LA-UR-17-21865
Approved for public release; distribution is unlimited.

Title: Contextual Multi-armed Bandits under Feature Uncertainty

Author(s): Yun, Seyoung
Nam, Jun Hyun
Mo, Sangwoo
Shin, Jinwoo

Intended for: Web

Issued: 2017-03-03

Contextual Multi-armed Bandits under Feature Uncertainty

Se-Young Yun[1], Jun Hyun Nam[2], Sangwoo Mo[3], and Jinwoo Shin[4]

[1]Los Alamos National Laboratory, Los Alamos, NM, USA
[2,3,4]School of Electrical Engineering, KAIST, South Korea

Abstract

We study contextual multi-armed bandit problems under linear realizability on rewards and uncertainty (or noise) on features. For the case of identical noise on features across actions, we propose an algorithm, coined *NLinRel*, having $O\left(T^{\frac{7}{8}}\left(\log\left(dT\right)+K\sqrt{d}\right)\right)$ regret bound for T rounds, K actions, and d-dimensional feature vectors. Next, for the case of non-identical noise, we observe that popular linear hypotheses including *NLinRel* are impossible to achieve such sub-linear regret. Instead, under assumption of Gaussian feature vectors, we prove that a greedy algorithm has $O\left(T^{\frac{2}{3}}\sqrt{\log d}\right)$ regret bound with respect to the optimal linear hypothesis. Utilizing our theoretical understanding on the Gaussian case, we also design a practical variant of *NLinRel*, coined *Universal-NLinRel*, for arbitrary feature distributions. It first runs *NLinRel* for finding the 'true' coefficient vector using feature uncertainties and then adjust it to minimize its regret using the statistical feature information. We justify the performance of *Universal-NLinRel* on both synthetic and real-world datasets.

1 Introduction

The multi-armed bandit (MAB) problem (or simply bandit problem) has received much attention due to a wide range of applications, e.g., clinical trials Thompson (1933), economics Schlag (1998), routing Awerbuch & Kleinberg (2004), and ranking Radlinski et al. (2008). The MAB problems are sequential decision problems, where at each round, a learner selects an action (or arm) from K candidates and receives a reward for the selected action. The learner makes decisions based on the observations such as the sequence of past rewards and the selected actions, and would like to maximize the cumulative reward, or equivalently to minimize regret, defined as the difference between the cumulative reward and that achieved by always playing the best arm/action.

The learner often can access to contextual information in addition to rewards and selected actions, which are referred as *contextual* MAB Langford & Zhang (2008). Examples include personalized recommendation Bouneffouf et al. (2012), web server defense Jung et al. (2012) and information retrieval Hofmann et al. (2011). For instance, the learner see feature vectors $z_1(t), z_2(t), \ldots, z_K(t) \in \mathbb{R}^d$ associated with each of K arms at every round t. To address the problem, one has to assume a hypothesis set consisting of functions from the feature vectors to an action that will give the best expected reward. The *linear* hypothesis set is simple and widely used, where each hypothesis is defined by a coefficient vector $\theta \in \mathbb{R}^d$ and predicts an optimal action as $\arg\max_{1\le i\le K} z_i(t)^\top \theta$. The linear hypothesis set assumes that the expected reward of action i at round t is defined by $z_i(t)^\top \theta^\star$ with a hidden coefficient vector $\theta^\star$ which is referred to as *linear payoff* and also called *linear realizability*. It is an online linear regression task balancing the trade-off between exploration and exploitation.

In this paper, we study the contextual MAB problems with linear payoffs under assuming uncertainty or noise on features. Specifically, we assume that the learner cannot observe the true feature vector $z_i(t)$ but noisy vector $x_i(t) = z_i(t) + \varepsilon_i(t)$ where random noise $\varepsilon_i(t)$ is independently drawn from some distribution. It can incorporate statistical uncertainties of linear hypotheses, and relax the strong linear assumption on rewards, i.e., enhance the power of linear models. Furthermore, it can incorporate recent remarkable progresses in Bayesian deep learning techniques Gal & Ghahramani (2016) that estimate feature uncertainties, i.e., the knowledge of noise distributions.

There are two main challenges in the noisy contextual MAB problem:

$\mathcal{I}1$. The learner might not extract the true hypothesis $\theta^\star$ from any sequence of observations using policies defined for the noiseless contextual MAB problem, e.g. *LinRel* Auer (2002).

$\mathcal{I}2$. Even if the learner could learn $\theta^\star$, it is still hard to design a good exploitation policy since every arm has some uncertainty due to the noise.

Therefore, we need to redesign learning policies considering the noisy feature vectors. To the best of our knowledge, this is the first work that aims to solve the contextual MAB problems under assuming such uncertainties on features.

Contribution. We first study the simplest, but non-trivial, case that every action has the identical noise vector, i.e., $\varepsilon_i(t) = \varepsilon(t)$ for all $i \in \{1, 2, \ldots, K\}$. This eliminates the issue $\mathcal{I}2$: the learner can find the best action after extracting the hidden coefficient vector since

$$\arg\max_{1\le i\le K} z_i(t)^\top\theta = \arg\max_{1\le i\le K} x_i(t)^\top\theta.$$

However, the issue $\mathcal{I}1$ remains, e.g., *LinRel* might not find $\theta^\star$. Furthermore, one has to design a new confidence interval due to the noise for balancing the exploration and exploitation trade-off. To address them, we propose a noisy version of *LinRel*, coined *NLinRel*, having regret bound $O\left(T^{\frac{7}{8}}\left(\log(dT) + K\sqrt{d}\right)\right)$ for T rounds. For the regret analysis, we use the tail inequalities of random matrices induced by noise vectors and bound the random matrix perturbation.

We next consider non-identical noise vectors, but assume that the feature vector at each round is independently drawn from some distribution $\mathcal{D}_{\texttt{feature}}$. The underlying reasoning for the statistical assumption is on our finding that this eliminates the issue $\mathcal{I}2$. Specifically, if $\mathcal{D}_{\texttt{feature}}$ is Gaussian, we derive a closed form formula of the optimal coefficient vector $\bar{\theta}$ using Bayesian analysis. Somewhat interestingly, the optimal $\bar{\theta}$ is not equal to the true $\theta^\star$, which cannot occur under noiseless settings. We further design a simple greedy algorithm that achieves $O\left(T^{\frac{2}{3}}\sqrt{\log d}\right)$ regret bound with respect to the optimal linear hypothesis. Here, one can easily observe that any linear hypothesis including the greedy algorithm and *NLinRel* cannot achieve a sub-linear regret with respect to optimal sequence of actions, and thus we analyze such a 'relative' regret. Our study on Gaussian features naturally motivates the question of whether $\bar{\theta}$ is also optimal for general feature distributions. To this end, we derive an optimization formulation (i.e., non-closed form) for the optimal coefficient vector for general, possibly non-Gaussian, setting, and numerically found that $\bar{\theta}$ is no longer optimal in this case. Finally, we design a new algorithm, coined *Universal-NLinRel*, for arbitrary distributions on features, where it searches the true $\theta^\star$ using *NLinRel* and adjusts the parameter to a gradient direction of the optimization objective. In our experiments, *Universal-NLinRel* outperforms *LinUCB* Chu et al. (2011), representing the known linear hypothesis designed for the noiseless contextual MAB problem, on both noisy synthetic and real-world datasets.

Related works. Although the name, contextual multi-armed bandit, first appeared in Langford & Zhang (2008), the problem setting has been studied under different names, e.g., bandit with covariates Woodroofe (1979); Sarkar (1991), associative reinforcement learning Kaelbling (1994), associative bandit Auer (2002); Strehl et al. (2006) and bandit with expert advice Auer et al. (2002). This paper, in particular, focuses on the linear hypothesis set and the linear payoff model, which was originally introduced in Abe & Long (1999) and developed in Auer (2002). Our algorithm design of *NLinRel* is actually motivated by *LinRel* Auer (2002) and *LinUCB* Chu et al. (2011). Both *LinRel* and *LinUCB* algorithms compute the expected rewards and their confidence intervals for controlling the exploration and exploitation trade-off. *Thompson sampling* was also studied for the linear payoff model Agrawal & Goyal (2013). The stochastic linear bandit optimization problem studied in Dani et al. (2008) and many following works are special cases of the contextual bandit with the linear payoff model having infinitely many arms. However, all the studies assume that the feature vectors are noiseless and we cannot directly apply their algorithms to our noise setting.

One can discretize the linear hypothesis set into an ε-net, $\mathcal{H}_\varepsilon$, such that $\|\theta_1 - \theta_2\|_2 \geq \varepsilon$ for all $\varepsilon_1 \neq \varepsilon_2 \in \mathcal{H}_\varepsilon$. With the ε-net, it is possible to use *EXP4*-type algorithms Auer et al. (2002); Beygelzimer et al. (2011) for the noisy contextual MAB problem studied in this paper. However, the computation costs of the algorithms are extremely expensive to use. The size of the ε-net is $\Omega\left(\varepsilon^{-(d-1)}\right)$ and *EXP4*-type algorithms have to update weights of all elements of the ε-net at every round. One can also possibly use *Epoch-Greedy* Langford & Zhang (2008) for our Gaussian setting mentioned earlier, but it also requires a huge amount of computations and memory space when computing the most likelihood hypothesis among the hypothesis set. If one uses the ε-net, *Epoch-Greedy* has the same issue with the *EXP4*-type algorithms, or all sequence of observations should be memorized to compute the maximum likelihood.

2 Preliminaries

We study a noisy version of the contextual multi-armed bandit (MAB) problem with linear payoffs. At each time $t = 0, 1, \ldots$, there are K possible actions and the learner observes a feature vector $x_i(t) \in \mathbb{R}^d$ for each possible action $1 \leq i \leq K$, i.e., the dimension of features is d and the number of arms is K. We assume that the observed features are noisy in the sense that the true hidden feature vector of action i at time t is denoted by $z_i(t) = x_i(t) - \varepsilon_i(t)$, for some independent random vector $\varepsilon_i(t)$ with $\mathbb{E}[\varepsilon_i(t)] = \mathbf{0}$. The learner selects an action $1 \leq a(t) \leq K$ and observe the reward $y(t) \in R$ for the selected action at time t. We assume that $y(t)$ is an independent sub-Gaussian random variables with finite variance and its distribution is determined by the true feature vector of the selected arm as

$$\mathbb{E}\left[y(t) | z_i(t) = z, a(t) = i\right] = z^\top \theta^\star.$$

This is called the linear payoff assumption, where there is a coefficient vector $\theta^\star \in \mathbb{R}^d$ with $\|\theta^\star\|_2 \leq 1$ is unknown to the learner a priori.

The learner uses some algorithm or policy selecting an action $a(t)$ at each time t given current observed feature vectors $x_i(t)$ for all action i and past information $a(s), x_{a(s)}(s), y(s)$ for all time $s < t$ so that it maximizes the cumulative reward $\sum_{t=1}^T y(t)$ up to time T. If the learner knows hidden information $\theta^\star$ and $z_i(t)$, the best choice of action to maximize the cumulative reward would be $a^\star(t) = \arg\max_{1\leq i\leq K} z_i(t)^\top \theta^\star$. We define a regret function of algorithm $\mathcal{A}$ compared to the oracle algorithm as follows:

$$R(T) = \sum_{t=1}^{T} \max_{1\leq i\leq K} z_i(t)^\top \theta^\star - z_{a(t)}(t)^\top \theta^\star \tag{1}$$

The objective of the learner's algorithm is to minimize the above regret.

