; \theta'].$$

Then the Bayesian oracle policy under the noisy features without missing entries selects arm $a_t^\star$ at time step $t \in [T]$ such that:

$$a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}[z_{a,t}^\top \theta^\star | x_{a,t}] = \arg\max_{a \in \mathcal{A}_t}[1; x_{a,t}]^\top \overline{\theta}$$
$$= \arg\max_{a \in \mathcal{A}_t} x_{a,t}^\top \theta'.$$

Proof. At each time step t, given noisy feature vectors, the Bayesian optimal decision is expressed as:

$$a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}[z_{a,t}^\top \theta^\star | x_{b,t}; \forall b \in \mathcal{A}]$$
$$= \arg\max_{a \in \mathcal{A}_t} \mathbb{E}[z_{a,t}|x_{a,t}]^\top \theta^\star.$$

From the Bayesian analysis of multivariate Gaussian random variables (Section 3.1 in Smith and Winkler (2006)), we can show that for any $a \in \mathcal{A}_t$,

$$\mathbb{E}[z_{a,t}|x_{a,t}]^\top \theta^\star = (x_{a,t} - \nu_f)^\top (\Sigma_f + \Sigma_n)^{-1}\Sigma_f \theta^\star + \nu_f^\top \theta^\star$$
$$= x_{a,t}^\top \theta' - \nu_f^\top \theta' + \nu_f^\top \theta^\star = [1; x_{a,t}]^\top \overline{\theta},$$
$$(1)$$

which concludes the proof. $\square$

From Proposition 1, we can observe that $\mathbb{E}[\varepsilon_{a,t}^\top \theta^\star | x_{a,t}] = x_{a,t}^\top \theta^\star - \mathbb{E}[z_{a,t}^\top \theta^\star | x_{a,t}] = x_{a,t}^\top \theta^\star - [1; x_{a,t}]^\top \overline{\theta} \neq 0$ in general, as we discussed in the previous section, which makes the problem different from the standard contextual linear bandits. Although $\mathbb{E}[\varepsilon_{a,t}^\top \theta^\star | x_{a,t}] \neq 0$, we can observe that interestingly, the noisy contextual bandit without missing entries still models the linear reward with respect to the modified contextual vector $[1; x_{a,t}]$, because the mean reward of an arm a given the observed feature is $\mathbb{E}[z_{a,t}^\top \theta^\star | x_{a,t}] = [1; x_{a,t}]^\top \overline{\theta}$ from (1). Thus, the regret of an algorithm selecting arm a_t at time step $t = 1, \dots, T$ can be expressed as:

$$R(T) = \sum_{t=1}^{T} \mathbb{E}\left[z_{a_t^\star,t}^\top \theta^\star - z_{a_t,t}^\top \theta^\star\right]$$
$$= \sum_{t=1}^{T} \mathbb{E}\left[[1; x_{a_t^\star,t}]^\top \overline{\theta} - [1; x_{a_t,t}]^\top \overline{\theta}\right].$$

Algorithm and regret analysis. From (1), the problem can be converted to estimating $\overline{\theta}$ using observed features

$[1; x_{a,t}]$ and reward. The observed reward can be considered as *Bayesian reward* whose mean reward is $[1; x_{a_t,t}]^\top \overline{\theta}$ given $x_{a,t}$'s. From the insight of the Bayesian analysis, we suggest using the strategy OFUL (Abbasi-Yadkori et al., 2011) that was proposed for standard contextual linear bandits. The algorithm is based on the principle of optimism in the face of uncertainty for reward estimation. To build the confidence set to estimate $\overline{\theta}$ in our setting, we consider $\|x_{a,t}\|_2 = O(\sqrt{d \log(KT)})$ for $a \in \mathcal{A}_t$ and $\|\overline{\theta}\|_2 \leq C_2$ for some constant $C_2 > 0$ and tune the bias term $\lambda > 0$ to control the variance of the feature vector. Define the confidence set as follows:

$$\mathcal{C}_t = \{\theta \in \mathbb{R}^{d+1} :$$
$$\|\widehat{\theta}_t - \theta\|_{V_t} \leq C_1 \sqrt{(d+1)\log((1+t)T)} + C_2 \lambda^{1/2}\},$$

where bias term $\lambda = C_3 d \log(KT)$, $V_t = \lambda I_{d+1} + \sum_{s=1}^t [1; x_{a_t,t}][1; x_{a_t,t}]^\top$, and $\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^t [1; x_{a_t,t}] y_t$ for some constants $C_1, C_3 > 0$. We provide a pseudocode of the algorithm in Appendix A.1.

The following proposition shows that OFUL can achieve a tight regret bound in the noisy setting without missing entries.

Proposition 2. OFUL *with* $\mathcal{C}_{t-1}$ *and modified feature vectors* $[1; x_{a,t}]$ *for all* $a \in \mathcal{A}$ *and* $t \in [T]$ *can guarantee* $R(T) = \tilde{O}(d\sqrt{T})$.

Following the proof steps of Theorem 3 in Abbasi-Yadkori et al. (2011) easily reveals the regret bound. We prove Proposition 2 in Appendix A.2.

Tightness of the regret bound. The result in Proposition 2, $\tilde{O}(d\sqrt{T})$, matches the regret lower bound for standard linear bandits, $\Omega(d\sqrt{T})$ (Li et al., 2010), up to poly-logarithmic factors, although our regret is defined based on the Bayesian oracle, unlike previous bandit literature. For the regret lower bound of the standard model, we additionally consider that the norms of true features are bounded by $\tilde{O}(\sqrt{d})$ as in our setting, which scales the regret $\sqrt{d}$ times more compared to the original form in Li et al. (2010). Furthermore, our regret bound is comparable to the results of previous work such as the achieved $\tilde{O}(d\sqrt{T})$ regret bound considering the contaminated mean reward (Krishnamurthy et al., 2018), or $\tilde{O}(d\sqrt{T})$ (Lamprier et al., 2018; Kirschner and Krause, 2019; Shariff and Sheffet, 2018) regret bounds considering feature noise. However, it is an open problem to obtain the problem-specific regret lower bound for our noisy setting under the Bayesian perspective.

4 NOISY FEATURES WITH MISSING ENTRIES

Now, we move on to our main contribution. Here, we consider a case where every component of each observed feature vector has been erased with *missing probability*

$1 - p \in (0, 1)$. Each component of the feature vector is masked with a random variable following a Bernoulli distribution with parameter p. Similar to the previous section, we first derive the Bayesian oracle strategy and suggest an algorithm with regret analysis.

Bayesian oracle strategy. Missing entries affect the information available to the oracle. The Bayesian oracle maximizes the expected reward, knowing the observed feature vectors with missing entries. Given a feature vector $x \in \mathbb{R}^d$, we define $\mathcal{S}(x)$ as the indexes for non-missing entries and $\mathcal{U}(x)$ as the indexes for missing entries. For simplicity, we use $\mathcal{S}$ for $\mathcal{S}(x)$ and $\mathcal{U}$ for $\mathcal{U}(x)$ if there is no confusion. Then, the Bayesian oracle strategy arises from the following theorem.

Theorem 1. *For any* $x \in \mathbb{R}^d, \nu_f \in \mathbb{R}^d$, *and* $\Sigma \in \mathbb{R}^{d \times d}$, *define* $\overline{x}(\nu, \Sigma, x) \in \mathbb{R}^d$ *as:*

$$\overline{x}(\nu, \Sigma, x)_\mathcal{S} := x_\mathcal{S} \quad and$$
$$\overline{x}(\nu, \Sigma, x)_\mathcal{U} := \nu_\mathcal{U} + \Sigma_{\mathcal{U},\mathcal{S}} \Sigma_{\mathcal{S},\mathcal{S}}^{-1} (x - \nu)_\mathcal{S}.$$

Then the Bayesian oracle strategy under the noisy features with missing entries selects arm $a_t^\star$ *at time step* $t \in [T]$ *such that:*

$$a_t^\star \in \operatorname*{argmax}_{a \in \mathcal{A}_t} \mathbb{E}[z_{a,t}^\top \theta^\star | x_{a,t}]$$
$$= \operatorname*{argmax}_{a \in \mathcal{A}_t} [1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})]^\top \overline{\theta}$$
$$= \operatorname*{argmax}_{a \in \mathcal{A}_t} \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})^\top \theta'.$$

We note that when $S = [d]$, we observe $\overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t}) = x_{a,t}$, which is the same as in the case of noisy features without missing entries. When $S = \emptyset$, we observe $\overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t}) = \nu_f$, which implies that the oracle estimates true features from the mean value directly because the oracle does not have noisy feature information.

Proof. Denote by $x'_{a,t}$ the noisy feature vector for arm a without masking such that $x'_{a,t} = z_{a,t} + \varepsilon_{a,t}$. Then we observe that $x'_{a,t}$ follows the Gaussian distribution with mean ν_f and covariance matrix $\Sigma_f + \Sigma_n$. Then by using the conditional Gaussian distribution from Proposition 3.13 in Eaton (1983), we can show that for all $a \in \mathcal{A}_t$,

$$\mathbb{E}[x'_{a,t} | x_{a,t}] = \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t}), \qquad (2)$$

which is the estimated feature vector for recovering missing values. Thus, from (1), (2), and $\overline{\theta} = [\nu^\top \theta^\star - \nu^\top \theta'; \theta']$, the expected reward given the observed feature vector can be expressed as

$$\mathbb{E}[z_{a,t}^\top \theta^\star | x_{a,t}] = \mathbb{E}\left[\mathbb{E}\left[z_{a,t}^\top \theta^\star | x'_{a,t}\right] | x_{a,t}\right]$$
$$= \mathbb{E}\left[(x'_{a,t} - \nu_f)^\top \theta' + \nu_f^\top \theta^\star | x_{a,t}\right]$$
$$= \mathbb{E}[x'_{a,t} | x_{a,t}]^\top \theta' + \nu_f^\top (\theta^\star - \theta')$$
$$= \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})^\top \theta' + \nu_f^\top (\theta^\star - \theta')$$
$$= [1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})]^\top \overline{\theta}, \qquad (3)$$

Algorithm 1 Contextual Linear Bandits on Bayesian Features (CLBBF)

Initialize: $Z \leftarrow \mathbf{0}_{d \times d}, \xi \leftarrow \mathbf{0}_{d \times 1}, n \leftarrow 0, i \leftarrow 1$.
$n \leftarrow n +$ the total number of non-missing entries in $x_{a,1}$ for all $a \in \mathcal{A}$.
$Z \leftarrow Z + \sum_{a \in \mathcal{A}} x_{a,1} x_{a,1}^\top, \xi \leftarrow \xi + \sum_{a \in \mathcal{A}} x_{a,1}$.
Select a_1 uniformly at random in $\mathcal{A}_1$.
Observe reward y_1.
for $t = 2$ **to** T **do**
  $n \leftarrow n +$ the total number of non-missing entries in $x_{a,t}$ for all $a \in \mathcal{A}$.
  $Z \leftarrow Z + \sum_{a \in \mathcal{A}} x_{a,t} x_{a,t}^\top, \xi \leftarrow \xi + \sum_{a \in \mathcal{A}} x_{a,t}$.
  Estimate parameters:
  $\widehat{p} \leftarrow \frac{\max\{1, n\}}{tdK}, \quad \widehat{\nu} \leftarrow \frac{1}{tK\widehat{p}} \xi$,
  $\widehat{\Sigma} \leftarrow \frac{1}{tK} Z \circ \left(\frac{\widehat{p}-1}{\widehat{p}^2} I_{d \times d} + \frac{1}{\widehat{p}^2} \mathbf{1}_{d \times d} \right) - \widehat{\nu} \widehat{\nu}^\top$.
  Estimate Bayesian features:
  $\widehat{z}_{a,t} \leftarrow [1; \overline{x}(\widehat{\nu}, \widehat{\Sigma}, x_{a,t})]$ for $a \in \mathcal{A}_t$.
  if $t = 2^i$ **then**
    Update selected Bayesian features:
    $\widehat{z}_{a_s,s} \leftarrow [1; \overline{x}(\widehat{\nu}, \widehat{\Sigma}, x_{a_s,s})]$ for all $s \in [t-1]$.
    $i \leftarrow i + 1$.
  end if
  Select $a_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_{t-1}} \langle \widehat{z}_{a,t}, \theta \rangle$.
  Observe reward y_t.
end for