Notation. Here, we define necessary matrix notation used throughout this paper. For any matrix $A \in \mathbb{R}^{m\times n}$, $A^\top$ and A^{-1} denote the transpose and inverse of A, respectively. The i-th eigenvalue and the i-th singular value of A are denoted by $\lambda_i(A)$ and $s_i(A)$, respectively. Let A_i denote the i-th column of A and $A_{i:j} = [A_i, A_{i+1}, \ldots, A_j]$ for $i \leq j$. We mean $\Delta(A, i)$ by the i-th diagonal value of A and $\Delta(A, i:j)$ by the $(j-i+1) \times (j-i+1)$ diagonal matrix consisting of $\Delta(A, i), \Delta(A, i+1), \ldots, \Delta(A, j)$.

3 Identical Feature Uncertainty

In this section, we assume every action shares the same noise feature vector $\varepsilon(t)$, i.e., $\varepsilon_i(t) = \varepsilon(t)$ for all $1 \le i \le K$. We also assume that $\{\varepsilon(t)\}_{t\ge1}$ are i.i.d. random vectors with $\mathbb{E}[\varepsilon(t)] = \mathbf{0}$ and the covariance of $\varepsilon(t)$, denoted by $\mathcal{E}(t) = \mathbb{E}[\varepsilon(t)\varepsilon(t)^\top]$, is known to the learner. As in Auer (2002), for the analysis, we assume that all feature vectors satisfy $\|z_i(t)\|_2 \le 1$ and the rewards $r_i(t)$ are bounded by a finite constant. Furthermore, we assume that the distribution of $\varepsilon(t)$ has a finite support.

Under the identical uncertainty assumption, we have

$$\arg\max_{1\le i\le K} z_i(t)^\top \theta^\star = \arg\max_{1\le i\le K} x_i(t)^\top \theta^\star, \quad \forall t \ge 1.$$

Hence, one can find the best action in terms of the expected reward after finding $\theta^\star$ even with noises on feature vectors. Namely, one can reduce the regret $R(T)$ by learning hidden coefficient vector $\theta^\star$ accurately. However, $R(T)$ could increase quickly if we spend too much time to learn $\theta^\star$, which is the popular exploitation-exploitation trade-off issue in the bandit problem.

3.1 Issues of *LinRel* under Noisy Features

When there is no noise on feature vectors, *LinRel* by Auer (2002) controls the exploitation-exploitation tradeoff very efficiently and guarantees a sub-linear $R(T)$ with resect to T. It executes the following procedures at round/time $t+1$:

1. Calculate

$$X(t) = \sum_{\tau=1}^{t} x_{a(\tau)}(\tau) x_{a(\tau)}(\tau)^\top$$
$$Y(t) = \sum_{\tau=1}^{t} y(\tau) x_{a(\tau)}(\tau)$$
$$\overline{X}(t)^{-1} = U_{1:k}\Delta(\Lambda, 1:k)^{-1} U_{1:k}^\top,$$

   where Λ is from the eigenvalue decomposition of $X(t) = U\Lambda U^\top$, and k is the index such that $\Delta(\Lambda, k) \ge 1$ and $\Delta(\Lambda, k+1) < 1$.

2. Compute $\widehat{\theta}(t) = \overline{X}(t)^{-1}Y(t)$, which is an estimator of $\theta^\star$. Then, compute the expected reward and the width of the confidence interval as follows: for all $1 \le i \le K$,

$$r_i(t) = z_i(t)^\top \widehat{\theta}$$
$$w_i(t) = \|U(t)_{k+1:d}^\top x_i(t)\|_2 + \sqrt{\|x_i(t)^\top \overline{X}^{-1} x_i(t)\|_2 \log(KT/\delta)}.$$

3. Select $a_i(t+1) = \arg\max_{1\le i\le K} r_i(t) + w_i(t)$. The expected reward $r_i(t)$ controls the exploitation and the width of the confidence interval $w_i(t)$ controls the exploration.

The following two facts of *LinRel* make $R(T)$ becomes sub-linear. First, one can check that $|x_i(t)^\top(\widehat{\theta} - \theta^\star)| \le w_i(t)$ for all $1 \le i \le K$ and $1 \le t \le T$. From this and the selection rule, we have $R(T) = O\left(\sum_{t=1}^T w_{a(t)}(t)\right)$. Second, it holds that $\sum_{t=1}^T w_{a(t)}(t) = O\left(\sqrt{T\log(KT)}\right)$. In other words, the sum of uncertainties of the observed arms increases sub-linearly. This is because *NLinRel* has a better estimation on $\theta^\star$ as playing actions having high uncertainties. Intuitively, $w_i(t)$ is the amount of information revealed by playing action i.

However, *LinRel* has very bad regret $R(t)$ under noise feature vectors. First, $\widehat{\theta}$ does not converges to $\theta^\star$. One can easily check that

$$\mathbb{E}[X(t)] = \sum_{\tau=1}^{t} \left(z_{a(t)}(t) z_{a(t)}(t)^\top + \mathcal{E}(t) \right)$$

$$\mathbb{E}[Y(t)] = \sum_{\tau=1}^{t} z_{a(t)}(t) z_{a(t)}(t)^\top \theta^\star$$

We have to remove $\mathcal{E}(t)$ to expect $\widehat{\theta} = X(t)^{-1}Y(t)$ to converge to $\theta^\star$. Second, the uncertainty indicator $w_i(t)$ strongly depends on $\varepsilon(t)$ and does not indicate the amount of information we can obtain from the action.

3.2 Redesigning *LinRel* for Noisy Features

In this section, we redesign *LinRel*, which is referred to as *NLinRel* and described formally in what follows.

Algorithm 1 NLinRel

$Z(0) \leftarrow \mathbf{0}_{d\times d}, Y(0) \leftarrow \mathbf{0}_{d\times 1}$
for $t = 1$ **to** T **do**
  $U\Sigma U^\top \leftarrow$ Eigenvalue decomposition of $Z(t-1)$
  $k \leftarrow \max\{i : \Delta(\Sigma, i) \geq t^\alpha, 1 \leq i \leq d\}$
  $\hat{\theta} \leftarrow U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top Y(t-1)$
  for all $i \in K$ **do**
    $r_i \leftarrow x_i(t)^\top \hat{\theta}$, $w_{i,j} \leftarrow \left\| U_{k+1:d}^\top (x_i(t) - x_j(t)) \right\|_2$
  end for
  $C \leftarrow \emptyset$, $R \leftarrow \emptyset$
  while $|(K \setminus C) \setminus R| > 0$ **do**
    $C \leftarrow C \cup \{c\}$ for $c \leftarrow \arg\max_{i\in(K\setminus C)\setminus R} a_i$
    $R \leftarrow R \cup R'$ for $R' \leftarrow \{i : a_i + w_{i,c} \leq a_c\}$
  end while
  $\hat{i}_t \leftarrow$ select from C uniformly at random
  $Z(t) \leftarrow Z(t-1) + x_{a(t)}(t)x_{a(t)}(t)^\top - \mathcal{E}(t)$
  $Y(t) \leftarrow Y(t-1) + x_{a(t)}(t)y(t)$
end for

We also prove that the above algorithm achieves the sub-linear regret bound.

Theorem 1 *Under the above identical noisy contextual MAB model, NLinRel has*

$$R(T) = O\left(T^{\frac{3}{2}-\alpha}\sqrt{\log\left(dT/\delta\right)} + KT^{(\alpha+2)/3}\sqrt{d}\right),$$

with probability at least $1-\delta$.

The proof of the above theorem is given in Appendix B. For instance, when $\alpha = 5/8$, the regret is

$$R(T) = O\left(T^{\frac{7}{8}}\left(\sqrt{\log\left(dT/\delta\right)} + K\sqrt{d}\right)\right).$$

In what follows, we provide our strategies on the algorithm design and the regret proof.

Redesigning components. For designing *NLinRel*, we introduce

$$Z(t) = \sum_{\tau=1}^{t} x_{a(t)}(t) x_{a(t)}(t)^\top - \mathcal{E}(t)$$

so that $\mathbb{E}[\sum_{\tau=1}^{t} y(\tau)x_{a(\tau)}(\tau)] = \mathbb{E}[Z(t)]\theta^\star$ and $\widehat{\theta} \to \theta^\star$ as $t \to \infty$. Let $U\Sigma U^\top$ be the eigenvalue decomposition of $Z(t)$. For given $\alpha > 1/2$, we use t^α as a threshold for the eigenvalues and let k be the largest index such that $\Delta(\Sigma, k) \geq t^\alpha$. We then estimate the coefficient vector and expected rewards of actions as follows:

$$\widehat{\theta} = \overline{Z}(t)^{-1} Y(t) \quad \text{and} \quad r_i(t) = x_i(t)^\top \widehat{\theta} \quad \forall i,$$

where $\overline{Z}(t)^{-1} = U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top$ and $Y(t) = \sum_{\tau=1}^{t} y(\tau)x_{a(\tau)}(\tau)$ as in *LinRel*.

Since the width of confidence interval for a single action $w_i(t)$ is not a good indicator to control the exploration due to noise vectors, we re-define the width of confidence interval for each action-pair (i, j), which is referred to $w_{i,j}(t)$ and approximates $|(x_i(t) - x_j(t))^\top(\widehat{\theta} - \theta^\star)|$. More precisely, we compute

$$w_{i,j}(t) = \|U_{k+1:d}^\top (x_i(t) - x_j(t))\|_2 \quad \forall 1 \leq i, j \leq K.$$

Note that $\varepsilon(t)$ is removed by $x_i(t) - x_j(t)$ and thus $w_{i,j}(t)$ is independent to the noise.

The exploration and exploitation tradeoff is controlled by $r_i(t)$ and $w_{i,j}(t)$. A candidate set C is generated so that $|r_i(t) - r_j(t)| \leq w_{i,j}(t)$ for all $i, j \in C$ and there exists $j \in C$ such that $r_j(t) - w_{i,j}(t) > r_i(t)$ for all $i \notin C$. Then, $a(t)$ is selected uniformly at random from C.

Proof strategy for Theorem 1. In *NLinRel*, $w_{a(t),a^\star(t)}(t)$ indicates the uncertainty between the best arm and the selected arm at t. The uncertainty is roughly proportional to the amount of revealed information by playing an arm $a(t)$. From that, we first bound the expected sum of uncertainties as follows:

$$\mathbb{E}\left[\sum_{t=1}^{T} w_{a(t),a^\star(t)}(t)\right] = O\left(T^{(\alpha+2)/3}\right).$$

We then connect the expected sum of uncertainties to the regret that has

$$\begin{aligned} R(T) =& O\left(\mathbb{E}\left[\sum_{t=1}^{T} w_{a(t),a^\star(t)}(t)\right] + t^{-\alpha+3/2}\right) \\ =& O\left(T^{(\alpha+2)/3} + t^{-\alpha+3/2}\right). \end{aligned}$$

In the above equation, we have an additional term $t^{-\alpha+3/2}$, which stems from the fact that $w_{i,j}(t)$ is just an approximation of $|(x_i(t) - x_j(t))^\top(\widehat{\theta} - \theta^\star)|$. From the definition of $\widehat{\theta}$, we have

$$\begin{aligned} \theta^\star - \widehat{\theta} =& \theta^\star - \overline{Z}(t)^{-1}Z(t)\theta^\star - \overline{Z}(t)^{-1}(Y(t) - Z(t)\theta^\star) \\ =& U_{k+1:d}U_{k+1:d}^\top\theta^\star - \overline{Z}(t)^{-1}(Y(t) - Z(t)\theta^\star), \end{aligned}$$

where the last term was not considered when we compute $w_{i,j}(t)$. We can bound the last term using a tail bound for sums of random matrices Tropp (2012). More precisely, using the matrix Azuma inequality, we show that $\|\overline{Z}(t)^{-1}(Y(t) - Z(t)\theta^\star)\|_2 = O(t^{-\alpha+1/2})$.