---

which concludes the proof. $\qquad\square$

From Theorem 1, the Bayesian oracle strategy selects an arm based on $\overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})$, which implies that the model is non-linear with respect to the observed feature $x_{a,t}$. This is a significant difference from the simple case of noisy features without missing entries. From the oracle strategy in Theorem 1 and (3), we can show that regret can be expressed as:

$$R(T) = \sum_{t=1}^{T} \mathbb{E} \left[z_{a_t^\star, t}^\top \theta^\star - z_{a_t, t}^\top \theta^\star \right]$$

$$= \sum_{t=1}^{T} \mathbb{E} \left[[1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a_t^\star, t})]^\top \overline{\theta} \right.$$
$$\left. - [1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a_t, t})]^\top \overline{\theta} \right].$$

Algorithms and regret analysis. From Theorem 1, the oracle policy is not linear to the observed feature vectors, which implies that we cannot naively use OFUL to guarantee a sublinear regret bound when the distribution latent parameters are not given to the algorithm. Therefore, we propose Algorithm 1, which includes a procedure to estimate *Bayesian features* $\mathbb{E}[z_{a,t}|x_{a,t}] = [1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})]$.

Algorithm 1 estimates the parameters of the distribution for the feature vectors from the information of the ob-

served features at each time. Using the estimated distribution parameters, the algorithm estimates the Bayesian feature $\mathbb{E}[z_{a,t}|x_{a,t}] = [1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a,t})]$ as $\widehat{z}_{a,t}$ for all $a \in \mathcal{A}_t$. Furthermore, the algorithm rarely updates the estimated features for previously chosen arms using the current estimated distribution parameters while reducing the computation cost. Finally, it takes a strategy based on the principle of optimism in the face of uncertainty using estimated feature vectors $\widehat{z}_{a,t}$. The algorithm selects a_t and observes reward y_t for each time t. We note that the algorithm considers the observed reward as Bayesian reward whose mean reward is $[1; \overline{x}(\nu_f, \Sigma_f + \Sigma_n, x_{a_t, t})]^\top \overline{\theta}$ from (3).

The necessary notations for Algorithm 1 are defined as follows. First we define $\lambda = C_4 d \log(KT)$, $V_t = \lambda I_{d+1} + \sum_{s=1}^{t} \widehat{z}_{a_s, s} \widehat{z}_{a_s, s}^\top$, $\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} \widehat{z}_{a_s, s} y_s$, and $X_t = \sum_{s=1}^{t} \|\widehat{z}_{a_s, s}\|_{V_t^{-1}}$ for some large constant $C_4 > 0$. Note that $\|\overline{\theta}\|_2 \leq C_2$ for some constant $C_2 > 0$ and that the bias term λ is tuned to control the variance of the observed feature norms. In addition, note that $\widehat{p}$ is an estimator for p at each time in the algorithm. Then, we define the confidence set for estimating a latent parameter $\overline{\theta}$ as

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^{d+1} : \|\widehat{\theta}_t - \theta\|_{V_t} \leq C_5 \sqrt{(d+1) \log((1+t)T)} \right.$$
$$\left. + C_2 \lambda^{1/2} + C_6 (d/\widehat{p})^{3/2} \sqrt{\frac{\log(KT)}{K}} X_t \right\},$$
(4)

for some large constants $C_5, C_6 > 0$. The estimation error of Bayesian features incurs that Algorithm 1 requires an additional confidence bound term, $C_6 (d/\widehat{p})^{3/2} \sqrt{\log(KT)/K} X_t$, in (4) compared to OFUL. By solving the convex problem to find a_t in Algorithm 1, as stated in Section 19.3 in Lattimore and Szepesvári (2020), the action in each round can be simply calculated as

$$a_t = \underset{a \in \mathcal{A}_t}{\operatorname{argmax}} \left\{ \widehat{z}_{a,t}^\top \widehat{\theta}_{t-1} + \left(C_5 \sqrt{(d+1) \log(tT)} + C_2 \lambda^{1/2} \right. \right.$$
$$\left. \left. + C_6 (d/\widehat{p})^{3/2} \sqrt{\frac{\log(KT)}{K}} X_{t-1} \right) \sqrt{\widehat{z}_{a,t}^\top V_{t-1}^{-1} \widehat{z}_{a,t}} \right\}.$$

Then, the following theorem provides regret for the algorithm.

Theorem 2. *Algorithm 1 achieves the regret bound of*

$$R(T) = \tilde{O} \left(d\sqrt{T} + \frac{d^2}{p^{3/2}} \sqrt{\frac{T}{K}} + \frac{d}{p^4 K} \right).$$

The first term for the regret bound in Theorem 2 comes from the analysis of optimism in the face of uncertainty, and the second and third terms come from the estimation error for Bayesian features, which is the main difference

from standard linear bandits where the true feature vectors are given. We can note that the second term dominates over the third one when T is large. An additional factor d compared to Proposition 2 comes from the estimation error of Bayesian features, and factor $1/p^{3/2}$ mainly comes from estimation error of $\widehat{\Sigma}$. From Theorem 2, when $K = \Omega(\max\{d^2/p^3, 1/(\sqrt{T}p^4)\})$ we get

$$R(T) = \tilde{O}(d\sqrt{T}).$$

Therefore, the regret becomes insensitive to p when the algorithm has sufficient observed feature information to estimate Bayesian features accurately. We provide a further discussion about the tightness of the bound with respect to p in Appendix A.4. We also note that by using the well-known doubling trick, we can run our algorithm without information of T a priori.

Proof sketch. We provide a proof sketch here; Appendix A.3 contains the full version. Let $\overline{z}_{a,t} = [1; \overline{x}(\nu, \Sigma_f + \Sigma_n, x_{a,t})]$ for simplicity, and instantaneous regret $r_t = \overline{z}_{a_t^\star,t}^\top \overline{\theta} - \overline{z}_{a_t,t}^\top \overline{\theta}$. For $s \leq t$, we write $\widehat{\nu}_t$, $\widehat{\Sigma}_t$, $\widehat{z}_{a,s}(t)$, and $\widehat{p}_t$ for $\widehat{\nu}$, $\widehat{\Sigma}$, $\widehat{z}_{a,s}$, and $\widehat{p}$ at time step t, respectively. From the condition for the active set, we have $\|\overline{z}_{a,t}\|_2 = O(\sqrt{d\log(KT)})$ for all $a \in \mathcal{A}_t$. The confidence set $\mathcal{C}_t$ is designed on this constraint for $\overline{z}_{a,t}$. However, the algorithm selects an arm based on estimated features $\widehat{z}_{a,t}(t)$ rather than $\overline{z}_{a,t}$. Hence, the first few steps require collecting feature information to estimate $\overline{z}_{a_s,s}$ as $\widehat{z}_{a_s,s}(t)$ to satisfy $\|\widehat{z}_{a_s,s}(t)\|_2 = O(\sqrt{d\log(KT)})$. Let $\tau = 2\lceil d(\log(T))^2/(Kp^4)\rceil$. Then, we show that for all $t > \tau$ and $1 \leq s \leq t$, with high probability we have

$$\|\widehat{z}_{a_s,s}(t)\|_2 = O\left(\sqrt{d\log(KT)}\right). \tag{5}$$

To obtain a regret bound before τ, we use the fact that x_a follows a Gaussian distribution given $m_{a,t}$. Using a bound for an expected maximum Gaussian variable, we can show that

$$\mathbb{E}\left[\overline{z}_{a_t^\star,t}^\top \overline{\theta} - \overline{z}_{a_t,t}^\top \overline{\theta}\right] = O\left(\sqrt{\log(K)}\right). \tag{6}$$

Then from (6), the regret bound for the first τ time steps can be obtained as

$$\sum_{t=1}^{\tau} \mathbb{E}\left[r_t\right] = O\left(\tau\sqrt{\log(K)}\right) = \tilde{O}\left(d/(Kp^4)\right). \tag{7}$$

Let $\tilde{\theta}_t = \arg\max_{\theta \in \mathcal{C}_{t-1}} \max_{a \in \mathcal{A}_t} \langle \widehat{z}_{a,t}(t), \theta \rangle$. Now, we analyze regret after the time step τ. Our algorithm selects an arm a_t based on the estimated features $\widehat{z}_{a,t}(t)$ and observes y_t whose mean reward is $\overline{z}_{a_t,t}^\top \overline{\theta}$ from the Bayesian view. Therefore, regret is influenced by both estimation errors for $\tilde{\theta}_t$ and $\widehat{z}_{a,t}(t)$, which is the main difference from the standard linear bandits in regret analysis. First, we try to

separate regret by the estimation errors; we show that with high probability,

$$r_t \leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \|\overline{z}_{a_t,t} - \widehat{z}_{a_t,t}(t)\|_2 \|\overline{\theta}\|_2$$
$$+ \|\overline{\theta} - \tilde{\theta}_t\|_{V_{t-1}} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}, \tag{8}$$

where the instantaneous regret bound comprises the estimation errors for $\widehat{z}_{a_t^\star,t}$, $\widehat{z}_{a_t,t}(t)$, and $\tilde{\theta}_t$. To obtain the error bounds for $\|\overline{z}_{a,t} - \widehat{z}_{a,t}(t)\|_2$ and $\|\overline{\theta} - \tilde{\theta}_t\|_2$, we consider the estimation errors for $\widehat{\nu}_t$ and $\widehat{\Sigma}_t$. Using vector and matrix concentration inequalities, we can show that with high probability

$$\|\overline{z}_{a,t} - \widehat{z}_{a,t}(t)\|_2 = O\left((d/p^{3/2})\sqrt{\log(KT)\log(T)/(tK)}\right). \tag{9}$$

From (5), (8), (9), and since $\widehat{p}_t = \Theta(p)$ with high probability from Hoeffding's inequality, we can show that $\overline{\theta} \in \mathcal{C}_{t-1}$ with high probability. In addition, we show that $X_{t-1} = O\left(\sqrt{td\log T}\right)$. Then, with high probability, we have

$$\|\overline{\theta} - \tilde{\theta}_t\|_{V_{t-1}} = O\left(\sqrt{d\log(tKT)}\right.$$
$$\left. + (d/p)^{3/2}\log(T)\sqrt{\log(KT)/K}\right). \tag{10}$$

From (8), (9), and (10), we have

$$r_t = \tilde{O}\left(\left(\sqrt{d} + (d/p)^{3/2}\sqrt{1/K}\right)\|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}\right.$$
$$\left. + (d/p^{3/2})\sqrt{1/(tK)}\right). \tag{11}$$

Using (5) and Lemma 11 in Abbasi-Yadkori et al. (2011), we can show that with high probability

$$\sum_{t=\tau+1}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2 = O\left(d\log(T\log(KT))\log(T)\right). \tag{12}$$

Finally, with (7), (11), and (12), using the Cauchy-Schwarz inequality, we have

$$\sum_{t=1}^{T} \mathbb{E}[r_t] = \tilde{O}\left(d\sqrt{T} + (d^2/p^{3/2})\sqrt{T/K} + d/(Kp^4)\right),$$

which concludes the proof. $\qquad\square$

Algorithm efficiency. Algorithm 1 rarely updates the estimated features for previously chosen arms, achieving the $O(T\log(T))$ computation cost over the horizon T. The update procedure requires $O(t)$ memory space for each time step t. This update is required to obtain well-estimated features for the previously chosen arms, which are then used to obtain V_t, $\widehat{\theta}_t$, and X_t. We propose a more efficient algorithm
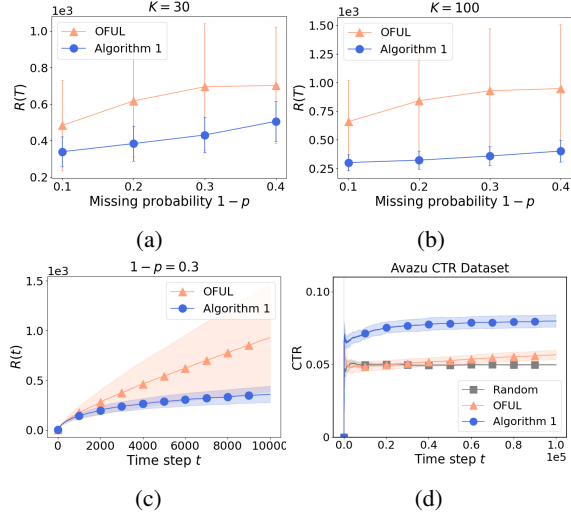
Figure 2: Performance of Algorithm 1 and OFUL: $R(T)$ versus missing probability $1-p$ with (a) $K=30$ and (b) $K=100$ in synthetic datasets; and $R(t)$ or CTR versus time step t in (c) synthetic and (d) real-world datasets.

(Algorithm 3 in Appendix A.5) to reduce the computation cost and the required memory space for Algorithm 1, achieving the $O(T)$ computation cost and $O(1)$ memory space. However, the algorithm requires information of $\alpha > 0$ that satisfies $\alpha \leq p$. We can show that this algorithm achieves $R(T) = \tilde{O}\left(d\sqrt{T} + (d^2/p^{3/2})\sqrt{T/K} + d/(\alpha^4 K)\right)$. Appendix A.5 contains a more detailed explanation and regret analysis of the algorithm.

5 NUMERICAL EXPERIMENTS

In this section, we present numerical results of experiments for synthetic and real-world datasets.[1] We repeat each experiment 10 times. First, we describe the experimental setting for synthetic datasets. For the latent parameters, we generate $\theta^\star$ and ν_f from the uniform distribution in $[0, 1]$ and normalize them using 2-norm. Then, we generate random matrices $A, B \in [0, 1]^{d \times d}$ from the uniform distribution and construct the covariance matrices as $\Sigma_f = A^\top A$ and $\Sigma_n = B^\top B$ with spectral norm normalization. Feature vectors are generated from $\mathcal{N}(\nu_f, \Sigma_f)$, noise vectors are generated from $\mathcal{N}(0, \Sigma_n)$, and missing entries are assigned following a Bernoulli distribution with probability p. Using the distributions, we generate K number of noisy feature vectors for each time. The reward noise η_t is generated from $\mathcal{N}(0, 1)$. We set $d = 2$ and $T = 10^4$.

For real-world experiments, we use Avazu click-through rate (CTR) dataset (Avazu, 2014) that contains 'clicked or

not' information for each advertisement recommendation. We use an autoencoder model to preprocess ad-user pair feature information for constructing item feature vectors. We set preprocessed feature dimension $d = 32$, $K = 20$, and $T = 10^5$. We erase some entries in the preprocessed feature vectors with missing probability $1 - p = 0.1$. A more detailed explanation is deferred to Appendix A.6.

We compare the performance of Algorithm 1 with OFUL described in Section 3. We first examine cumulative regret over the time horizon T for each algorithm in synthetic datasets varying $1 - p$ from 0 to 0.4, where the number of missing entries is likely to increase as the missing probability increases. Figure 2 (a,b) confirms that our algorithm performs better than OFUL for various $1 - p$. In Figure 2 (a) when $K = 30$, the regret of our algorithm increases rapidly as the missing probability increases, while in Figure 2 (b) when $K = 100$, we can observe that our algorithm becomes more robust to the missing probability. This result is consistent with Theorem 2, which shows that our algorithm is insensitive to p with $R(T) = \tilde{O}(d\sqrt{T})$ when K is sufficiently large. Furthermore, we can observe that the regret variance for Algorithm 1 is smaller than that for OFUL.

Next, in Figure 2 (c), we examine the cumulative regret over time steps $t \in [T]$ when $1 - p = 0.3$ and $K = 100$ in synthetic datasets. We can observe that Algorithm 1 achieves much better performance with sub-linearly increasing regret with low variance, whereas OFUL exhibits almost linearly increasing regret with large variance. In Figure 2 (d), we demonstrate that our algorithm achieves higher CTRs than OFUL in the real-world dataset. Further experiments for various real-world datasets are provided in Appendix A.6.

6 DISCUSSION ON EXTENSIONS

In our setting, we consider that the features of the arms are generated from the same Gaussian distribution. However, our method can be easily extended to the case where the features of the arms are generated from different Gaussian distributions with each other. In such a case, the algorithm needs to estimate latent distribution parameters for each arm. Then, with a slight modification to our algorithm, we can achieve $R(T) = \tilde{O}((d^2/p^{3/2})\sqrt{T} + d/p^4)$. In addition, in our setting, we consider that the features of arms at the same time are independent. We can extend our setting to the dependent case such that the K number of features is generated from a Gaussian distribution with parameters $\nu_f \in \mathbb{R}^{Kd}$ and $\Sigma_f \in \mathbb{R}^{Kd \times Kd}$ having a Gaussian noise with $\Sigma_n \in \mathbb{R}^{Kd \times Kd}$. With a slight modification of our algorithm, we can achieve $R(T) = \tilde{O}(((Kd)^2/p^{3/2})\sqrt{T} + (dK)/p^4)$.

In our setting, we consider the Gaussian conjugate distribution of features for Bayesian analysis. It would be of interest to extend our setting to other conjugate distributions.

---

[1]Our code is available at https://github.com/junghunkim7786/contextual_linear_bandits_under_noisy_features

7 CONCLUSION

In this paper, we studied contextual linear bandits under noisy features that contain Gaussian noise and missing entries. We analyzed Bayesian oracles and subsequently proposed an algorithm to achieve $\tilde{O}(d\sqrt{T})$ when the number of arms is large. Lastly, we demonstrated the performance of our algorithm using synthetic and real-world datasets.

Limitation: We leave several open questions. For instance, extensions to other distributions beyond the Gaussian distribution for Bayesian analysis would be of interest. In addition, obtaining a regret lower bound would be useful in understanding the fundamental limitations of our model.

Potential negative societal impacts: This study focuses on theoretical analysis; therefore, we could not see any negative social consequences.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Alibaba. Taobao.com dataset. https://tianchi.aliyun.com/dataset/dataDetail?dataId=56&lang=en-us, 2018. This dataset is provided by Alibaba and distributed under the CC BY-NC-SA 4.0 license.

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

Avazu. Avazu ctr dataset. https://www.kaggle.com/competitions/avazu-ctr-prediction/data, 2014. This dataset is provided by Avazu for KDD competition 'Click-Through Rate Prediction'.

Baruch Awerbuch and Robert D Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 45–53. ACM, 2004.

Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *IJCAI*, pages 5802–5804, 2018.

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

Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *International Conference on Neural Information Processing*, pages 324–331. Springer, 2012.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pages 208–214, 2011.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation—TAMC*, volume 4978, pages 1–19. Springer Verlag, April 2008. URL https://www.microsoft.com/en-us/research/publication/differential-privacy-a-survey-of-results/.

Morris L Eaton. Multivariate statistics: A vector space approach. *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 512*, 1983.

Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Jean Tarbouriech, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, and Predag Klasnja. Action centered contextual bandits. *Advances in neural information processing systems*, 30, 2017.

Peisong Han. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109(507):1159–1173, 2014.

F Maxwell Harper and Joseph A Konstan. The movie-lens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015. https://grouplens.org/datasets/movielens/100k/.

Katja Hofmann, Shimon Whiteson, Maarten de Rijke, et al. Contextual bandits for information retrieval. In *NIPS 2011 Workshop on Bayesian Optimization, Experimental Design, and Bandits, Granada*, volume 12, page 2011, 2011.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Tobias Jung, Sylvain Martin, Damien Ernst, and Guy Leduc. Contextual multi-armed bandits for web server defense.

In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE, 2012.

Gi-Soo Kim and Myunghee Cho Paik. Contextual multi-armed bandit algorithm for semiparametric reward model. In *International Conference on Machine Learning*, pages 3389–3397, 2019.

Johannes Kirschner and Andreas Krause. Stochastic bandits with context distributions. In *Advances in Neural Information Processing Systems*, pages 14090–14099, 2019.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785, 2018.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Sylvain Lamprier, Thibault Gisselbrecht, and Patrick Gallinari. Profile-based bandit with unknown profiles. *The Journal of Machine Learning Research*, 19(1):2060–2099, 2018.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in Neural Information Processing Systems*, 24, 2011.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

Eduardo Pavez and Antonio Ortega. Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Transactions on Information Theory*, 67(2):1201–1215, 2020.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791. ACM, 2008.

Marco Ramoni and Paola Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.

Karl H Schlag. Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits. *Journal of economic theory*, 78(1):130–156, 1998.

Fanhua Shang, Yuanyuan Liu, James Cheng, and Hong Cheng. Robust principal component analysis with missing data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1149–1158, 2014.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *Advances in Neural Information Processing Systems*, 31, 2018.

James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

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

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Jeffrey M Wooldridge. Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281–1301, 2007.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33:19075–19087, 2020.

Kai Zheng, Tianle Cai, Weiran Huang, Zhenguo Li, and Liwei Wang. Locally differentially private (contextual) bandits learning. *Advances in Neural Information Processing Systems*, 33:12300–12310, 2020.

A APPENDIX

A.1 Pseudocode of OFUL

Algorithm 2 OFUL (Abbasi-Yadkori et al., 2011)

Select a_1 uniformly at random in $\mathcal{A}_1$
Observe reward y_1
for $t = 2$ **to** T **do**
    Select $a_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_{t-1}} [1; x_{a,t}]^\top \theta$
    Observe reward y_t
end for

A.2 Proof of Proposition 2

From OFUL in Abbasi-Yadkori et al. (2011), the algorithm selects an arm a_t at each time t such that

$$(a_t, \tilde{\theta}_t) = \underset{(a,\theta) \in \mathcal{A}_t \times \mathcal{C}_{t-1}}{\operatorname{argmax}} [1; x_{a,t}]^\top \theta.$$

For the completeness, we provide a proof following the proof steps of Theorem 3 in Abbasi-Yadkori et al. (2011). We first provide a lemma about a property of the confidence set $\mathcal{C}_{t-1}$. From Section 3.1 in Smith and Winkler (2006), we observe that y_{a_t} given $x_{a,t}$'s follows a C-subGuassian distribution with mean $[1; x_{a,t}]^\top \overline{\theta}$ for some constant $C > 0$.