4 Non-identical Feature Uncertainty

When the noise feature vectors are not identical, i.e., $\varepsilon_i(t) \neq \varepsilon_j(t)$, any algorithm based on a linear hypothesis is impossible to guarantee a sub-linear regret function. The regret function can grow linearly even though we know the hidden coefficient vector $\theta^\star$ exactly. To see why, suppose each $\varepsilon_i(t)$ is drawn under a normal distribution. Then, there exists some constant $\delta > 0$ such that for any given set of feature vectors $\{z_i(t)\}_{1\le i\le K}$, with probability δ, $\arg\max_{1\le i\le K} z_i(t)^\top\theta^\star \neq \arg\max_{1\le i\le K} x_i(t)^\top\theta^\star$. For coefficient vector $\theta \in \mathbb{R}^d$, let $R_\theta(t)$ be the corresponding expected regret function when the learner decides actions as follows:

$$a_\theta(t) = \arg\min_{1\le i\le K} x_i(t)^\top\theta.$$

In this section, we do not aim for designing an algorithm of a sub-linear regret with respect to the optimal sequence $\{a^\star(t)\}_{t\ge1}$, but study how to find an optimal linear hypotheis θ that minimizes $R_\theta(t)$. Somewhat interestingly, we found that the choice $\theta = \theta^\star$ is not always the best, i.e., there could exist $\theta \neq \theta^\star$ such that $R_\theta(t) > R_{\theta^\star}(t)$ for all $t \ge 1$. In order to describe the intuition why $\theta^\star$ is not the optimal choice, we first consider a noisy Gaussian contextual MAB model in Section 4.1. Under the model, $\overline{\theta}$ that minimizes $R_\theta(t)$ is represented in a closed form and we prove that a very simple algorithm has a 'relative' sub-linear regret bound with respect to the optimal linear hypothesis. The optimal closed form for Gaussian models might no longer be true for non-Gaussian ones, which is discussed in Section 4.2.

4.1 Gaussian Features

In this section, we consider the following noisy Gaussian contextual MAB model. The true feature vectors $\{z_i(t) : i \in K, 1 \le t \le T\}$ are i.i.d. random vectors drawn from the normal distribution $\mathcal{N}(\mathbf{0}_{d\times1}, \Sigma_{\texttt{feature}})$ where $\Sigma_{\texttt{feature}}$ is a $d \times d$ positive-definite matrix. The noises are defined by i.i.d. multivariate Gaussian random vectors as well: for all $i \in K$ and $1 \le t \le T$, $\varepsilon_i(t)$ follows $\mathcal{N}(\mathbf{0}_{d\times1}, \Sigma_{\texttt{noise}})$ with a $d \times d$ positive-definite matrix $\Sigma_{\texttt{noise}}$. Since both $\Sigma_{\texttt{feature}}$ and $\Sigma_{\texttt{noise}}$ are positive-definite matrices, we have inverse matrices not only for $\Sigma_{\texttt{feature}}$ and $\Sigma_{\texttt{noise}}$ but also for $(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})$ and $(\Sigma^{-1}_{\texttt{feature}} + \Sigma^{-1}_{\texttt{noise}})$.
Optimal linear hypothesis. We would like to find $\overline{\theta}$ such that $R_\theta(t) \ge R_{\overline{\theta}}(t)$ for all $t \ge 1$ and $\theta \in \mathbb{R}^d$. The following theorem obtains a closed form of an optimal choice $\overline{\theta}$.

Theorem 2 *Under the noisy Gaussian contextual MAB model, $R_\theta(t) \ge R_{\overline{\theta}}(t)$ for all $t \ge 1$ and $\theta \in \mathbb{R}^d$, where*

$$\overline{\theta} := (\Sigma_{\textit{feature}} + \Sigma_{\textit{noise}})^{-1}\Sigma_{\textit{feature}}\theta^\star. \quad (2)$$

The proof of the above theorem is provided in Appendix C. Here we provide its high-level sketch. At each round, the learner receives noisy feature vectors $x_1(t), \dots, x_K(t)$. When one knows $\theta^\star$ and the distributions of $z_i(t)$ and $\varepsilon_i(t)$ for all i, the optimal decision from the given feature vectors can be computed as:

$$\begin{aligned} i^\star(t) &= \arg\max_{1\le i\le K} \mathbb{E}[y_i(t)|x_1(t), \dots, x_K(t)] \\ &= \arg\max_{1\le i\le K} \mathbb{E}[z_i(t)|x_i(t)]^\top\theta^\star, \end{aligned}$$

where the last equality comes from the independence between actions and the linearity of the expectation. In the proof of Theorem 2, we obtain using Bayesian analysis that

$$\begin{aligned} &\mathbb{E}[z_i(t)|x_i(t)]^\top\theta^\star \\ &\quad = x_i(t)^\top(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}\Sigma_{\texttt{feature}}\theta^\star. \end{aligned}$$

Therefore, one can easily find the optimal action with $\overline{\theta} = (\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}\Sigma_{\texttt{feature}}\theta^\star$ under

$$i^\star(t) = \arg\max_{1\le i\le K} x_i(t)^\top \overline{\theta}.$$

From the optimal action, we define the following 'relative' regret function:

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

Greedy algorithm. We now propose a very simple greedy algorithm that operates only with observations $a(t)$, $x_{a(t)}(t)$, $y(t)$ and can find $\overline{\theta}$ very accurately. The simple greedy algorithm consists of two parts, each for *exploration* and *exploitation*, as stated formally in what follows.

Algorithm 2 Simple greedy algorithm

$\tau \leftarrow \lfloor T^{\frac{2}{3}} \rfloor$, $X(0) \leftarrow \mathbf{0}_{d\times d}$, $Y \leftarrow \mathbf{0}_{d\times 1}$
for $t = 1$ **to** τ **do**
  Randomly select $a(t)$
  $X(t) \leftarrow X(t-1) + x_{a(t)}(t)x_{a(t)}(t)^\top$
  $Y(t) \leftarrow Y(t-1) + x_{a(t)}(t)y_{a(t)}(t)$
end for
$\widehat{\theta} = X^{-1}Y$
for $t = \tau + 1$ **to** T **do**
  $a(t) \leftarrow \arg\max_{i\in K} x_i(t)^\top\widehat{\theta}$
end for

The first τ selections of the above algorithm are used for the exploration to learn $\widehat{\theta} = X(\tau)^{-1}Y(\tau)$, where $X(\tau) = \sum_{t=1}^{\tau} x_{a(t)}x_{a(t)}^\top$ and $Y(\tau) = \sum_{t=1}^{\tau} y(t)x_{a(t)}$. The remaining selections exploit $\widehat{\theta}$ for their decision:

$$a(t) = \arg\max_{1\le i\le K} x_i(t)^\top\widehat{\theta}.$$

Observe that the above algorithm do not utilize the information $\Sigma_{\texttt{feature}}$, $\Sigma_{\texttt{noise}}$ and $\theta^\star$. Nevertheless, we indeed show that it finds the optimal $\overline{\theta}$ and a sub-linear regret $\overline{R}(T)$.

Theorem 3 *Under the above noisy Gaussian contextual MAB model, the simple greedy algorithm has*

$$\overline{R}(T) = O\left(T^{\frac{2}{3}}\sqrt{\log\left(\frac{d}{\delta}\right)}\right),$$

with probability $1-\delta$.

The proof of the above theorem is provided in Appendix D. Here we provide its high-level sketch. In the exploration part, at each time instance, the learner selects an action uniformly at random so that the selection and noisy feature vectors become independent. Then, $z_{a(\tau)}(1), \ldots, z_{a(\tau)}(\tau)$ are i.i.d. random vectors following $\mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{feature}})$ and $\varepsilon_{a(\tau)}(1), \ldots, \varepsilon_{a(\tau)}(\tau)$ are also i.i.d. random vectors following $\mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{noise}})$. Therefore, we have

$$\begin{aligned}
\mathbb{E}[X(\tau)] =& \mathbb{E}\left[\sum_{t=1}^{\tau} z_{a(t)}z_{a(t)}^\top\right] + \mathbb{E}\left[\sum_{t=1}^{\tau} \varepsilon_{a(t)}\varepsilon_{a(t)}^\top\right] \\
=& \tau\left(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}}\right) \\
\mathbb{E}[Y(\tau)] =& \mathbb{E}\left[\sum_{t=1}^{\tau} z_{a(t)}z_{a(t)}^\top\theta^\star\right] = \tau\Sigma_{\texttt{feature}}\theta^\star.
\end{aligned}$$

From the matrix Azuma inequality, we show that residual matrices $X(\tau) - \mathbb{E}[X(\tau)]$ and $Y(\tau) - \mathbb{E}[Y(\tau)]$ are negligible compared with $X(\tau)$ and $Y(\tau)$, respectively, with respect to their spectral norms. From the facts, we derive

$$\widehat{\theta} = X(\tau)^{-1}Y(\tau) \approx (\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}\Sigma_{\texttt{feature}}\theta^\star.$$

This will leads to the conclusion of Theorem 3.

4.2 Non-Gaussian Features

In this section, we consider that the true feature vectors drawn under an arbitrary, possibly non-Gaussian, distribution in this case, the proof of Theorem 2 is no longer true and it is not easy to analyze whether $\overline{\theta}$ defined in (2) is optimal in any sense. Formally, we assume that $\{z_i(t) : i \in K, 1 \leq t \leq T\}$ are i.i.d. random vectors drawn from some (possibly, non-Gaussian) distribution $\mathcal{D}$ with mean $\mathbf{0}_{d\times 1}$ and covariance $\Sigma_{\texttt{feature}}$ where $\Sigma_{\texttt{feature}}$ is a $d \times d$ positive-definite matrix. The noise model is same as that of Gaussian contextual MAB model in the previous section. We focus on verifying numerically whether $\overline{\theta}$ defined in (2) is optimal under the the non-Gaussian setting.

To this end, one can observe that the optimal θ minimizes the following given the information $\theta^\star$, $\mathcal{D}$ and $\Sigma_{\texttt{noise}}$:

$$\begin{aligned} R_\theta(t) - R_\theta(t-1) &= \mathbb{E}\left[z_{a^\star(t)}\theta^\star - z_{a_{\theta(t)}}\theta^\star\right] \\ &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{noise}})}\left[z_{a^\star(t)}\theta^\star - z_{a_\theta(t)}\theta^\star \mid z(t)\right]\right]. \end{aligned} \tag{3}$$

The solution of this optimization might not be given as a closed form as like (2) unless $\mathcal{D}$ is Gaussian/normal. Furthermore, computing a gradient is a non-trivial task depending on $\mathcal{D}$ and the knowledge of $\mathcal{D}$ might not be given in practical scenarios. Hence, we estimate it via the following Monte Carlo method:

$$\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta \mathbb{E}_{\mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{noise}})}\left[z_{a^\star(t)}\theta^\star - z_{a_\theta(t)}\theta^\star \mid z(t) = z^{(i)}\right], \tag{4}$$

where $z^{(1)}, z^{(2)}, \ldots, z^{(N)}$ are randomly generated samples from the distribution $\mathcal{D}$ or real feature vectors observed in practice. It is elementary to check that each gradient in (4) can be expressed as an integral form with respect to the probability density function of $\mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{noise}})$.

Under several different choices of feature distribution $\mathcal{D}$, we compute (4) at $\theta = \overline{\theta}$ to confirm whether it is optimal or not. In all the experiments, the number of arm $K = 5$, the dimension of feature $d = 10$, the number of samples $N = 10^6$, and each element in the feature vector is an i.i.d. random variable. We also choose each element of $\theta^\star$ uniformly at random in the interval $[-1, 1]$, i.e., Uniform(-1,1). For the distribution of noise, we use $\mathcal{N}(\mathbf{0}_{10\times 1}, \text{diag}(0.1, 0.2, \cdots, 1.0))$. The numerical results are reported in Table 1, which implies that (2) might be far from being optimal unless $\mathcal{D}$ is Gaussian.