Lemma 1 (Theorem 2 in Abbasi-Yadkori et al. (2011)). *With probability at least $1 - 1/T$, for all $t \geq 0$, $\overline{\theta}$ lies in the set $\mathcal{C}_{t-1}$.*

Let instantaneous regret $r_t = [1; x_{a_t^\star, t}]^\top \overline{\theta} - [1; x_{a_t, t}]^\top \overline{\theta}$. Then from Lemma 1, with probability at least $1 - 1/T$, for all $t \in [T]$ we have

$$
\begin{aligned}
r_t &= [1; x_{a_t^\star, t}]^\top \overline{\theta} - [1; x_{a_t, t}]^\top \overline{\theta} \\
&\leq [1; x_{a_t, t}]^\top \tilde{\theta}_t - [1; x_{a_t, t}]^\top \overline{\theta} \\
&\leq \|[1; x_{a_t, t}]\|_{V_{t-1}^{-1}} \|\tilde{\theta}_t - \overline{\theta}\|_{V_{t-1}},
\end{aligned}
\tag{13}
$$

where the first inequality comes from the strategy of the algorithm for selecting a_t under $\overline{\theta} \in C_{t-1}$.

From Lemma 1, we can obtain that with probability at least $1 - 1/T$, for all $t \in [T]$,

$$\|\tilde{\theta}_t - \overline{\theta}\|_{V_{t-1}} \leq C_1 \sqrt{(d+1)\log(tT)} + C_2 \lambda^{1/2}, \tag{14}$$

where $\lambda = C_3 d \log(KT)$ for some constants $C_1, C_2, C_3 > 0$. We denote by E the event of (14). From (13), (14), and the fact that $r_t = O(\sqrt{d \log(KT)})$ from the scale condition for $\mathcal{A}_t$, we can obtain

$$\sum_{t=1}^{T} \mathbb{E}[r_t] = \tilde{O}\left(\sqrt{d} \sum_{t=1}^{T} \mathbb{E}\left[\|[1; x_{a_t, t}]\|_{V_{t-1}^{-1}} \mid E \right] + \sqrt{d} \right). \tag{15}$$

In what follows, we provide a lemma for a bound of $\sum_{t=1}^{T} \|[1; x_{a_t, t}]\|_{V_{t-1}^{-1}}$.

Lemma 2. *We have*

$$\sum_{t=1}^{T} \|[1; x_{a_t, t}]\|_{V_{t-1}^{-1}}^2 = O\left(d \log(T \log(KT)) \right)$$

Proof. From the scale condition of $\|x_{a,t}\|_2$ for $a \in \mathcal{A}_t$, there exists a constant $C > 0$ such that

$$\|[1; x_{a_t,t}]\|_2^2 \le Cd\log(KT)$$

Then for sufficiently large $C_3 > 0$ with $\lambda = C_3 d\log(KT)$, we have

$$\|[1; x_{a_t,t}]\|_{V_{t-1}^{-1}}^2 \le \|[1; x_{a_t,t}]\|_2^2 \|V_{t-1}^{-1}\|_2 \le \|[1; x_{a_t,t}]\|_2^2/\lambda = \|[1; x_{a_t,t}]\|_2^2/(C_3 d\log(KT)) \le 1. \tag{16}$$

From (16), by following the proof steps in Lemma 11 in Abbasi-Yadkori et al. (2011), we have

$$\sum_{t=1}^{T} \|[1; x_{a_t,t}]\|_{V_{t-1}^{-1}}^2 \le 2\sum_{t=1}^{T} \log(1 + \|[1; x_{a_t,t}]\|_{V_{t-1}^{-1}}^2)$$
$$\le 2\log\det(V_T) \le 2(d+1)\log(C(T+1)\log(KT)). \tag{17}$$

$\square$

Finally from (15) and Lemma 2, we can obtain

$$R(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$$
$$= \tilde{O}\left(\sqrt{d}\mathbb{E}\left[\sum_{t=1}^{T} \|[1; x_{a_t,t}]\|_{V_{t-1}^{-1}} \mid E\right] + \sqrt{d}\right)$$
$$\le \tilde{O}\left(\sqrt{d}\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T} \|[1; x_{a_t,t}]\|_{V_{t-1}^{-1}}^2} \mid E\right] + \sqrt{d}\right)$$
$$= \tilde{O}(d\sqrt{T}).$$

Further discussion. OFUL achieved $\tilde{O}(d\sqrt{T})$ Abbasi-Yadkori et al. (2011) in the standard linear bandits, with mean reward scale bounded by 1. Therefore, it may be of interest that OFUL, in our setting, still achieves $\tilde{O}(d\sqrt{T})$ for the case of mean reward scale $|[1; x_{a,t}]^\top\theta| = \tilde{O}(\sqrt{d})$ with $\|[1; x_{a,t}]\|_2 = \tilde{O}(\sqrt{d})$ and $\|\theta\|_2 = O(1)$. This can be obtained from that we tune a bias term λ properly for dealing with the variance of $\|[1; x_{a_t,t}]\|_{V_{t-1}^{-1}}^2$ shown in (16).

A.3 Proof of Theorem 2

For the notational simplicity, let $\Sigma = \Sigma_f + \Sigma_n$ and $\overline{z}_{a,t} = [1; \overline{x}(\nu, \Sigma, x_{a,t})]$. Also, for $s \le t$, we write $\widehat{z}_{a,s}(t)$ for $\widehat{z}_{a,s}$ at time step t. From the condition for the active set, we observe that $\|\overline{z}_{a,t}\|_2 = O(\sqrt{d\log(KT)})$. The confidence set $\mathcal{C}_t$ is designed considering this constraint to get a tight bound. However, the algorithm selects an arm based on estimated features $\widehat{z}_{a,t}(t)$ instead of $\overline{z}_{a,t}$. Hence, for the first few steps, it requires to collect feature information to estimate $\widehat{z}_{a_s,s}(t)$ for satisfying $\|\widehat{z}_{a_s,s}(t)\|_2 = O(\sqrt{d\log(KT)})$. Recall that $\theta' = (\Sigma_f + \Sigma_n)^{-1}\Sigma_f\theta^\star$ and $\overline{\theta} = [\nu^\top\theta^\star - \nu^\top\theta'; \theta']$. Let $\tau = 2\lceil d(\log(T))^2/Kp^4\rceil$. Then for any $t > \tau$ and $0 < s \le t$ we have $\|\widehat{z}_{a_s,s}(t)\|_2 = O(\sqrt{d\log(KT)})$ with a high probability, which will be shown later. The regret of Algorithm 1 can be decomposed into the followings:

$$R(T) = \sum_{t=1}^{T} \mathbb{E}\left[[1; \overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})]^\top\overline{\theta} - [1; \overline{x}(\nu_f, \Sigma, x_{a_t,t})]^\top\overline{\theta}\right]$$
$$= \sum_{t=1}^{\tau} \mathbb{E}\left[\overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})^\top\theta' - \overline{x}(\nu_f, \Sigma, x_{a_t,t})^\top\theta'\right]$$
$$+ \sum_{t=\tau+1}^{T} \mathbb{E}\left[[1, \overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})]^\top\overline{\theta} - [1, \overline{x}(\nu_f, \Sigma, x_{a_t,t})]^\top\overline{\theta}\right]. \tag{18}$$

We use the following lemma to get a regret bound before τ time steps.

Lemma 3. *For any time $t > 0$, we have*

$$\mathbb{E}\left[\overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})^\top \theta' - \overline{x}(\nu_f, \Sigma, x_{a_t,t})^\top \theta'\right] = O\left(\sqrt{\log(K)}\right). \tag{19}$$

Proof. For any given $m_{a,t}$, $\overline{x}(\nu_f, \Sigma, x_{a,t})^\top \theta'$ is a random variable with a Gaussian distribution which mean is $\nu_f^\top \theta'$ for all $a \in \mathcal{A}$. For analyzing the random variable, we provide a bound for the expected maximum value of random variables according to an i.i.d Gaussian distribution in the following.

Lemma 4. *Let X_i be an independent random variable with $\mathcal{N}(\mu, \sigma_i^2)$ for all $i \in \{1, ..., n\}$. Define $X^\star = \max\limits_{i \in \{1,...,n\}} X_i$ and $\sigma^\star = \max\limits_{i \in \{1,...,n\}} \sigma_i$. Then,*

$$\mathbb{E}[X^\star] \leq \mu + \sigma^\star \sqrt{2\log(n)}.$$

Proof. For any $\lambda \geq 0$,

$$\exp(\lambda \mathbb{E}[X^\star]) \leq \mathbb{E}[\exp(\lambda X^\star)]$$
$$= \mathbb{E}\left[\max_{i \in [n]} \exp(\lambda X_i)\right]$$
$$\leq \sum_{i=1}^n \mathbb{E}[\exp(\lambda X_i)]$$
$$\leq n \exp(\lambda \mu + \lambda^2 (\sigma^\star)^2 / 2).$$

Set $\lambda = \frac{\sqrt{2\log(n)}}{\sigma^\star}$, then we get

$$\mathbb{E}[X^\star] \leq \mu + \sigma^\star \sqrt{2\log(n)}.$$

$\square$

Denote by $Var(X)$ the variance of a random variable X and $V^\star$ the maximum value among $Var(\overline{x}(\nu_f, \Sigma, x_{a,t})^\top \theta')$ for all $a \in \mathcal{A}$. Then using Lemma 4 and $\mathcal{A}_t \subseteq \mathcal{A}$, we have

$$\mathbb{E}\left[\overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})^\top \theta' - \overline{x}(\nu_f, \Sigma, x_{a_t,t})^\top \theta'\right]$$
$$\leq \mathbb{E}\left[\max_{a \in \mathcal{A}} \left(\overline{x}(\nu_f, \Sigma, x_{a,t})^\top \theta'\right) + \max_{a \in \mathcal{A}} \left(-\overline{x}(\nu_f, \Sigma, x_{a,t})^\top \theta'\right)\right]$$
$$\leq \mathbb{E}[2\sqrt{2V^\star \log(K)}], \tag{20}$$

where $V^\star$ is bounded as follows. Define $\mathcal{E} = \{1, ..., d\}$. For any $a \in \mathcal{A}$ and any given $m_{a,t}$, we have

$$Var(\overline{x}(\nu_f, \Sigma, x_{a,t})^\top \theta') = Var((\nu_f + \Sigma_{\mathcal{ES}}\Sigma_{\mathcal{SS}}^{-1}(x_{a,t} - \nu_f)_{\mathcal{S}})^\top (\Sigma^{-1}\Sigma_f \theta^\star))$$
$$= Var((\Sigma_{\mathcal{ES}}\Sigma_{\mathcal{SS}}^{-1}(x_{a,t})_{\mathcal{S}})^\top \Sigma^{-1}\Sigma_f \theta^\star)$$
$$= \mathbb{E}\left[((x_{a,t} - \nu_f)_{\mathcal{S}}^\top \Sigma_{\mathcal{SS}}^{-1}\Sigma_{\mathcal{ES}}^\top \Sigma^{-1}\Sigma_f \theta^\star)^2\right]$$
$$= O\left(\theta^{\star\top}\Sigma_f \Sigma^{-1}\Sigma_{\mathcal{ES}}\Sigma_{\mathcal{SS}}^{-1}\Sigma_{\mathcal{ES}}^\top \Sigma^{-1}\Sigma_f \theta^\star\right)$$
$$= O\left(\|\theta^\star\|_2^2 \|\Sigma_f\|_2^2 \|\Sigma^{-1}\|_2^2 \|\Sigma_{\mathcal{ES}}\|_2^2 \|\Sigma_{\mathcal{SS}}^{-1}\|_2\right)$$
$$= O\left(\|\theta^\star\|_2^2 \|\Sigma\|_2^4 \|\Sigma^{-1}\|_2^2 \|\Sigma_{\mathcal{SS}}^{-1}\|_2\right)$$
$$= O\left(1\right).$$

$\square$

Therefore from Lemma 3, the regret over the first τ time steps in (18) is replaced by

$$\sum_{t=1}^{\tau} \mathbb{E}\left[\overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})^\top \theta' - \overline{x}(\nu_f, \Sigma, x_{a_t,t})^\top \theta'\right] = O\left(\tau\sqrt{\log(K)}\right). \tag{21}$$

Now we provide a regret bound after τ time steps. Recall that instantaneous regret $r_t = \overline{z}_{a_t^\star,t}^\top \overline{\theta} - \overline{z}_{a_t,t}^\top \overline{\theta}$. In the algorithm, at each time step, the covariance matrix Σ is estimated by the scaled empirical covariance matrix $\widehat{\Sigma}$ and the mean of feature vector ν_f is estimated by $\widehat{\nu}$. For ease of presentation, we write $\widehat{p}_t$, $\widehat{\nu}_t$, and $\widehat{\Sigma}_t$ for $\widehat{p}$, $\widehat{\nu}$, and $\widehat{\Sigma}$ at time step t, respectively. For analyzing the regret, we can decompose the regret according to the estimation errors with high probability as follows:

$$r_t \leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \|\tilde{\theta}_t - \overline{\theta}\|_{V_{t-1}} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} + \|\widehat{z}_{a_t,t}(t) - \overline{z}_{a_t,t}\|_2 \|\overline{\theta}\|_2,$$

which will be shown later. The instantaneous regret bound consists of the estimation errors of $\widehat{z}_{a_t^\star,t}(t)$, $\widehat{z}_{a_t,t}(t)$, and $\tilde{\theta}_t$. Therefore we first focus on providing bounds for $\|\overline{z}_{a,t} - \widehat{z}_{a,t}(t)\|_2$ and $\|\overline{\theta} - \tilde{\theta}_t\|_2$. From matrix concentration inequalities, we have the following lemma.

Lemma 5. *For all $t > \tau/2$, with probability at least $1 - 3/T$, we have*

$$\|\nu_f - \widehat{\nu}_t\|_2 = O\left(\frac{1}{p}\sqrt{\frac{d\log(T)}{tK}}\right), \left\|\Sigma - \widehat{\Sigma}_t\right\|_2 = O\left(\frac{1}{p^2}\sqrt{\frac{d\log(T)}{tK}}\right), \text{ and } |\widehat{p}_t - p| = O\left(\sqrt{\frac{\log(T)}{dtK}}\right)$$

Proof. Recall $x'_{a,t} = z_{a,t} + \varepsilon_{a,t}$. We define $v_{a,t} = x'_{a,t} \circ m_{a,t} - \nu_f p$ for all $a \in \mathcal{A}$. Let z_j be the j-th entry in an arbitrary vector z. Then, for any $u \in \mathbb{R}^d$ such that $\|u\|_2 = 1$ and for any $\lambda \in \mathbb{R}$, we get

$$\mathbb{E}[\exp(\lambda v_{a,t}^\top u)]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda((x'_{a,t} - \nu_f) \circ m_{a,t})^\top u\right) | m_{a,t}\right] \exp(\lambda(\nu_f \circ m_{a,t} - \nu_f p)^\top u)\right]$$
$$\leq \exp(\|\Sigma\|_2 \lambda^2/2) \mathbb{E}\left[\exp(\lambda(\nu_f \circ m_{a,t} - \nu_f p)^\top u)\right]$$
$$\leq \exp(\|\Sigma\|_2 \lambda^2/2) \prod_{j=1}^{d} \mathbb{E}\left[\exp(\lambda(\nu_f \circ m_{a,t} - \nu_f p)_j u_j)\right]$$
$$\leq \exp(\|\Sigma\|_2 \lambda^2/2) \exp(\sum_{j=1}^{d} |(\nu_f)_j u_j|^2 \lambda^2/2)$$
$$\leq \exp(\|\Sigma\|_2 \lambda^2/2) \exp(\sum_{j=1}^{d} |(\nu_f)_j u_j|^2 \lambda^2/2)$$
$$\leq \exp((\|\Sigma\|_2 + 1)\lambda^2/2),$$

where the first inequality is from the normal moment generating function and the third inequality is from Hoeffding's lemma. From the definition of sub-gaussian vector, we can find that $v_{a,t}$ for $a \in \mathcal{A}$ and $t > 0$ are according to independent sub-gaussian with variance proxy $\|\Sigma\|_2 + 1$. According to Theorem 2.1 of Hsu et al. (2012), for all $k > 0$ we get

$$\mathbb{P}\left(\left\|\sum_{s=1}^{t}\sum_{a\in\mathcal{A}}(x'_{a,s} \circ m_{a,s} - \nu_f p)\right\|_2 \geq \sqrt{tK(\|\Sigma\|_2 + 1)\left(d + 2\sqrt{dk} + 2k\right)}\right) \leq \exp(-k). \tag{22}$$

Then we have with probability at least $1 - 1/T^2$

$$\left\|\frac{1}{tK}\sum_{s=1}^{t}\sum_{a\in\mathcal{A}}(x_{a,s} - p\nu_f)\right\|_2 = O\left(\sqrt{\frac{d + \log(T) + \sqrt{d\log(T)}}{tK}}\right)$$

By Hoeffding's inequality, we also have

$$|\widehat{p}_t - p| = O\left(\sqrt{\frac{\log(T)}{dtK}}\right) \quad \text{w.p. at least } 1 - 1/T^2. \tag{23}$$

Recall that $\tau/2 = \lceil d(\log(T))^2/(Kp^4) \rceil$. Then for $t > \tau/2$, we have $\widehat{p}_t = \Theta(p)$ at least probability $1 - 1/T^2$. Using the above inequalities, we provide bounds for $\|\nu_f - \widehat{\nu}_t\|_2$ and $\|\Sigma - \widehat{\Sigma}_t\|_2$. With probability at least $1 - 1/T^2$, for $t > \tau/2$ we have

$$
\begin{aligned}
\|\widehat{\nu}_t - \nu_f\|_2 &= \left\| \widehat{\nu}_t - \frac{1}{tKp} \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} + \frac{1}{tKp} \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} - \nu_f \right\|_2 \\
&\leq \left| \frac{p - \widehat{p}_t}{p\widehat{p}_t} \right| \left(\left\| \frac{1}{tK} \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} (x_{a,s} - p\nu_f) \right\|_2 + \|p\nu_f\|_2 \right) + \left\| \frac{1}{tKp} \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} - \nu_f \right\|_2 \\
&= O \left(\frac{1}{p} \sqrt{\frac{d \log(T)}{tK}} \right).
\end{aligned}
$$

Now we provide a bound for $\|\Sigma - \widehat{\Sigma}_t\|_2$. We define $w = \nu_f p$. By using $x_{a,t} = v_{a,t} + w$, we get

$$
\begin{aligned}
\left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} x_{a,s}^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} x_{a,s}^\top \right] \right\|_2 &\leq \left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} v_{a,s}^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} v_{a,s}^\top \right] \right\|_2 \\
&\quad + 2 \left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top \right] \right\|_2.
\end{aligned} \tag{24}
$$

For the first term in the RHS of (24), from Proposition 2.1 of Vershynin (2012) with sub-Gaussian v_a, with probability at least $1 - 1/T^2$ we get

$$
\left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} v_{a,s}^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} v_{a,s}^\top \right] \right\|_2 = O \left(\sqrt{tKd \log(T)} \right). \tag{25}
$$

For the second term in the RHS of (24),

$$
\begin{aligned}
\left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top \right] \right\|_2 &= \left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top \right\|_2 \\
&\leq \left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} (x'_{a,s} \circ m_{a,s} - \nu_f p) \right\|_2 \|\nu_f p\|_2.
\end{aligned} \tag{26}
$$

From the results of (26) and (22), with probability at least $1 - 1/T^2$ we get

$$
\begin{aligned}
\left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} v_{a,s} w^\top \right] \right\|_2 &= O \left(\left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} (x'_{a,s} \circ m_{a,s} - \nu_f p) \right\|_2 p \right) \\
&= O \left(p \sqrt{tKd \ln(T)} \right).
\end{aligned} \tag{27}
$$

By putting the results of (24), (25), and (27), with probability at least $1 - 1/T^2$ we get

$$
\left\| \sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} x_{a,s}^\top - \mathbb{E} \left[\sum_{s=1}^{t} \sum_{a \in \mathcal{A}} x_{a,s} x_{a,s}^\top \right] \right\|_2 = O \left(\sqrt{tKd \log(T)} \right). \tag{28}
$$

We write Z_t for Z at time step t in the algorithm. Then, (28) directly implies that

$$
\|Z_t - \mathbb{E}[Z_t]\|_2 = O \left(\sqrt{tKd \log(T)} \right). \tag{29}
$$

Let $\mathbf{1}_{d\times d}$ be a matrix $\in \mathbb{R}^{d\times d}$ with all 1 entries and $P = (p - p^2)I_d + p^2\mathbf{1}_{d\times d}$. Then from Proposition 1 in Pavez and Ortega (2020), we can show that

$$\|\mathbb{E}[Z_t]\|_2 = tK\|\Sigma \circ P + (\nu\nu^\top) \circ P\|_2 = O(tKp) \tag{30}$$

Lastly, with probability at least $1 - 1/T^2$, we get

$$
\begin{aligned}
\|\widehat{\Sigma}_t - \Sigma\|_2 &\leq \left\|\widehat{\Sigma}_t - \left(\frac{1}{tK}Z_t \circ \left(\frac{p-1}{p^2}I_d + \frac{1}{p^2}\mathbf{1}_{d\times d}\right) - \nu_f\nu_f^\top\right)\right.\\
&\quad \left. + \left(\frac{1}{tK}Z_t \circ \left(\frac{p-1}{p^2}I_d + \frac{1}{p^2}\mathbf{1}_{d\times d}\right) - \nu_f\nu_f^\top\right) - \Sigma\right\|_2\\
&\leq \left\|\widehat{\Sigma}_t - \left(\frac{1}{tK}Z_t \circ \left(\frac{p-1}{p^2}I_d + \frac{1}{p^2}\mathbf{1}_{d\times d}\right) - \nu_f\nu_f^\top\right)\right\|_2\\
&\quad + \left\|\left(\frac{1}{tK}Z_t \circ \left(\frac{p-1}{p^2}I_d + \frac{1}{p^2}\mathbf{1}_{d\times d}\right) - \nu_f\nu_f^\top\right) - \Sigma\right\|_2\\
&= O\left(\frac{1}{tK}\frac{|p - \widehat{p}_t|}{p^3}(\|Z_t - \mathbb{E}[Z_t]\|_2 + \|\mathbb{E}[Z_t]\|_2) + \|\widehat{\nu}_t - \nu_f\|_2\right.\\
&\quad \left. + \frac{1}{tK}\left\|(Z_t - \mathbb{E}[Z_t]) \circ \left(\frac{p-1}{p^2}I_d + \frac{1}{p^2}\mathbf{1}_{d\times d}\right)\right\|_2\right)\\
&= O\left(\frac{1}{p^2}\sqrt{\frac{d\log(T)}{tK}}\right).
\end{aligned}
$$

Therefore, using the union bound for all time $t > \tau/2$, we can conclude the proof. $\qquad\square$