This motivates to design a new algorithm, completely different from Algorithm 2, for non-Gaussian feature distribution $\mathcal{D}$. For the purpose, we propose the following algorithm, called *Universal-NLinRel*.

Table 1: ℓ_2-norms of gradients at $\theta = \overline{\theta}$

Feature distribution $\mathcal{D}$	ℓ_2-norm of gradient
Gaussian(0,1)	0.000
Uniform(-1,1)	0.013
Laplace(0,1)	0.032
Exponential(1)	0.413
LogNormal(0,1)	0.648
Mixture of Gaussian[1]	0.320
Mixture of Uniform[2]	0.273

[1] 0.3 * Gaussian(10,1) + 0.7 * Gaussian(-10,1)
[2] 0.3 * Uniform(9,11) + 0.7 * Uniform(-11,-9)

Algorithm 3 Universal-NLinRel

$Z(0) \leftarrow \mathbf{0}_{d\times d}, Y(0) \leftarrow \mathbf{0}_{d\times 1}$
Randomly select initial θ
for $t = 1$ **to** T **do**
  $U\Sigma U^\top \leftarrow$ Eigenvalue decomposition of $Z(t-1)$
  $k \leftarrow \max\{i : \Delta(\Sigma, i) \geq t^\alpha, 1 \leq i \leq d\}$
  $\theta^\dagger \leftarrow U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top Y(t-1)$
  Randomly sample z from $\mathcal{D}$
  $\theta \leftarrow \theta - \alpha \cdot \nabla_\theta \mathbb{E}\left[z_{a^\star(t)}\theta^\dagger - z_{a_\theta(t)}\theta^\dagger \,\middle|\, z(t) = z\right]$
  $a(t) \leftarrow \arg\max_{i\in K} x_i(t)^\top\theta + \alpha \left\|U_{k+1:d}^\top x_i(t)\right\|_2$
  $Z(t) \leftarrow Z(t-1) + x_{a(t)}(t)x_{a(t)}(t)^\top - \mathcal{E}(t)$
  $Y(t) \leftarrow Y(t-1) + x_{a(t)}(t)y(t)$
end for

In the above, $\alpha \geq 0$ is some *UCB*-like constant as like *LinUCB* Chu et al. (2011). The main idea on the algorithm design is that it runs *NLinRel* for estimating the true coefficient vector $\theta^\star$, and then update the current θ to a stochastic gradient direction by replacing $\theta^\star$ by the estimation $\theta^\dagger$. Although *NLinRel* has its theoretical value, *Universal-NLinRel* uses a practical variant of *NLinRel* by introducing parameter α since too many initial explorations might hurt its regret unless an extremely large enough number of time instances is allowed. In the following section, We measure the regret performance of *Universal-NLinRel*.

5 Experimental Results

In this section, we report experimental results comparing the regret performances of *Univeral-NLinRel* with the following algorithms. First, *LinUCB* Chu et al. (2011) represents known algorithms designed for the noiseless contextual MAB problem.[1] Seocnd, *Oracle-GD* is identical to *Univeral-NLinRel*, except for using the true coefficient vector $\theta^\star$ instead of the estimated one $\theta^\dagger$. Finally, *Oracle-TC* and *Oracle-CF* are linear hypotheses choosing arm $a(t) \in \arg\min_{1\leq i\leq K} x_i(t)^\top\theta$ where they consider the true coefficient vector $\theta = \theta^\star$ and the closed form $\theta = \bar{\theta}$ defined in (2), respectively.
Synthetic dataset. We follow the same synthetic setups described in Section 4.2, and the experimental

[1] We choose $\alpha = 0.25$ for both *LinUCB* and *Universal-NLinRel* in all our experiments, but the choice is not sensitive for their performances in all our settings.

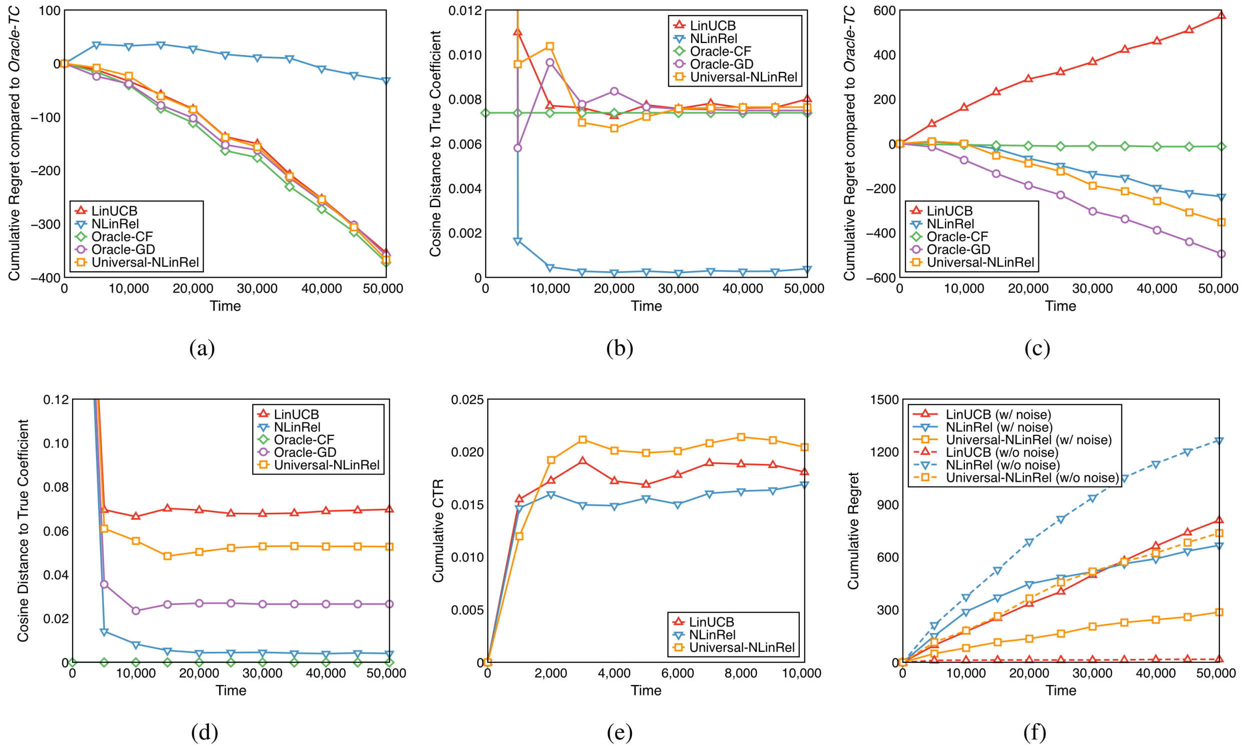

Figure 1: Comparisons of algorithms on synthetic (a)/(b)/(c)/(d) and real-world (e)/(f) datasets. (a)/(b) and (c)/(d) are measured under the choices of feature distributions as Gaussian(0,1) and 0.3 * Uniform(9,11) + 0.7 * Uniform(-11,-9), respectively. (a)/(c) report cumulative regrets of algorithms deducted by that of *Oracle-TC*. (b)/(d) report the cosine distances between coefficient vectors maintained by algorithms and the true one $\theta^\star$. (e) and (f) are for Yahoo and mushroom datasets, respectively.

comparisons among *Universal-NLinRel*, *LinUCB*, *Oracle-GD*, *Oracle-TC* and *Oracle-CF* are reported in Figure 1. In the case of the Gaussian distribution, as reported in Figure 1 (a), one can observe that both *LinUCB* and *Univeral-NLinRel* are close to the optimal *Oracle-CF* in this setting. The near-optimality of *LinUCB* can be explained as its similarity to the simple greedy algorithm in Section 4.1. In the case of the mixture of uniform distribution, as reported in Figure 1 (c), one can observe that *LinUCB* has the worst regret and is significantly outperformed by *Univeral-NLinRel*. Figure 1 (b) and (d) show that *NLinRel* finds the true coefficient vector $\theta^\star$ well in both Gaussian and non-Gaussian setups. This explains why *Univeral-NLinRel* can perform well (since *Univeral-NLinRel* uses *NLinRel* as its subroutine for tracking the true parameter).

Yahoo dataset. We use Yahoo Webscope R6A dataset Li et al. (2010), which contains the history of Yahoo! Front Page Module. The "Featured" tab of the Module highlights one article from the human edited candidate set of size 20. The log contains user context, arm context, candidate set, chosen arm, and reward (click or not). We consider an article as an arm. As a pre-processing step, we removed the lines which are incomplete (contains an arm whose context is not recorded). Then, we clustered the lines by the user. Then, each user can observe several candidate arms whose rewards can be calculated as their empirical CTRs (Click-Through Rates). For example, if user u observed an arm a for N times and clicked it M times, we assumed the reward of the context vector $x_{u,a}$ is M/N. We only consider users whose candidates/arms are of size larger than 2. The number of users after this pruning processing is 11,352, and MAB algorithm iterated 10,000 of them without duplication. Both user and article are represented by a six-dimensional real vector. We used the inner product of two features $x_{u,a} \in \mathbb{R}^{36}$ as a context of each arm as in Li et al. (2010).

We remark that our reported CTRs are different from those in Li et al. (2010); the authors uses different parameter θ for each article, but we instead use a single universal θ under which our algorithms and their theoretical reasoning have been developed.

We run *LinUCB*, *NLinRel* and *Universal-NLinRel* on the pre-processed Yahoo dataset. Compared to our synthetic setting, computing gradients in *Universal-NLinRel* becomes more expensive due to the larger number of candidate arms. Hence, we estimate each integral in gradients by Monte Carlo of 100 samples. In addition, since we do not have the knowledge of noise and feature distributions, we use the current context as a random sample z in *Universal-NLinRel*, and set the noise variance as 10% of the sample variance of contexts in the entire dataset. Under the Yahoo data, *Universal-NLinRel*, *LinUCB* and *NLinRel* perform better in their orders, as reported in Figure 1 (e).

Mushroom dataset. We use mushroom dataset Bache & Lichman (2013) which was used in the contextual MAB experiment in Blundell et al. (2015). Each mushroom has 22 categorical features and labeled as edible or poisonous. As in Blundell et al. (2015), we used 126 dimensional binary vectors as features. At each round, we sample one edible mushroom and 4 poisonous ones. Thus, the learner searches one edible mushroom from 5 candidate mushrooms. If the agent chooses an edible mushroom, the regret does not change, and if the agent chooses a poisonous one, the regret increases by 1. We experimented two different settings. The first one uses raw data and assumes the noise of context is 10% of the sample variance (as we do for Yahoo dataset). The second one added artificial noise to features, similar to the synthetic experiment. For each dimension, we added Gaussian noise with mean 0 and variance $i/126$, where $i \in \{1, ..., 126\}$. The results are reported in Figure 1 (f). In the first experiment without noise, we observe that *LinUCB* performs quite well, almost zero regret, since this data is almost linearly separable, i.e., the best setting for *LinUCB*. However, in the second experiment with artificial noise, *Universal-NLinRel* definitely outperforms *LinUCB*. This experiment shows that in some scenarios, it is important to learn/know the statistical information on noise for the performance of *Universal-NLinRel*. We leave this for further exploration in the future.

6 Conclusion

In this paper, we study contextual multi-armed bandit problems under assuming linear payoffs and uncertainty on features. Based on our theoretical understandings on the special cases of identical noise and Gaussian features, we could develop *Universal-NLinRel* for general scenarios. We believe that utilizing model uncertainties as addressed in this paper would provide an important direction for designing more practical algorithms for the bandit task.

References

Abe, Naoki and Long, Philip M. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pp. 3–11, 1999.

Agrawal, Shipra and Goyal, Navin. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pp. 127–135, 2013.

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

Auer, Peter, Cesa-Bianchi, Nicolo, Freund, Yoav, and Schapire, Robert E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Awerbuch, Baruch and Kleinberg, Robert D. 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*, pp. 45–53. ACM, 2004.

Bache, Kevin and Lichman, Moshe. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Beygelzimer, Alina, Langford, John, Li, Lihong, Reyzin, Lev, and Schapire, Robert E. Contextual bandit algorithms with supervised learning guarantees. In *AISTATS*, pp. 19–26, 2011.

Blundell, Charles, Cornebise, Julien, Kavukcuoglu, Koray, and Wierstra, Daan. Weight uncertainty in neural network. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 1613–1622, 2015.

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

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

Dani, Varsha, Hayes, Thomas P, and Kakade, Sham M. Stochastic linear optimization under bandit feedback. In *COLT*, pp. 355–366, 2008.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1050–1059, 2016.

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

Jung, Tobias, Martin, Sylvain, Ernst, Damien, and Leduc, Guy. Contextual multi-armed bandits for web server defense. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8. IEEE, 2012.

Kaelbling, Leslie Pack. Associative reinforcement learning: Functions ink-dnf. *Machine Learning*, 15(3): 279–298, 1994.

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

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

Paulsen, Vern. *Completely bounded maps and operator algebras*, volume 78. Cambridge University Press, 2002.

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

Sarkar, Jyotirmoy. One-armed bandit problems with covariates. *The Annals of Statistics*, pp. 1978–2002, 1991.

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

Strehl, Alexander L, Mesterharm, Chris, Littman, Michael L, and Hirsh, Haym. Experience-efficient learning in associative bandit problems. In *ICML*, pp. 889–896, 2006.

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

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

Woodroofe, Michael. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

A Tail Bounds of Sums of Random Matrices

In the proof of Theorem 1, we have matrix martingales and require to find their spectral norms to complete our proofs. When a matrix martingale is a sum of random matrices having bounded spectral norms, we can use matrix Azuma inequality which is Theorem 7.1 of Tropp (2012).

Theorem 4 (Matrix Azuma) *Let $\{X(t)\}_{1\le t\le T}$ be a finite sequence of self-adjoint matrices in dimension d that satisfy*

$$\mathbb{E}[X(t+1)|\{X(i)\}_{1\le i\le t}] = \mathbf{0}_{d\times d} \quad \textit{and} \quad \left\|X(t)^2\right\|_2 \le B.$$

Then, for all T,

$$\mathbb{P}\left\{\left\|\sum_{t=1}^{T} X(t)\right\|_2 \ge x\right\} \le 2de^{-x^2/8BT}.$$

For the proof of Theorem 1, we should study matrix martingales $\{N_1(t)\}_{1\le t\le T}$, $\{N_2(t)\}_{1\le t\le T}$, and $\{N_3(t)\}_{1\le t\le T}$, which are defined as follows:

$$\begin{aligned} N_1(t) &= \sum_{\tau=1}^{t} z_{a(\tau)}(\tau)\varepsilon(\tau)^\top, \\ N_2(t) &= \sum_{\tau=1}^{t} \varepsilon(\tau)\varepsilon(\tau)^\top - \mathcal{E}(\tau), \quad \text{and} \quad N_3(t) = \sum_{\tau=1}^{t} x_{a(\tau)}(\tau)(y(\tau) - z_{a(\tau)}(\tau)^\top\theta^\star). \end{aligned} \tag{5}$$

We cannot directly apply matrix Azuma inequality to bound the spectral norms of $\{N_1(t)\}_{1\le t\le T}$ and $\{N_3(t)\}_{1\le t\le T}$, since they are not self-adjoint. To resolve this problem, we introduce an operator $\mathscr{S}$, called *dilations* by Paulsen (2002), so that

$$\mathscr{S}(A) := \begin{bmatrix} \mathbf{0}_{m\times m} & A \\ A^\top & \mathbf{0}_{n\times n} \end{bmatrix} \qquad \forall A \in \mathbb{R}^{m\times n}. \tag{6}$$

It is known that *dilations* preserves the spectral norm, i.e.

$$\|\mathscr{S}(A)\|_2 = \|A\|_2. \tag{7}$$

Let $\delta(t) = y(t) - z_{a(t)}(t)^\top \theta^\star$ and $\delta_{\max} = \max_{1\le t\le T} \delta(t)$. Let $\varepsilon_{\max} = \max_{1\le t\le T} \|\varepsilon(t)\|_2$ and $\mathcal{E}_{\max} = \max_{1\le t\le T} \|\mathcal{E}(t)\|_2$. Using matrix Azuma inequality and *dilations* operator, we can bound the spectral norm of N_1, N_2, and N_2 as follows:

1. ($\|N_1(t)\|_2$) Let $\mathcal{N}(t) = \mathscr{S}(N_1(t))$. Then, $\{\mathcal{N}(\tau) - \mathcal{N}(\tau-1)\}_{1\le\tau\le t}$ is a sequence of self-adjoint matrices in dimension $2d$ that satisfy

$$\mathbb{E}\left[\mathcal{N}(\tau) - \mathcal{N}(\tau-1) | \{\mathcal{N}(i)\}_{1\le i\le\tau-1}\right] = 0 \qquad \text{and}$$

$$\begin{aligned}\left\|(\mathcal{N}(\tau) - \mathcal{N}(\tau-1))^2\right\|_2 &= \left\|\begin{bmatrix} z_{a(\tau)}(\tau)\varepsilon(\tau)^\top\varepsilon(\tau)z_{a(\tau)}(\tau)^\top & 0 \\ 0 & \varepsilon(\tau)z_{a(\tau)}(\tau)^\top z_{a(\tau)}(\tau)\varepsilon(\tau)^\top\end{bmatrix}\right\|_2 \\ &\le \|\varepsilon(\tau)\|_2^2 \left\|z_{a(\tau)}(\tau)\right\|_2^2 \le \varepsilon_{\max}^2.\end{aligned}$$

From matrix Azuma inequality,

$$\mathbb{P}\left\{\|\mathcal{N}(t)\|_2 \ge \sqrt{8\varepsilon_{\max}^2 t \log\frac{4d}{\delta_1}}\right\} \le \delta_1. \tag{8}$$

2. ($\|N_2(t)\|_2$) Let $\overline{\mathcal{E}}(\tau) = \epsilon(\tau)\epsilon(\tau)^\top - \mathcal{E}(\tau)$. Then, $N_2(t) = \sum_{\tau=1}^t \overline{\mathcal{E}}(\tau)$ and $\left\{\overline{\mathcal{E}}(\tau)\right\}_{1\le\tau\le t}$ is a sequence of self-adjoint matrices in dimension d that satisfy

$$\mathbb{E}\left[\overline{\mathcal{E}}(\tau) | \{\overline{\mathcal{E}}(i)\}_{1\le i\le\tau-1}\right] = 0 \qquad \text{and}$$

$$\begin{aligned}\left\|\overline{\mathcal{E}}(\tau)^2\right\|_2 &= \left\|\varepsilon(\tau)\varepsilon(\tau)^\top\varepsilon(\tau)\varepsilon(\tau)^\top - \mathcal{E}(\tau)\varepsilon(\tau)\varepsilon(\tau)^\top - \varepsilon(\tau)\varepsilon(\tau)^\top\mathcal{E}(\tau) + \mathcal{E}(\tau)^2\right\|_2 \\ &\le \|\varepsilon(\tau)\|_2^4 + 2\|\mathcal{E}(\tau)\|_2\|\varepsilon(\tau)\|_2^2 + \|\mathcal{E}(\tau)\|_2^2 \le (\varepsilon_{\max}^2 + \mathcal{E}_{\max})^2.\end{aligned}$$

From matrix Azuma inequality,

$$\mathbb{P}\left\{\|N_2(t)\|_2 \ge \sqrt{8(\varepsilon_{\max}^2 + \mathcal{E}_{\max})^2 t \log\frac{2d}{\delta_2}}\right\} \le \delta_2 \tag{9}$$

3. ($\|N_3(t)\|_2$) Let $\mathcal{N}(t) = \mathscr{S}(N_3(t))$. Then, $\{\mathcal{N}(\tau) - \mathcal{N}(\tau-1)\}_{1\le\tau\le t}$ is a sequence of self-adjoint matrices in dimension $d+1$ that satisfy

$$\mathbb{E}\left[\mathcal{N}(\tau) - \mathcal{N}(\tau-1) | \{\mathcal{N}(i)\}_{1\le i\le\tau-1}\right] = 0 \qquad \text{and}$$

$$\begin{aligned}\left\|(\mathcal{N}(\tau) - \mathcal{N}(\tau-1))^2\right\|_2 &= \begin{bmatrix} \delta(\tau)^2 x_{a(\tau)}(\tau)x_{a(\tau)}(\tau)^\top & 0 \\ 0 & \delta(\tau)^2 x_{a(\tau)}(\tau)^\top x_{a(\tau)}(\tau)\end{bmatrix} \\ &\le \delta_{\max}^2(\varepsilon_{\max} + z_{\max})^2.\end{aligned}$$

From matrix Azuma inequality,

$$\mathbb{P}\left\{\|N_3(t)\|_2 \ge \sqrt{8\delta_{\max}^2(\varepsilon_{\max}+1)^2 t \log\frac{2(d+1)}{\delta_3}}\right\} \le \delta_3 \tag{10}$$

From the union bound and equations (8), (9), and (10) with $\delta_1 = \delta_2 = \delta_3 = \frac{\delta}{6T}$, we can make Claim 5. In Claim 5, c_1 is a function of $\delta_{\max}$, $\varepsilon_{\max}$, and $\mathcal{E}_{\max}$, which are finite constants from the model.

Claim 5 *There exists a constant $c_1 > 0$ such that with probability $1 - \frac{\delta}{2}$,*

$$\|N_1(t)\|_2 + \|N_2(t)\|_2 + \|N_3(t)\|_2 \le c_1 \sqrt{t \log\left(\frac{dT}{\delta}\right)} \quad \textit{for all} \quad 1 \le t \le T.$$

In the proof of Theorem 3, we have to find spectral norms of

$$X - \mathbb{E}[X] = \sum_{t=1}^{\tau} \left(x_{a(t)} x_{a(t)}^\top - \mathbb{E}[x_{a(t)} x_{a(t)}^\top] \right)$$
$$Y - \mathbb{E}[Y] = \sum_{t=1}^{\tau} \left(y_{a(t)} y_{a(t)}^\top - \mathbb{E}[y_{a(t)} y_{a(t)}^\top] \right).$$

Both matrices are sums of i.i.d random matrices with zero mean and thus can be considered as matrix martingales. However, we cannot use matrix Azuma inequality since the spectral norm of each random matrix cannot be bounded. Indeed, $x_{a(t)}$ and $y_{a(t)}$ follow multivariate normal distributions and thus, $\|x_{a(t)}\|_2^2$ and $\|y_{a(t)}\|_2^2$ follow subexponential distributions. Then, matrix Bernstein can be applied to obtain the spectral norm of the sum. For readers' convenience, we write matrix Bernstein (Theorem 6.2 of Tropp (2012)) in the bellow.