From Lemma 5, we define an event

$$
\begin{aligned}
E_1 = \Bigg\{ &\|\nu_f - \widehat{\nu}_t\|_2 = O\left(\frac{1}{p}\sqrt{\frac{d\log(T)}{tK}}\right), \left\|\Sigma - \widehat{\Sigma}_t\right\|_2 = O\left(\frac{1}{p^2}\sqrt{\frac{d\log(T)}{tK}}\right),\\
&\text{and } |\widehat{p}_t - p| = O\left(\sqrt{\frac{\log(T)}{dtK}}\right) \forall t \in [\tau/2 + 1, T] \Bigg\},
\end{aligned}
\tag{31}
$$

which holds true with at least probability $1 - 3/T$. Using Weyl's inequality, we can show that $\sigma_d(\widehat{\Sigma}_t) \geq \sigma_d(\Sigma) - \|\widehat{\Sigma}_t - \Sigma\|_2$. Under E_1, for all $t > \tau/2$, we have $\|\Sigma - \widehat{\Sigma}_t\|_2 = o(1)$. Then, we have

$$\|\widehat{\Sigma}_t\|_2 \leq \|\widehat{\Sigma}_t - \Sigma\|_2 + \|\Sigma\|_2 = O(\|\Sigma\|_2) \quad \text{and}$$

$$\|\widehat{\Sigma}_t^{-1}\|_2 = \frac{1}{\sigma_d(\widehat{\Sigma}_t)} \leq \frac{1}{\sigma_d(\Sigma) - \|\widehat{\Sigma}_t - \Sigma\|_2} = O\left(\frac{1}{\sigma_d(\Sigma)}\right) = O\left(\|\Sigma^{-1}\|_2\right). \tag{32}$$

By following the same steps, for any non empty set $\mathcal{S}$, we can also show that

$$\|(\widehat{\Sigma}_t)_{\mathcal{S}\mathcal{S}}^{-1}\|_2 = O(\|\Sigma_{\mathcal{S}\mathcal{S}}^{-1}\|_2). \tag{33}$$

Then, under E_1, from the definition of $\widehat{z}_{a,t}(t)$ and (32), for any $t > \tau$ and $0 < s \leq t$ we have

$$\|\widehat{z}_{a_s,s}(t)\|_2 = O(\sqrt{d\log(KT)}).$$

This is because the algorithm updates the estimated features of previously chosen arms once within $\tau/2 < t \leq \tau$. In what follows, we provide a bound for feature estimators using (32). For simplicity, we define $\bar{z}_{a,t} := [1; \bar{x}(\nu, \Sigma, x_{a,t})]$. In the algorithm, recall that for $s \leq t$, $\widehat{z}_{a,s}(t) = [1; \bar{x}(\widehat{\nu}, \widehat{\Sigma}, x_{a,s})]$ at time t.

Lemma 6. *Under E_1, for all $t > \tau/2$ and $a \in \mathcal{A}_t$, we have*

$$\|\bar{z}_{a,t} - \widehat{z}_{a,t}(t)\|_2 = O\left(d\sqrt{\frac{1}{p^3}\frac{\log(KT)\log(T)}{tK}}\right).$$

Proof. First, at time step t, we have

$$\Sigma_{\mathcal{US}}\Sigma_{\mathcal{SS}}^{-1}$$
$$= \Sigma_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1}(\widehat{\Sigma}_t)_{\mathcal{SS}}\Sigma_{\mathcal{SS}}^{-1}$$
$$= \Sigma_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1} + \Sigma_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1}((\widehat{\Sigma}_t)_{\mathcal{SS}} - \Sigma_{\mathcal{SS}})\Sigma_{\mathcal{SS}}^{-1}$$
$$= (\widehat{\Sigma}_t)_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1} + (\Sigma_{\mathcal{US}} - (\widehat{\Sigma}_t)_{\mathcal{US}})(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1} + \Sigma_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1}((\widehat{\Sigma}_t)_{\mathcal{SS}} - \Sigma_{\mathcal{SS}})\Sigma_{\mathcal{SS}}^{-1}. \tag{34}$$

Thus with (34), (32), and (33), under E_1 for $t > \tau/2$

$$\left\| \Sigma_{\mathcal{US}}\Sigma_{\mathcal{SS}}^{-1} - (\widehat{\Sigma}_t)_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1} \right\|_2$$
$$\leq \left\| (\Sigma_{\mathcal{US}} - (\widehat{\Sigma}_t)_{\mathcal{US}})(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1} \right\|_2 + \left\| \Sigma_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1}((\widehat{\Sigma}_t)_{\mathcal{SS}} - \Sigma_{\mathcal{SS}})\Sigma_{\mathcal{SS}}^{-1} \right\|_2$$
$$= O\left(\|\Sigma - \widehat{\Sigma}_t\|_2 \|\Sigma_{\mathcal{SS}}^{-1}\|_2 + \|\Sigma\|_2 \|\Sigma_{\mathcal{SS}}^{-1}\|_2 \|\widehat{\Sigma}_t - \Sigma\|_2 \|\Sigma_{\mathcal{SS}}^{-1}\|_2 \right)$$
$$= O\left(\|\widehat{\Sigma}_t - \Sigma\|_2 \right).$$

Therefore, for any $t > \tau/2$ and $a \in \mathcal{A}_t$, it follows that under E_1,

$$\|\widehat{z}_{a,t}(t) - \overline{z}_{a,t}\|_2^2 \leq 2 \left\| \Sigma_{\mathcal{US}}\Sigma_{\mathcal{SS}}^{-1} - (\widehat{\Sigma}_t)_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1} \right\|_2^2 \|(x_{a,t} - \nu_f)_{\mathcal{S}}\|_2^2$$
$$+ 2\left(\|\nu_f - \widehat{\nu}_t\|_2 + \|(\widehat{\Sigma}_t)_{\mathcal{US}}(\widehat{\Sigma}_t)_{\mathcal{SS}}^{-1}(\nu_f - \widehat{\nu}_t)_{\mathcal{S}}\|_2 \right)^2$$
$$= O\left(\|\Sigma - \widehat{\Sigma}_t\|_2^2 \mathbb{E}[\|m_{a,t}\|_2^2] \log(KT) + \|\nu_f - \widehat{\nu}_t\|_2^2 \right)$$
$$= O\left(pd\|\Sigma - \widehat{\Sigma}_t\|_2^2 \log(KT) + \|\nu_f - \widehat{\nu}_t\|_2^2 \right)$$
$$= O\left(d^2(1/p^3) \log(T) \log(KT)/tK \right).$$

$\square$

Now, we provide a lemma for showing a good property of the confidence set $\mathcal{C}_t$.

Lemma 7. *Under E_1, for all $t > \tau$, with probability at least $1 - 1/T$, we have $\overline{\theta} \in \mathcal{C}_{t-1}$.*

Proof. Since, under E_1, $\|\widehat{\Sigma}\|_2 = O(\|\Sigma\|_2)$ and $\|\widehat{\Sigma}^{-1}\|_2 = O(\|\Sigma^{-1}\|_2)$, we have, for all $t > \tau$ and $1 \leq s \leq t$ considering rarely updating estimators in the algorithm,

$$\|\widehat{z}_{a_s,s}(t)\|_2 = O(\sqrt{d \log(KT)}).$$

This is because the algorithm updates the estimated features of previously chosen arms once within $\tau/2 < t \leq \tau$. Then for $t > \tau$ we can observe that $\|\widehat{z}_{a_s,s}(t)\|_2^2 \leq C_1 d \log(KT)$ and $\|\overline{\theta}\|_2 \leq C_2$ for some constants C_1 and $C_2 > 0$. We also observe that noise of reward η_t is independent to E_1. Using the facts, Lemma 6, and Theorem 2 in Abbasi-Yadkori et al. (2011), we can easily prove that for all $t > \tau$, with probability at least $1 - 1/T$, $\overline{\theta}$ lies in $\mathcal{C}_{t-1}$. In what follows, we provide details. From Section 3.1 in Smith and Winkler (2006), Proposition 3.13 in Eaton (1983), and 1-subGaussian η_t, we observe that y_{a_t} given $x_{a,t}$ follows a C_3-subGaussian distribution with mean $\overline{x}^\top \overline{\theta}$ for some constant $C_3 > 0$. We denote by $\tilde{\eta}_t$ the noise term for y_{a_t} given $x_{a,t}$ at time t. Then, we have

$$\widehat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} \widehat{z}_{a_s,s}(t+1)\left(\overline{z}_{a_s,s}^\top \overline{\theta} + \tilde{\eta}_s \right)$$
$$= V_t^{-1} \sum_{s=1}^{t} \widehat{z}_{a_s,s}(t+1)\left(\widehat{z}_{a_s,s}(t+1)^\top \overline{\theta} + \tilde{\eta}_s \right) + V_t^{-1} \sum_{s=1}^{t} \widehat{z}_{a_s,s}(t+1)(\overline{z}_{a_s,s} - \widehat{z}_{a_s,s}(t+1))^\top \overline{\theta}$$
$$= V_t^{-1} \sum_{s=1}^{t} \widehat{z}_{a_s,s}(t+1)\tilde{\eta}_s + \overline{\theta} - \lambda V_t^{-1}\overline{\theta} + V_t^{-1} \sum_{s=1}^{t} \widehat{z}_{a_s,s}(t+1)(\overline{z}_{a_s,s} - \widehat{z}_{a_s,s}(t+1))^\top \overline{\theta}.$$

Then $\forall x \in \mathbb{R}^{d+1}$, we have

$$|x^\top \widehat{\theta}_t - x^\top \overline{\theta}| \leq \|x\|_{V_t^{-1}} \left(\left\| \sum_{s=1}^t \widehat{z}_{a_s,s}(t+1)\widetilde{\eta}_s \right\|_{V_t^{-1}} + \lambda^{1/2}\|\overline{\theta}\|_2 + \left\| \sum_{s=1}^t \widehat{z}_{a_s,s}(t+1)(\overline{z}_{a_s,s} - \widehat{z}_{a_s,s}(t+1))^\top \overline{\theta} \right\|_{V_t^{-1}} \right)$$

$$\leq \|x\|_{V_t^{-1}} \left(\left\| \sum_{s=1}^t \widehat{z}_{a_s,s}(t+1)\widetilde{\eta}_s \right\|_{V_t^{-1}} + \lambda^{1/2}\|\overline{\theta}\|_2 + \sum_{s=1}^t \|\widehat{z}_{a_s,s}(t+1)\|_{V_t^{-1}} \|\overline{z}_{a_s,s} - \widehat{z}_{a_s,s}(t+1)\|_2 \|\overline{\theta}\|_2 \right).$$

By setting $x = V_t(\widehat{\theta}_t - \overline{\theta})$, we have

$$\|\widehat{\theta}_t - \overline{\theta}\|_{V_t} \leq \left\| \sum_{s=1}^t \widehat{z}_{a_s,s}(t+1)\widetilde{\eta}_s \right\|_{V_t^{-1}} + \lambda^{1/2}\|\overline{\theta}\|_2 + \sum_{s=1}^t \|\widehat{z}_{a_s,s}(t+1)\|_{V_t^{-1}} \|\overline{z}_{a_s,s} - \widehat{z}_{a_s,s}(t+1)\|_2 \|\overline{\theta}\|_2.$$

From Lemma 9 in Abbasi-Yadkori et al. (2011), with probability at least $1 - 1/T$ we have

$$\left\| \sum_{s=1}^t \widehat{z}_{a_s,s}(t+1)\widetilde{\eta}_s \right\|_{V_t^{-1}} \leq C_3 \sqrt{2(d+1)\log(T(t+1))}.$$

We can observe that $\widehat{p}_t = \Theta(p)$ when $t > \tau$ under E_1. Putting them together, for some constant $C_5, C_6 > 0$, with probability at least $1 - 1/T$, we have

$$\|\widehat{\theta}_t - \overline{\theta}\|_{V_t} \leq C_5 \sqrt{(d+1)\log(T(t+1))} + C_2 \lambda^{1/2} + \sum_{s=1}^t \|\widehat{z}_{a_s,s}(t+1)\|_{V_t^{-1}} \frac{d}{p^{3/2}} \sqrt{\frac{\log(T)\log(KT)}{(t+1)K}}$$

$$\leq C_5 \sqrt{(d+1)\log(T(t+1))} + C_2 \lambda^{1/2} + C_6 \sum_{s=1}^t \|\widehat{z}_{a_s,s}(t+1)\|_{V_t^{-1}} \frac{d}{\widehat{p}^{3/2}} \sqrt{\frac{\log(T)\log(KT)}{(t+1)K}}.$$

$\square$

From the above lemma, we define an event

$$E_2 = \left\{ \overline{\theta} \in \mathcal{C}_{t-1}, \forall t \in [\tau+1, T] \right\}.$$

Then we provide a lemma for decomposing the instantaneous regret according to the estimation errors of $\widehat{z}_{a_t^\star,t}(t)$, $\widehat{z}_{a_t,t}(t)$, and $\widetilde{\theta}_t$.