Theorem 6 (Matrix Bernstein: Subexponential Case) *Let $X_1, \ldots, X_T$ be a finite sequence of independent, random, self-adjoint matrices with dimension d, satisfying that*

$$\mathbb{E}[X_k] = 0 \quad \textit{and} \quad \mathbb{E}[X_k^p] \preceq \frac{p!}{2} R^{p-2} A_k^2 \ \textit{for} p = 2, 3, \ldots, T.$$

and the variance parameter

$$\sigma_2 := \left\| \sum_{k=1}^{T} A_k^2 \right\|_2 \le S$$

Then the following chain of inequalities holds for all $x \ge 0$,

$$\mathbb{P}\left\{ \left\| \sum_{k=1}^{T} X_k \right\|_2 \ge x \right\} \le 2d \exp\left(-\frac{x^2/2}{S + Rx} \right).$$

B Proof of Theorem 1

In this proof, we use $C(t)$ and k_t to denote the C and k at time t of *NLinRel*. Then,

$$w_{i,j}(t) = \left\| U_{k_t+1:d}^\top (z_i(t) - z_j(t)) \right\|_2.$$

We use Claim 5 in many plcase. Recall that with probability $1 - \frac{\delta}{2}$,

$$\|N_1(t)\|_2 + \|N_2(t)\|_2 + \|N_3(t)\|_2 \le c_1 \sqrt{t \log\left(\frac{dT}{\delta}\right)} \quad \text{for all} \quad 1 \le t \le T, \tag{11}$$

where N_1, N_2, and N_3 are defined in Section A.

Estimation Error. From the estimated expected rewards $r_1(t), \dots, r_K(t)$, we can estimate the difference between the hidden expected rewards of any two actions. Lemma 7 states the error of the estimation, which can be bounded by $w_{i,j}(t)$ and a time decaying term $t^{\frac{1}{2}-\alpha}$.

Lemma 7 *For all time $t \in \{1, 2, \dots, T\}$ and for all $i, j \in \{1, 2, \dots, K\}$,*

$$\left|(r_i(t) - r_j(t)) - (z_i(t) - z_j(t))^\top \theta^\star\right| \leq t^{-\alpha} \left(\|N_1(t)\|_2 + \|N_2(t)\|_2 + \|N_3(t)\|_2\right) + w_{i,j}(t).$$

From (11) and Lemma 7, we have that for all $t \in \{1, \dots, T\}$ and all $i \in \{1, \dots, K\}$,

$$\left|(r_{a^\star(t)}(t) - r_i(t)) - (z_{a^\star(t)}(t) - z_i(t))^\top \theta^\star\right| \leq c_1 t^{\frac{1}{2}-\alpha} \sqrt{\log\left(\frac{dT}{\delta}\right)} + w_{a^\star(t),i}(t).$$

From the above inequality, the regret can be bounded as follows:

$$\begin{aligned} R(T) &= \sum_{t=1}^{T} (z_{a^\star(t)} - z_{a(t)})^\top \theta^\star \\ &\leq \sum_{t=1}^{T} \left((r_{a^\star(t)}(t) - r_{a(t)}(t)) + w_{a^\star(t),a(t)}(t) + c_1 t^{\frac{1}{2}-\alpha} \sqrt{\log\left(\frac{dT}{\delta}\right)} \right). \end{aligned} \tag{12}$$

An alternative form of (12)**.** Still, (12) is hard to analyze. Here, we find an alternative representation of (12) that can be analyzed from results obtained in Auer (2002). To this aim, we use the following properties of $C(t)$:

1. When both i and j belong to $C(t)$,

$$|r_i(t) - r_j(t)| \leq w_{i,j}(t). \tag{13}$$

2. For all $i \notin C(t)$, there exists $\ell \in C(t)$ such that

$$r_\ell(t) - r_i(t) > w_{i,\ell}(t). \tag{14}$$

First, when $a(t)^\star \in C(t)$, we have

$$\begin{aligned} (r_{a^\star(t)}(t) - r_{a(t)}(t)) + w_{a^\star(t),a(t)}(t) &\overset{(a)}{\leq} 2 \left\| U_{k_t+1:d}^\top (z_{a^\star(t)}(t) - z_{a(t)}(t)) \right\|_2 \\ &\overset{(b)}{\leq} 2 \sum_{i \in C(t)} \left\| U_{k_t+1:d}^\top z_i(t) \right\|_2 \\ &\overset{(c)}{\leq} 2K \mathbb{E}\left[\left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2 \right], \end{aligned} \tag{15}$$

where (a) comes from (13), (b) is obtained from the triangle inequality, and (c) holds because

$$\mathbb{E}\left[\left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2 \right] = \frac{1}{|C(t)|} \sum_{i \in C(t)} \left\| U_{k_t+1:d}^\top z_i(t) \right\|_2$$

since $a(t)$ is selected uniformly at random from $C(t)$.

Second, when $a^\star(t) \notin C(t)$, from (14), there exists $\ell \in C(t)$ such that

$$r_\ell(t) - r_{a^\star(t)}(t) \geq w_{a^\star(t),\ell} = \left\| U_{k_t+1:d}^\top (z_{a^\star(t)}(t) - z_\ell(t)) \right\|_2 .$$

Therefore,

$$\begin{aligned} &(r_{a^\star(t)}(t) - r_{a(t)}(t)) + w_{a^\star(t),a(t)}(t) \\ \leq &(r_{a^\star(t)}(t) - r_{a(t)}(t)) + \left(\left\| U_{k_t+1:d}^\top (z_{a^\star(t)}(t) - z_\ell(t)) \right\|_2 + \left\| U_{k+1:d}^\top (z_\ell(t) - z_j(t)) \right\|_2 \right) \\ \leq &(r_{a^\star(t)}(t) - r_{a(t)}(t)) + \left(r_\ell(t) - r_{a^\star(t)}(t) + \left\| U_{k_t+1:d}^\top (z_\ell(t) - z_j(t)) \right\|_2 \right) \\ \leq &r_\ell(t) - r_{a(t)}(t) + \left\| U_{k_t+1:d}^\top (z_\ell(t) - z_j(t)) \right\|_2 \\ \leq &2K \mathbb{E} \left[\left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2 \right], \end{aligned} \tag{16}$$

where the last inequality can be obtained analogously to (15).

Putting (15) and (16) onto (12), we have with probability $1 - \delta$,

$$R(T) \leq 2K \sum_{t=1}^{T} \mathbb{E} \left[\left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2 \right] + c_1 \sum_{t=1}^{T} t^{\frac{1}{2}-\alpha} \sqrt{\log\left(\frac{dT}{\delta}\right)}. \tag{17}$$

Sum of uncertainties. In this part, we study $\sum_{t=1}^{T} \mathbb{E}\left[\left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2 \right]$. We first consider $\sum_{t=1}^{T} \left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2$ and then, connect this to $\sum_{t=1}^{T} \mathbb{E}\left[\left\| U_{k_2+1:d}^\top z_{a(t)}(t) \right\|_2 \right]$ at the end of this part.

In Section 4 of Auer (2002), we can find an upper bound analysis for $\sum_{t=1}^{T} \|\tilde{v}_{a(t)}(t)\|_2$ where $\tilde{v}_{a(t)}(t)$ corresponds to $U_{k_t+1:d}^\top z_{a(t)}(t)$ in our analysis. However, we cannot directly apply their analysis to bound $\sum_{t=1}^{T} \left\| U_{k_t+1:d}^\top z_{a(t)}(t) \right\|_2$. In Section 4 of Auer (2002), $Z(T)$ is a sum of positive semi-definite matrices and their proofs rely on the properties of positive semi-definite matrices. In contrast, here, $Z(T) = \sum_{t=1}^{T} \left(x_{a(t)}(t) x_{a(t)}(t)^\top - \mathcal{E}(t) \right)$ is not a sum of positive semi-definite matrices due to the $\mathcal{E}(t)$ term.

To use the results in Auer (2002), we introduce

$$\overline{Z}(t) = \sum_{\tau=1}^{t} z_{a(\tau)}(\tau) z_{a(\tau)}(\tau)^\top$$

and let $\overline{U}(t)\overline{\Lambda}(t)\overline{U}(t)^\top$ be the singular value decomposition of $\overline{Z}(t)$. Then, $\overline{Z}(t) = Z(t) - N_1(t) - N_1(t)^\top - N_2(t)$. Let $\overline{\lambda}_k(t)$ be the k-th largest singular value of $\overline{Z}(t)$ and $\bar{k}_t$ be the constant such that $\overline{\lambda}_{\bar{k}_t}(t) \geq (t+1)^\beta$ and $\overline{\lambda}_{\bar{k}_t+1}(t) < (t+1)^\beta$ for a given $\beta > \alpha$. From Lemma 11 of Auer (2002),

$$\|\overline{U}_{\bar{k}_t+1:d}^\top z_{a(t)}(t)\|_2^2 \leq 4 \sum_{\lambda_j(t) \leq 5t^\beta} \lambda_j(t) - \lambda_j(t-1). \tag{18}$$

From the above inequality, we have

$$\begin{aligned} \sum_{t=1}^{T} \|\overline{U}_{\bar{k}_t+1:d}^\top z_{a(t)}(t)\|_2^2 \leq &\sum_{t=1}^{T} \left(4 \sum_{\lambda_j(t) \leq 5t^\beta} \lambda_j(t) - \lambda_j(t-1) \right) \\ \leq &20dT^\beta. \end{aligned} \tag{19}$$

Since $x_1 = \cdots = x_T = \sqrt{\frac{D}{T}}$ is the solution of $\max_{\boldsymbol{x}\in\mathbb{R}^T:\|\boldsymbol{x}\|_2^2\le D} \sum_{t=1} x_t$, we have from (19) that

$$\sum_{t=1}^T \|\overline{U}_{\bar{k}_t+1:d}^\top z_{a(t)}(t)\|_2 \le \sum_{t=1}^T \sqrt{\frac{20dT^\beta}{T}} = \sqrt{20d}T^{(\beta+1)/2}. \tag{20}$$

We now make an upper bound of $\sum_{t=1}^T \|U_{k_t+1:d}^\top z_{a(t)}(t)\|_2$ from $\sum_{t=1}^T \|\overline{U}_{\bar{k}_t+1:d}^\top z_{a(t)}(t)\|_2$. To this aim, we study the eigenvector perturbation of $\overline{Z(t)}$ by adding $N_1(t) + N_1(t)^\top + N_2(t)$. From matrix perturbation theories and the spectral norm of noise matrices $N_1(t)$, $N_2(t)$, and $N_3(t)$ studied in Section A, we obtain Lemma 8.

Lemma 8 *Let* $\eta(t) = t^{\alpha-\beta} + t^{-\beta}\left(2\left\|N_1(t)\right\|_2 + \left\|N_2(t)\right\|_2\right)$. *Then,*

$$\max\left\{0, \sqrt{1-\eta(t)^2}\left\|U_{k_t+1:d}^\top z_{a(t)}(t)\right\|_2 - \eta(t)\right\} \le \left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top z_{a(t)}(t)\right\|_2.$$

Since we have from (11) that $2\left\|N_1(t)\right\|_2 + \left\|N_2(t)\right\|_2 \le 2c_1\sqrt{t\log(\frac{dT}{\delta})}$,

$$\eta(t) \le t^{\alpha-\beta} + 2c_1 t^{\frac{1}{2}-\beta}\sqrt{\log(\frac{dT}{\delta})}. \tag{21}$$

Let $\overline{T}$ be a constant such that

$$\overline{T}^{\alpha-\beta} + 2c_1\overline{T}^{\frac{1}{2}-\beta}\sqrt{\log(\frac{dT}{\delta})} \le \sqrt{3}/2.$$

Then, from (20), (21), and Lemma 8,

$$\begin{aligned}
\sum_{t=1}^T \|U_{k_t+1:d}^\top z_{a(t)}(t)\|_2 &= \sum_{t=1}^{\overline{T}-1} \|U_{k_t+1:d}^\top z_{a(t)}(t)\|_2 + \sum_{t=\overline{T}}^T \|U_{k_t+1:d}^\top z_{a(t)}(t)\|_2 \\
&\le \sum_{t=1}^{\overline{T}-1} \|U_{k_t+1:d}^\top z_{a(t)}(t)\|_2 + \sum_{t=\overline{T}}^T \left(2\|\overline{U}_{\bar{k}_t+1:d}^\top z_{a(t)}(t)\|_2 + 2\eta(t)\right) \\
&\le \sum_{t=1}^{\overline{T}-1} \|U_{k_t+1:d}^\top z_{a(t)}(t)\|_2 + T^{\frac{\beta+1}{2}}\sqrt{80d} + 2\sum_{t=1}^T \eta(t) \\
&= O\left(T^{\frac{\beta+1}{2}}\sqrt{d} + T^{1+\alpha-\beta}\right),
\end{aligned} \tag{22}$$

where the last equality comes from $\alpha > 1/2$.

From Azuma's inequality,

$$\mathbb{P}\left(\sum_{t=1}^T \left(\mathbb{E}\left[\left\|U_{k_t+1:d}^\top z_{a(t)}(t)\right\|_2\right] - \left\|U_{k_t+1:d}^\top z_{a(t)}(t)\right\|_2\right) \ge \sqrt{4T\log(2/\delta)}\right) \le \frac{\delta}{2}. \tag{23}$$

Therefore, (22) and (23) conclude that with probability $1 - \frac{\delta}{2}$,

$$\sum_{t=1}^T \mathbb{E}\left[\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2\right] = O\left(T^{\frac{\beta+1}{2}}\sqrt{d} + T^{1+\alpha-\beta}\right). \tag{24}$$

Regret Analysis. From (17) and (24), when $\beta = \frac{1+2\alpha}{3}$,

$$R(T) = O\left(T^{\frac{3}{2}-\alpha}\sqrt{d\log\left(dT/\delta\right)} + KT^{(\alpha+2)/3}\sqrt{d}\right).$$

B.1 Proof of Lemma 7

Let $U\Sigma U^\top$ be the eigenvalue decomposition of $Z(t-1)$ and $k = k_{t-1}$. From the triangle inequality, we have

$$
\begin{aligned}
&\left|(r_i(t) - r_j(t)) - (z_i(t) - z_j(t))^\top \theta^\star\right| \\
=&\left|(r_i(t) - r_j(t)) - (z_i(t) - z_j(t))^\top (U_{1:k}U_{1:k}^\top + U_{k+1:d}U_{k+1:d}^\top)\theta^\star\right| \\
\le&\left|(r_i(t) - r_j(t)) - (z_i(t) - z_j(t))^\top U_{1:k}U_{1:k}^\top\theta^\star\right| + \left|(z_i(t) - z_j(t))^\top U_{k+1:d}U_{k+1:d}^\top\theta^\star\right| \\
\le&\left|(r_i(t) - r_j(t)) - (z_i(t) - z_j(t))^\top U_{1:k}U_{1:k}^\top\theta^\star\right| + w_{i,j}(t)\left\|U_{k+1:d}^\top\right\|_2 \|\theta^\star\|_2 \\
=&\left|(z_i(t) - z_j(t))^\top \left(\widehat{\theta} - U_{1:k}U_{1:k}^\top\theta^\star\right)\right| + w_{i,j}(t) \\
\le&\|z_i(t) - z_j(t)\|_2 \left\|\widehat{\theta} - U_{1:k}U_{1:k}^\top\theta^\star\right\|_2 + w_{i,j}(t) \\
=&\|z_i(t) - z_j(t)\|_2 \left\|U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top(Y(t-1) - Z(t-1)\theta^\star)\right\|_2 + w_{i,j}(t),
\end{aligned}
\tag{25}
$$

where the last equality is obtained from the fact that $U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top Z(t-1) = U_{1:k}U_{1:k}^\top$.

From (25), it suffices to show that

$$
\begin{aligned}
&\left\|U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top(Y(t-1) - Z(t-1)\theta^\star)\right\|_2 \\
\le& t^{-\alpha}\left(\|N_1(t-1)\|_2 + \|N_2(t-1)\|_2 + \|N_3(t-1)\|_2\right).
\end{aligned}
\tag{26}
$$

Since $\Delta(\Sigma, k) \ge t^\alpha$ from the definition of k,

$$
\begin{aligned}
&\left\|U_{1:k}\Delta(\Sigma, 1:k)^{-1}U_{1:k}^\top(Y(t-1) - Z(t-1)\theta^\star)\right\|_2 \\
\le&\|U_{1:k}\|_2 \left\|\Delta(\Sigma, 1:k)^{-1}\right\|_2 \left\|U_{1:k}^\top\right\| \|(Y(t-1) - Z(t-1)\theta^\star)\|_2 \\
\le& t^{-\alpha}\|(Y(t-1) - Z(t-1)\theta^\star)\|_2.
\end{aligned}
\tag{27}
$$

We can rewrite $Y(t-1) - Z(t-1)\theta^\star$ using $N_1(t)$, $N_2(t)$, and $N_3(t)$ as follows:

$$
\begin{aligned}
&Y(t-1) - Z(t-1)\theta^\star \\
=&\sum_{\tau=1}^{t-1} x_{a(\tau)}(\tau)y(\tau) - Z(t-1)\theta^\star \\
=&\sum_{\tau=1}^{t-1} x_{a(\tau)}(\tau)z_{a(\tau)}(\tau)^\top\theta^\star + \sum_{\tau=1}^{t-1} x_{a(\tau)}(\tau)(y(\tau) - z_{a(\tau)}(\tau)^\top\theta^\star) - Z(t-1)\theta^\star \\
=&\sum_{\tau=1}^{t-1}\left(-x_{a(\tau)}(\tau)\varepsilon(\tau)^\top + \mathcal{E}(\tau)\right)\theta^\star + \sum_{\tau=1}^{t-1} x_{a(\tau)}(\tau)(y(\tau) - z_{a(\tau)}(\tau)^\top\theta^\star) \\
=&\left(\sum_{\tau=1}^{t-1}\left(-\varepsilon(\tau)\varepsilon(\tau)^\top + \mathcal{E}(\tau)\right) - \sum_{\tau=1}^{t-1} z_{a(\tau)}(\tau)\varepsilon(\tau)^\top\right)\theta^\star + \sum_{\tau=1}^{t-1} x_{a(\tau)}(\tau)(y(\tau) - z_{a(\tau)}(\tau)^\top\theta^\star) \\
=&-\left(N_1(t-1) + N_2(t-1)\right)\theta^\star + N_3(t-1).
\end{aligned}
\tag{28}
$$

Therefore, from (28) and the triangle inequality, we have that

$$
\|Y(t-1) - Z(t-1)\theta^\star\|_2 \le \|N_1(t-1)\|_2 + \|N_2(t-1)\|_2 + \|N_3(t-1)\|_2. \tag{29}
$$

By putting (29) onto (27), we can obtain (26) and thus can conclude this proof.

B.2 Proof of Lemma 8

When $\eta(t) \geq 1$ or $\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 = 0$, we have trivial inequalities:

$$\left(\max\left\{0, \sqrt{1-\eta(t)^2}\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 - \eta(t)\right\}\right)^2 = 0 \quad \text{and}$$
$$\left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top z_{a(t)}(t)\right\|_2 \geq 0.$$

Thus, Lemma 8 holds when $\eta(t) \geq 1$ or $\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 = 0$.

We now assume that $\eta(t) < 1$ and $\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 > 0$. In this proof, we denote by $U\Sigma U^\top$ the eigenvalue decomposition of $Z(t-1)$ and we simply use $\eta = \eta(t)$ and $k = k_{t-1}$. Indeed, η is an upper bound of $\left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top U_{k+1:d}\right\|_2$. Since the columns of $\overline{U}_{1:\bar{k}_{t-1}}(t-1)$ are the top $\bar{k}_{t-1}$ eigenvectors of $\overline{Z}(t-1)$ and $\overline{\lambda}_{\bar{k}_{t-1}}(t-1) \geq t^\beta$, we have

$$\begin{aligned}\left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top U_{k+1:d}\right\|_2 &= \left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top \overline{Z}(t-1)^{-1}\overline{Z}(t-1)U_{k+1:d}\right\|_2 \\ &\leq t^{-\beta}\left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top \overline{Z}(t-1)U_{k+1:d}\right\|_2 \\ &= t^{-\beta}\left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top (Z(t) - N_1(t) - N_1(t)^\top - N_2(t))U_{k+1:d}\right\|_2 \\ &\leq t^{\alpha-\beta} + t^{-\beta}\left(2\left\|N_1(t)\right\|_2 + \left\|N_2(t)\right\|_2\right) = \eta < 1. \end{aligned} \tag{30}$$

To prove this lemma, we first define some useful unit vectors $h \in \mathbb{R}^{d-k}$ and $f \in \mathbb{R}^{d-\bar{k}_{t-1}}$ as follows:

$$h = \frac{1}{\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2} U_{k+1:d}^\top z_{a(t)}(t) \quad \text{and}$$
$$f = \frac{1}{\left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} h\right\|_2}\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} h.$$

Then, from (30), we have

$$\left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top U_{k+1:d} h\right\|_2 \leq \left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top U_{k+1:d}\right\|_2 = \eta \quad \text{and} \tag{31}$$
$$\left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} h\right\|_2 \geq \sqrt{1 - \left\|\overline{U}_{1:\bar{k}_{t-1}}(t-1)^\top U_{k+1:d}\right\|_2^2} = \sqrt{1-\eta^2} > 0. \tag{32}$$

From (32), we can make the following inequalities:

$$\left\|U_{k+1:d} h\right\|_2 \geq \left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} h\right\|_2 \geq \sqrt{1-\eta^2} \qquad \text{and thus}$$
$$\left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{1:k} h\right\|_2 \leq \left\|U_{1:k} h\right\|_2 \leq \sqrt{1 - \left\|U_{k+1:d} h\right\|_2^2} \leq \eta. \tag{33}$$

From (32) and definition of h, we have

$$\begin{aligned}\frac{\left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} U_{k+1:d}^\top z_{a(t)}(t)\right\|_2}{\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2} &= \left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} h\right\|_2 \\ &= \left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d} h\right\|_2 \\ &\geq \sqrt{1-\eta^2}. \end{aligned} \tag{34}$$

We can show Lemma 8 using f. Since $\|f\|_2 = 1$,

$$
\begin{aligned}
&\left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top z_{a(t)}(t)\right\|_2 \\
\geq &\left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top z_{a(t)}(t)\right\|_2 \\
= &\left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top (U_{1:k}U_{1:k}^\top + U_{k+1:d}U_{k+1:d}^\top) z_{a(t)}(t)\right\|_2 \\
\overset{(a)}{\geq} &\left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{k+1:d}U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 - \left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{1:k}U_{1:k}^\top z_{a(t)}(t)\right\|_2 \\
\overset{(b)}{\geq} &\sqrt{1-\eta^2}\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 - \left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{1:k}U_{1:k}^\top z_{a(t)}(t)\right\|_2 \\
\overset{(c)}{\geq} &\sqrt{1-\eta^2}\left\|U_{k+1:d}^\top z_{a(t)}(t)\right\|_2 - \eta\left\|z_{a(t)}(t)\right\|_2,
\end{aligned}
$$

where (a) is obtained from the triangle inequality, (b) stems from (34), and (c) holds because

$$
\begin{aligned}
&\left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{1:k}U_{1:k}^\top z_{a(t)}(t)\right\|_2 \\
\leq &\left\|f^\top \overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{1:k}h\right\|_2 \left\|U_{1:k}^\top z_{a(t)}(t)\right\|_2 \\
= &\left\|\overline{U}_{\bar{k}_{t-1}+1:d}(t-1)^\top U_{1:k}h\right\|_2 \left\|U_{1:k}^\top z_{a(t)}(t)\right\|_2 \\
\overset{(d)}{\leq} &\eta\left\|U_{1:k}^\top z_{a(t)}(t)\right\|_2 \leq \eta\left\|z_{a(t)}(t)\right\|_2,
\end{aligned}
$$

where (d) is obtained from (33).