Lemma 8. *Under E_2, we have*

$$r_t \leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \|\widetilde{\theta}_t - \overline{\theta}\|_{V_{t-1}} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} + \|\widehat{z}_{a_t,t}(t) - \overline{z}_{a_t,t}\|_2 \|\overline{\theta}\|_2$$

Proof. Under E_2, from the fact that $\overline{\theta} \in \mathcal{C}_{t-1}$ and considering that a_t is the chosen arm at time t, we have $\widehat{z}_{a_t^\star,t}(t)^\top \overline{\theta} \leq \widehat{z}_{a_t,t}(t)^\top \widetilde{\theta}_t$. Then we have

$$\begin{aligned}
r_t &= \overline{z}_{a_t^\star,t}^\top \overline{\theta} - \overline{z}_{a_t,t}^\top \overline{\theta} \\
&= \overline{z}_{a_t^\star,t}^\top \overline{\theta} - \widehat{z}_{a_t^\star,t}(t)^\top \overline{\theta} + \widehat{z}_{a_t^\star,t}(t)^\top \overline{\theta} - \overline{z}_{a_t,t}^\top \overline{\theta} \\
&\leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \widehat{z}_{a_t^\star,t}(t)^\top \overline{\theta} - \overline{z}_{a_t,t}^\top \overline{\theta} \\
&\leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \widehat{z}_{a_t,t}(t)^\top \widetilde{\theta}_t - \overline{z}_{a_t,t}^\top \overline{\theta} \\
&\leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \widehat{z}_{a_t,t}(t)^\top \widetilde{\theta}_t - \widehat{z}_{a_t,t}(t)^\top \overline{\theta} + \widehat{z}_{a_t,t}(t)\overline{\theta} - \overline{z}_{a_t,t}\overline{\theta} \\
&\leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2 \|\overline{\theta}\|_2 + \|\widetilde{\theta}_t - \overline{\theta}\|_{V_{t-1}} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} + \|\widehat{z}_{a_t,t}(t) - \overline{z}_{a_t,t}\|_2 \|\overline{\theta}\|_2.
\end{aligned}$$

$\square$

Then, we provide a bound for $\sum_{t=\tau+1}^{T} \mathbb{E}[r_t]$ in the following.

Lemma 9. *We have*

$$\sum_{t=\tau+1}^{T} \mathbb{E}[r_t] = \tilde{O}\left(\sqrt{d}\mathbb{E}\left[\sum_{t=\tau+1}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} \mid E_1 \cap E_2\right] + \frac{d}{p^{3/2}}\sqrt{\frac{T}{K}} + \sqrt{d}\right).$$

(35)

Proof. Under E_1 and E_2, we have $\tilde{\theta}_t \in \mathcal{C}_{t-1}$. Then we get

$$\|\tilde{\theta}_t - \overline{\theta}\|_{V_{t-1}} \leq C_5\sqrt{(d+1)\log(Tt))} + C_2\lambda^{1/2} + C_6\sum_{s=1}^{t-1}\|\widehat{z}_{a_s,s}(t)\|_{V_{t-1}^{-1}}\frac{d}{\widehat{p}^{3/2}}\sqrt{\frac{\log(T)\log(KT)}{tK}}$$

$$\leq C_5\sqrt{2(d+1)\log(Tt)} + C_2\lambda^{1/2} + C_6\sqrt{t\sum_{s=1}^{t-1}\|\widehat{z}_{a_s,s}(t)\|_{V_{s-1}^{-1}}^2}\frac{d}{\widehat{p}^{3/2}}\sqrt{\frac{\log(T)\log(KT)}{tK}}$$

$$= O\left(\sqrt{d\log(Tt)} + \lambda^{1/2} + \frac{d^{3/2}}{p^{3/2}}\log(T)\sqrt{\frac{\log(KT)}{K}}\right),$$

where the last equality comes from Lemma 11 in Abbasi-Yadkori et al. (2011).

Then with Lemma 8, we have

$$r_t \leq \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2\|\overline{\theta}\|_2 + \|\tilde{\theta}_t - \overline{\theta}\|_{V_{t-1}}\|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} + \|\widehat{z}_{a_t,t}(t) - \overline{z}_{a_t,t}\|_2\|\overline{\theta}\|_2$$

$$= \tilde{O}\left(\left(\sqrt{d} + (d/p)^{3/2}\sqrt{1/K}\right)\|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} + \|\overline{z}_{a_t^\star,t} - \widehat{z}_{a_t^\star,t}(t)\|_2\|\overline{\theta}\|_2 + \|\widehat{z}_{a_t,t}(t) - \overline{z}_{a_t,t}\|_2\|\overline{\theta}\|_2\right),$$

(36)

where the equality comes from $\tilde{\theta}_t \in \mathcal{C}_{t-1}$. Then from Lemma 6, we have

$$r_t = \tilde{O}\left(\left(\sqrt{d} + (d/p)^{3/2}\sqrt{1/K}\right)\|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} + (d/p^{3/2})\sqrt{1/tK}\right).$$

(37)

We note that under the complement event $E_1^c \cup E_2^c$ which holds at most probability $2/T$, for all $t > \tau$, we have $r_t \leq |\overline{z}_{a_t^\star,t}^\top\theta'| + |\overline{z}_{a_t,t}^\top\theta'| = O(\sqrt{d\log(KT)})$. Therefore with (36), we have

$$\sum_{t=\tau+1}^{T} \mathbb{E}[r_t] = \tilde{O}\left(\left(\sqrt{d} + (d/p)^{3/2}\sqrt{1/K}\right)\mathbb{E}\left[\sum_{t=\tau+1}^{T}\|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} \mid E_1 \cap E_2\right] + (d/p^{3/2})\sqrt{T/K} + \sqrt{d}\right)$$

□

Now we provide a lemma to bound the first term in (35).

Lemma 10. *Under E_1, we have*

$$\sum_{t=\tau+1}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2 = O\left(d\log(T\log(KT))\log(T)\right)$$

Proof. Under E_1, for some sufficiently large constant $C > 0$ with $\lambda = Cd\log(KT)$, for any $t \geq \tau + 1$ and $s \leq t$ we have

$$\|\widehat{z}_{a_s,s}(t)\|_2^2 \leq Cd\log(KT) \text{ and}$$

$$\|\widehat{z}_{a_s,s}(t)\|_{V_{t-1}^{-1}}^2 \leq \|\widehat{z}_{a_s,s}(t)\|_2^2\|V_{t-1}^{-1}\|_2 \leq \|\widehat{z}_{a_s,s}(t)\|_2^2/Cd\log(KT) \leq 1.$$

Let 2^l be the smallest time step after $\tau + 1$ for an integer l and 2^n be the largest time step before T for an integer n. Then we have $n = O(\log T)$. We note that estimated features of previously selected arms in V_{t-1} are not updated during the time steps between $t = 2^i + 1$ and $2^{i+1} - 1$ for $i \in [n]$. Then by following the proof steps in Lemma 11 in Abbasi-Yadkori et al. (2011) and considering rarely updating procedure in the algorithm, we can obtain

$$
\begin{aligned}
\sum_{t=\tau+1}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2 &= \sum_{t=\tau+1}^{2^l-1} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2 + \sum_{i=l}^{n-1} \sum_{t=2^i}^{2^{i+1}-1} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2 + \sum_{t=2^n}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2 \\
&\leq 2 \sum_{t=\tau+1}^{2^l-1} \log\left(1 + \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2\right) \\
&\quad + 2 \sum_{i=l}^{n-1} \sum_{t=2^i}^{2^{i+1}-1} \log\left(1 + \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2\right) + 2 \sum_{t=2^n}^{T} \log\left(1 + \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2\right) \\
&\leq 2(\log(\det(V_{2^l-1})) + \sum_{i=l}^{n-1} 2(\log(\det(V_{2^{i+1}-1})) + 2(\log(\det(V_T)) \\
&\leq 2(n+2)(d+1)\log(C(T+1)\log(KT)) \\
&= O\left(d \log(T \log(KT)) \log(T)\right),
\end{aligned}
\tag{38}
$$

where the first inequality is obtained from $x \leq 2\log(1+x)$ when $x \in [0,1]$, the second and last inequality is obtained from Lemma 11 in Abbasi-Yadkori et al. (2011), and the last equality is obtained from $n = O(\log T)$. $\qquad\square$

Finally using (18), (21), Lemmas 9 and 10, and Cauchy-Schwarz inequality, we get

$$
\begin{aligned}
R(T) &= \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \\
&= \tilde{O}\left(\tau + \mathbb{E}\left[\left(\sqrt{d} + (d/p)^{3/2}\sqrt{1/K}\right) \sum_{t=\tau+1}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}} \mid E_1 \cap E_2\right] + \frac{d}{p^{3/2}}\sqrt{T/K} + \sqrt{d}\right) \\
&= \tilde{O}\left(d/Kp^4 + \mathbb{E}\left[\left(\sqrt{d} + (d/p)^{3/2}\sqrt{1/K}\right)\sqrt{T \sum_{t=\tau+1}^{T} \|\widehat{z}_{a_t,t}(t)\|_{V_{t-1}^{-1}}^2} \mid E_1 \cap E_2\right] + \frac{d}{p^{3/2}}\sqrt{T/K} + \sqrt{d}\right) \\
&= \tilde{O}\left(\frac{d}{Kp^4} + d\sqrt{T} + \frac{d^2}{p^{3/2}}\sqrt{\frac{T}{K}}\right).
\end{aligned}
$$

A.4 Tightness of the regret bound of Theorem 2 with respect to p

Dependency on p in regret analysis comes from variance of estimator $\widehat{\Sigma}_t$ which includes $1/\widehat{p}_t^2$ term for reducing bias. In more detail, from (8) we can observe that regret depends on the estimation error for Bayesian feature estimator $\widehat{z}_a$, and this error depends on the variance of $\widehat{\Sigma}_t$ and $\|x_{a,t}\|_2$ (See proof of Lemma 6). Then, from Lemma 5, we get $\|\widehat{\Sigma}_t - \Sigma\|_2 = O((1/p^2)\sqrt{d\log(T)/tK})$, and from Gaussian distribution we get $\|x_{a,t}\|_2 = \tilde{O}(\sqrt{pd})$. Therefore, $1/p^{3/2}$ in the regret bound seems to be inevitable considering $1/p^2$ in the variance of $\widehat{\Sigma}_t$ and $\sqrt{p}$ from $\|x_{a,t}\|_2$.