C Proof of Theorem 2

Since contextual vectors $z_1(t), z_2(t), \ldots, z_K(t)$ and noise vectors $\varepsilon_1(t), \varepsilon_2(t), \ldots, \varepsilon_K(t)$ are independent each other,

$$
\mathbb{E}[y_i(t)|x_1(t), \ldots, x_K(t)] = \mathbb{E}[y_i(t)|x_i(t)]. \tag{35}
$$

Since $y_i(t)$ is a random variable with mean $z_i(t)^\top\theta^\star$ and independent to $\varepsilon_i(t)$, we have

$$
\begin{aligned}
\mathbb{E}[y_i(t)|x_i(t)] =&\mathbb{E}[z_i(t)^\top\theta^\star|x_i(t)] \\
=&\mathbb{E}[z_i(t)|x_i(t)]^\top\theta^\star.
\end{aligned} \tag{36}
$$

We can obtain $\mathbb{E}[z_i(t)|x_i(t)]$ using Bayes' theorem. Since $z_i(t) \sim \mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{feature}})$ and $(x_i(t) - z_i(t)) \sim \mathcal{N}(\mathbf{0}_{d\times 1}, \Sigma_{\texttt{noise}})$,

$$
\begin{aligned}
\mathbb{P}(z_i(t)|x_i(t)) \propto&\mathbb{P}(x_i(t)|z_i(t))\mathbb{P}(z_i(t)) \\
\propto&\exp\left(-\frac{1}{2}(x_i(t)-z_i(t))^\top\Sigma_{\texttt{noise}}^{-1}(x_i(t)-z_i(t))\right)\exp\left(\frac{1}{2}z_i(t)^\top\Sigma_{\texttt{feature}}^{-1}z_i(t)\right) \\
\propto&\exp\left(-\frac{1}{2}\left(z_i(t)-\tilde{x}_i(t)\right)^\top\left(\Sigma_{\texttt{feature}}^{-1}+\Sigma_{\texttt{noise}}^{-1}\right)\left(z_i(t)-\tilde{x}_i(t)\right)\right),
\end{aligned}
$$

where $\tilde{x}_i(t) = \left(\Sigma_{\texttt{feature}}^{-1}+\Sigma_{\texttt{noise}}^{-1}\right)^{-1}\Sigma_{\texttt{noise}}^{-1}x_i(t)$. Therefore, $\mathbb{P}(z_i(t)|x_i(t))$ is a Gaussian distribution with mean

$$
\mathbb{E}[z_i(t)|x_i(t)] = \left(\Sigma_{\texttt{feature}}^{-1}+\Sigma_{\texttt{noise}}^{-1}\right)^{-1}\Sigma_{\texttt{noise}}^{-1}x_i(t). \tag{37}
$$

From (35), (36), and (37),

$$\begin{aligned}\mathbb{E}[y_i(t)|x_1(t),\dots,x_K(t)] =&\mathbb{E}[z_i(t)|x_i(t)]^\top\theta^\star\\ =&\left(\left(\Sigma_{\texttt{feature}}^{-1}+\Sigma_{\texttt{noise}}^{-1}\right)^{-1}\Sigma_{\texttt{noise}}^{-1}x_i(t)\right)^\top\theta^\star\\ =&x_i(t)^\top\overline{\theta},\end{aligned}\tag{38}$$

where the last equality is obtained from Lemma 9.

Lemma 9 *When both Σ_1 and Σ_2 are full rank matrices,*

$$(\Sigma_1+\Sigma_2)^{-1}\Sigma_1=\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}.$$

C.1 Proof of Lemma 9

From the associative property of the matrix product, we have

$$\begin{aligned}&(\Sigma_1+\Sigma_2)^{-1}\Sigma_1\\ =&\left(\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\right)\left(\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\right)^{-1}(\Sigma_1+\Sigma_2)^{-1}\Sigma_1\\ =&\left(\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\right)\left(\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\Sigma_1\right)\\ =&\left(\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\right)\left(\left(\Sigma_1^{-1}\Sigma_2+\boldsymbol{I}_{d\times d}\right)(\Sigma_1+\Sigma_2)^{-1}\Sigma_1\right)\\ =&\left(\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\right)\left(\left(\Sigma_1^{-1}\Sigma_2+\Sigma_1^{-1}\Sigma_1\right)(\Sigma_1+\Sigma_2)^{-1}\Sigma_1\right)\\ =&\left(\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1}\right)\left(\Sigma_1^{-1}(\Sigma_1+\Sigma_2)\left(\Sigma_1+\Sigma_2\right)^{-1}\Sigma_1\right)\\ =&\Sigma_2^{-1}\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)^{-1},\end{aligned}$$

which is the end of this proof.

D Proof of Theorem 3

The simple greedy algorithm clearly separete the exploration part and the exloitation part. From Lemma 10, the exploration part can learn $\overline{\theta}$ and the output $\widehat{\theta}$ satisfies that

$$\|\overline{\theta}-\widehat{\theta}\|_2=O\left(T^{-\frac{1}{3}}\sqrt{\log\left(\frac{d}{\delta}\right)}\right)\quad\text{w.p.}\quad 1-\delta.\tag{39}$$

The proof of Lemma 10 is given in Section D.1.

Lemma 10 *Under the noisy Gaussian contextual bandit model, with probability $1-\delta$,*

$$\|\overline{\theta}-\widehat{\theta}\|_2=O\left(\sqrt{\frac{\log(d/\delta)}{\tau}}\right),$$

where $\widehat{\theta}$ is the output of the simple greedy algorithm after τ.

The regret of the simple greedy algorithm is bounded as follows:

$$\begin{aligned}\overline{R}(T) =& \sum_{t=1}^{T} \mathbb{E}\left[\left(x_{i^\star(t)}(t) - x_{a(t)}(t)\right)^\top \overline{\theta}\right] \\ =& \sum_{t=1}^{\tau} \mathbb{E}\left[\left(x_{i^\star(t)}(t) - x_{a(t)}(t)\right)^\top \overline{\theta}\right] + \sum_{t=\tau+1}^{T} \mathbb{E}\left[\left(x_{i^\star(t)}(t) - x_{a(t)}(t)\right)^\top \overline{\theta}\right] \\ =& \sum_{t=\tau+1}^{T} \mathbb{E}\left[\left(x_{i^\star(t)}(t) - x_{a(t)}(t)\right)^\top \overline{\theta}\right] + O\left(T^{\frac{2}{3}}\right). \end{aligned} \tag{40}$$

Since $a(t) = \arg\max_{1\leq i\leq K} x_i(t)^\top \widehat{\theta}$ for $t \geq \tau + 1$,

$$\begin{aligned} x_{i^\star(t)}(t)^\top \overline{\theta} - x_{a(t)}(t)^\top \overline{\theta} =& x_{i^\star(t)}(t)^\top \left(\overline{\theta} - \widehat{\theta}\right) + x_{i^\star(t)}(t)^\top \widehat{\theta} - x_{a(t)}(t)^\top \overline{\theta} \\ \leq& x_{i^\star(t)}(t)^\top \left(\overline{\theta} - \widehat{\theta}\right). \end{aligned} \tag{41}$$

From (39), (40), and (41), we have Theorem 3,

$$\overline{R}(T) = O\left(T^{\frac{2}{3}} \sqrt{\log\left(\frac{d}{\delta}\right)}\right) \quad \text{w.p.} \quad 1-\delta.$$

D.1 Proof of lemma 10

After τ, the expected matrices of X and Y are

$$\begin{aligned} \mathbb{E}[X] =& \mathbb{E}\left[\sum_{t=1}^{\tau} x_{a(t)}(t) x_{a(t)}(t)^\top\right] \\ =& \mathbb{E}\left[\sum_{t=1}^{\tau} (z_{a(t)}(t) + \varepsilon_{a(t)}(t))(z_{a(t)}(t) + \varepsilon_{a(t)}(t))^\top\right] \\ =& \mathbb{E}\left[\sum_{t=1}^{\tau} \left(z_{a(t)}(t) z_{a(t)}(t)^\top + z_{a(t)}(t)\varepsilon_{a(t)}(t)^\top + \varepsilon_{a(t)}(t) z_{a(t)}(t)^\top + \varepsilon_{a(t)}(t)\varepsilon_{a(t)}(t)^\top\right)\right] \\ =& \tau(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}}) \qquad \text{and} \end{aligned} \tag{42}$$

$$\begin{aligned} \mathbb{E}[Y] =& \mathbb{E}\left[\sum_{t=1}^{\tau} x_{a(t)}(t) y_{a(t)}(t)\right] \\ =& \mathbb{E}\left[\sum_{t=1}^{\tau} (z_{a(t)}(t) + \varepsilon_{a(t)}(t))(z_{a(t)}(t)^\top \theta^\star + y_{a(t)}(t) - z_{a(t)}(t)^\top \theta^\star)\right] \\ =& \tau \Sigma_{\texttt{feature}} \theta^\star. \end{aligned} \tag{43}$$

We can write $\widehat{\theta}$ using (42) and (43) as follows:

$$
\begin{aligned}
\widehat{\theta} =& X^{-1}Y \\
=& X^{-1}\mathbb{E}[X]\mathbb{E}[X]^{-1}Y \\
=& \mathbb{E}[X]^{-1}Y + X^{-1}(\mathbb{E}[X] - X)\mathbb{E}[X]^{-1}Y \\
=& \mathbb{E}[X]^{-1}\mathbb{E}[Y] + \mathbb{E}[X]^{-1}(Y - \mathbb{E}[Y]) + X^{-1}(\mathbb{E}[X] - X)\mathbb{E}[X]^{-1}Y \\
=& (\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}\Sigma_{\texttt{feature}}\theta^\star + \frac{1}{\tau}(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}(Y - \mathbb{E}[Y]) \\
&+ \frac{1}{\tau}X^{-1}(\mathbb{E}[X] - X)(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}Y \\
=& \overline{\theta} + \frac{1}{\tau}(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}(Y - \mathbb{E}[Y]) + \frac{1}{\tau}X^{-1}(\mathbb{E}[X] - X)(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}Y, \quad (44)
\end{aligned}
$$

where the last equality is obtained from Lemma 9. Therefore,

$$
\|\overline{\theta} - \widehat{\theta}\|_2 \le \|(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}\|_2 \frac{\|Y - \mathbb{E}[Y]\|_2 + \|X^{-1}\|_2\|X - \mathbb{E}[X]\|_2\|Y\|_2}{\tau}. \quad (45)
$$

From the matrix Bernstein inequality (Theorem 6), with probability $1 - \delta$,

$$
\|X - \mathbb{E}[X]\|_2 = O\left(\|\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}}\|_2\sqrt{\tau\log(d/\delta)}\right) \quad \text{and} \quad (46)
$$

$$
\|Y - \mathbb{E}[Y]\|_2 = O\left(\|\Sigma_{\texttt{noise}}\|_2\sqrt{\tau\log(d/\delta)}\right). \quad (47)
$$

From (46) and (47), we have

$$
\|Y\|_2 \le \|\mathbb{E}[Y]\|_2 + \|Y - \mathbb{E}[Y]\|_2 = \tau\|\Sigma_{\texttt{noise}}\|_2 + O\left(\|\Sigma_{\texttt{noise}}\|_2\sqrt{\tau\log(d/\delta)}\right) \quad (48)
$$

$$
\|X^{-1}\|_2 \le \frac{1}{\tau\lambda_d(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}}) - \|X - \mathbb{E}[X]\|_2} = O\left(\frac{1}{\tau}\|(\Sigma_{\texttt{feature}} + \Sigma_{\texttt{noise}})^{-1}\|_2\right) \quad (49)
$$

Putting (46), (47), (48), and (49) onto (45), we can conclude that with probability $1 - \delta$,

$$
\|\overline{\theta} - \widehat{\theta}\|_2 = O\left(\sqrt{\frac{\log(d/\delta)}{\tau}}\right).
$$