Next, we explain why $1/p^4$ appears in the regret bound. We can observe that the instantaneous regret depends on $\|\widehat{z}_{a_t}\|_2$ from (11) and Lemma 10. For bounding the variance $\|\widehat{z}_{a_t}\|_2$, we need to bound $\|\widehat{\Sigma}_t^{-1}\|_2$ from the Bayesian feature definition. Using Weyl's inequality, we have $\|\widehat{\Sigma}_t^{-1}\|_2 \leq 1/(\sigma_d(\Sigma) - \|\widehat{\Sigma}_t - \Sigma\|_2)$ in (32). Therefore, for getting a tight constant bound for $\|\widehat{\Sigma}_t^{-1}\|_2$, we need at least $t = \Theta(1/p^4)$ time steps to collect enough information for reducing $\|\widehat{\Sigma}_t - \Sigma_t\|_2$ from Lemma 5. This is why we have $1/p^4$ in the regret bound. We believe that there may be some room to improve the dependency of $1/p^4$ by controlling $\|\widehat{\Sigma}_t^{-1}\|_2$, which can be a future work.

A.5 Algorithm 3 and its regret bound

Algorithm 3 Efficient Contextual Linear Bandits on Bayesian Features (E-CLBBF)

Input: τ'; **Initialize:** $Z \leftarrow \mathbf{0}_{d \times d}$; $\xi \leftarrow \mathbf{0}_{d \times 1}$; $n \leftarrow 0$
for $t = 1$ **to** τ' **do**
    Select a_t uniformly at random in $\mathcal{A}_t$
    Update $n \leftarrow n + $ the total number of non-missing entries in x_a for all $a \in \mathcal{A}$
    Update $Z \leftarrow Z + \sum_{a \in \mathcal{A}} x_{a,t} x_{a,t}^\top, \xi \leftarrow \xi + \sum_{a \in \mathcal{A}} x_{a,t}$
end for
for $t = \tau' + 1$ **to** T **do**
    Update $n \leftarrow n + $ the total number of non-missing entries in x_a for all $a \in \mathcal{A}$
    Update $Z \leftarrow Z + \sum_{a \in \mathcal{A}} x_{a,t} x_{a,t}^\top, \xi \leftarrow \xi + \sum_{a \in \mathcal{A}} x_{a,t}$
    Estimate parameters:
    $\widehat{p} \leftarrow \frac{\max\{1,n\}}{tdK}; \widehat{\nu} \leftarrow \frac{1}{tK\widehat{p}}\xi; \widehat{\Sigma} \leftarrow \frac{1}{tK} Z \circ \left(\frac{\widehat{p}-1}{\widehat{p}^2} I_{d \times d} + \frac{1}{\widehat{p}^2} \mathbf{1}_{d \times d} \right) - \widehat{\nu}\widehat{\nu}^\top$
    Estimate features $\widehat{z}_{a,t} \leftarrow [1; \overline{x}(\widehat{\nu}, \widehat{\Sigma}, x_{a,t})]$ for $a \in \mathcal{A}_t$.
    Select $a_t = \arg\max_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_{t-1}} \langle \widehat{z}_{a,t}, \theta \rangle$
    Observe reward y_t.
end for

We first define some notations in Algorithm 3. Let $\lambda = Cd\log(KT)$, $V_t = \lambda I_{d+1} + \sum_{s=\tau'+1}^t \widehat{z}_{a_s,s}\widehat{z}_{a_s,s}^\top$, and $\widehat{\theta}_t = V_t^{-1} \sum_{s=\tau'+1}^t \widehat{z}_{a_s,s}^\top y_s$ for some sufficiently large constant $C > 0$. We note that $\widehat{z}_{a_s,s}$ only for $s > \tau'$ are used in V_t and $\widehat{\theta}_t$. We also define the confidence set for estimating a latent parameter $\overline{\theta}$ as

$$\mathcal{C}_t = \{\theta \in \mathbb{R}^{d+1} : \|\widehat{\theta}_t - \theta\|_{V_t} \leq \sqrt{(d+1)\log((1+t)T)} + \lambda^{1/2}\}.$$

In Algorithm 1, $\widehat{z}_{a_s,s}$ in V_t and $\widehat{\theta}_t$ is required to be updated after time step s because estimated features in early stages may not be accurate. However, by introducing an explicit exploration phase for the first τ' time steps, Algorithm 3 does not use the early estimated features for V_t and $\widehat{\theta}_t$. Therefore, in Algorithm 3, $\widehat{z}_{a_s,s}$ in V_t and $\widehat{\theta}_t$ is computed at time s and it will not be updated anymore in later time steps.

The algorithm introduces an *explicit* exploration phase over τ' time steps, which guarantees the well-estimated features after the phase. Therefore, it does not require updating estimated features for previously selected arms, which is the main bottleneck of the Algorithm 1. However, setting a proper τ' for the exploration phase requires knowing an α that satisfies $\alpha \leq p$, and the regret bound depends on α. By setting $\tau' = \lceil d(\log(T))^2/(K\alpha^4)\rceil$, we can show that the regret bound of the efficient algorithm is $R(T) = \tilde{O}\left(d\sqrt{T} + (d^2/p^{3/2})\sqrt{T/K} + d/(\alpha^4 K)\right)$. The last term in the regret bound is larger than that in the regret of Theorem 2 due to the additional exploration phase in the efficient algorithm.

Now we provide a proof for the regret bound.

Regret bound of Algorithm 3   From Lemma 3, regret from $t = 1$ to τ' is bounded as

$$\sum_{t=1}^{\tau'} \mathbb{E}\left[\overline{x}(\nu_f, \Sigma, x_{a_t^\star,t})^\top \theta' - \overline{x}(\nu_f, \Sigma, x_{a_t,t})^\top \theta' \right] = O\left(\tau'\sqrt{\log(K)}\right).$$

Then from Lemma 5, for any time $t > \tau'$, with probability at least $1 - 1/T$, we have $\|\Sigma - \widehat{\Sigma}_t\|_2 = o(1)$, which incurs to have

$$\|\widehat{\Sigma}\|_2 \leq \|\widehat{\Sigma} - \Sigma\|_2 + \|\Sigma\|_2 = O\left(\|\Sigma\|_2\right) \quad \text{and}$$

$$\|\widehat{\Sigma}^{-1}\|_2 = \frac{1}{\sigma_d(\widehat{\Sigma})} \leq \frac{1}{\sigma_d(\Sigma) - \|\widehat{\Sigma} - \Sigma\|_2} = O\left(\frac{1}{\sigma_d(\Sigma)}\right) = O\left(\|\Sigma^{-1}\|_2\right).$$

Therefore, for all $t > \tau'$, we have

$$\|\widehat{z}_{a_t,t}\|_2 = O(\sqrt{d\log(KT)}).$$

Since we can get the bound for the estimated features after τ', the algorithm utilizes the estimated features of chosen arms only after τ' for getting V_t and $\widehat{\theta}_t$. Then by following the proof steps in Theorem 2, we can easily get

$$R(T) = \tilde{O}\left(d\sqrt{T} + \frac{d^2}{p^{3/2}}\sqrt{\frac{T}{K}} + \frac{d}{K\alpha^4}\right).$$
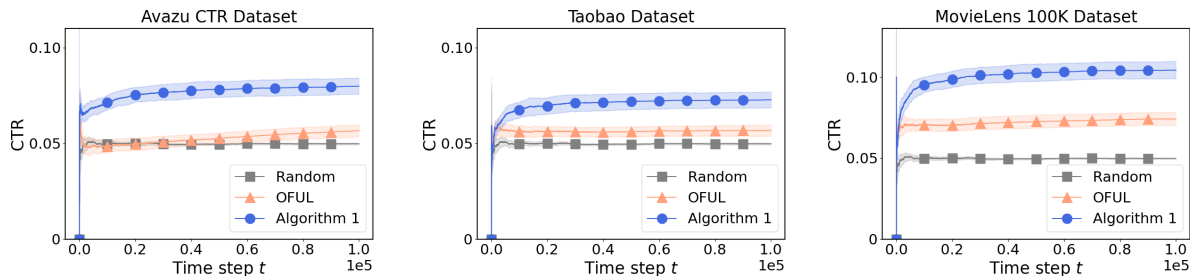
A.6 Real-world experiments

Figure 3: Click-through rate (CTR) for Algorithm 1, `OFUL`, and random policy for real-world datasets.

Here we present numerical results for real-world datasets. We compare our algorithm with `OFUL` and the random policy for click-through rate (CTR) at different time steps. For the comparison, we use Avazu CTR (Avazu, 2014), Taobao.com (Alibaba, 2018), and MovieLens 100K (Harper and Konstan, 2015) datasets that contain CTR or rating information collected from advertising or movie recommendation systems. We use autoencoder models to preprocess feature information, constructing item feature vectors for each item. We set preprocessed feature dimension $d = 32$, number of available items at each time, $K = 20$, and time horizon $T = 10^5$. We erase some entries in the preprocessed feature vectors using masking vectors with missing probability $1 - p = 0.1$. We examine average CTR for each algorithm over time steps $t \in [T]$. In Figure 3, we observe that our algorithm outperforms `OFUL` and the random policy for all the datasets *without generating both feature and noise vectors from Gaussian distributions externally*. In what follows, we describe the detailed experiment settings for real-world datasets.

Avazu CTR. This dataset contains mobile advertisement recommendation log data. Each data contains user, advertisement (ad), and click information. Each user-ad pair feature contains information about device type, site category, category of visited website, banner position, etc. For modeling reward payoffs, we use click information for each recommendation (user-ad); 0 for non-click and 1 for click. Using a pre-trained autoencoder model, we preprocess each user-ad feature vector to reduce dimension to 32 (output of the encoder), which we refer to as an item feature. The autoencoder is trained using user-ad feature information without reward information. Then, we erase each entry of the preprocessed item features with probability 0.1 for modeling the missing data scenario. We then divide the items (user-ad pairs) into two sets according to whether the ads are clicked or not; randomly select 1000 items to construct each set. For each time, algorithms get an available item set, which is constructed by randomly selecting $K = 20$ items from the two sets. In detail, the available item set consists of one selected from the clicked set and the others selected from the non-clicked set. Then, there must be one best item in the available item set. At each time, an algorithm selects an item from the available item set and get a reward according to the click information 0 or 1. Selecting an item can be represented as recommending an ad to a user.

Taobao.com. This dataset contains advertisement display/click log data on the website of Taobao.com. As in the Avazu dataset, each item (user-ad pair) contains user and ad information such as gender, age, consumption grade, brand, category, etc. Also there exist click data 1 or 0 for each recommendation. For each item feature, using a pre-trained autoencoder, we get a prepocessed feature vector with dimension 32. The rest of the experiment setting is the same as the case for the Avazu CTR dataset.

MovieLens 100K. This dataset contains movie ratings from users collected through the MovieLens website. Each rating has an integer value from 1 to 5. For modeling binary reward payoffs, we treat rating 5 as reward 1 and otherwise 0. Each user and movie data contain feature information such as age, gender, movie genre, etc. The rest of the experiment setting is the same as the experiment settings for the above datasets.