이학박사 학위논문

Efficient Linear Contextual Bandit Algorithms with Improved Regret Bounds

성능이 개선된 효율적인 선형 다중 슬롯 머신 알고리즘

2022년 2월

서울대학교 대학원

통계학과

김 원 영

Efficient Linear Contextual Bandit Algorithms with Improved Regret Bounds

지도교수 **Myunghee Cho Paik**

이 논문을 이학박사 학위논문으로 제출함
2021년 10월

서울대학교 대학원
통계학과
김 원 영

김원영의 이학박사 학위논문을 인준함
2021년 12월

위 원 장 _____ 오 희 석 _____ (인)

부위원장 _____ Myunghee Cho Paik _____ (인)

위 원 _____ 원 중 호 _____ (인)

위 원 _____ 오 민 환 _____ (인)

위 원 _____ 김 지 수 _____ (인)

Efficient Linear Contextual Bandit Algorithms with Improved Regret Bounds

By

Wonyoung Kim

A Thesis

Submitted in fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Statistics

Department of Statistics

College of Natural Sciences

Seoul National University

February, 2022

Abstract
Efficient Linear Contextual Bandit Algorithms with Improved Regret Bounds

Wonyoung Kim
The Department of Statistics
The Graduate School
Seoul National University

This thesis contains two proposed efficient algorithms: (i) Doubly Robust Thompson Sampling (`DRTS`) and (ii) Hybridization by Randomization (`HyRan`).

`DRTS` employs the doubly-robust method used in missing data literature to Thompson Sampling with contexts (`LinTS`). A challenging aspect of the bandit problem is that a stochastic reward is observed only for the chosen arm and the rewards of other arms remain missing. The dependence of the arm choice on the past context and reward pairs compounds the complexity of regret analysis. Different from previous works relying on missing data techniques [Dimakopoulou et al., 2019, Kim and Paik, 2019], the proposed algorithm is designed to allow a novel additive regret decomposition leading to an improved regret bound with the order of $\tilde{O}(\phi^{-2}\sqrt{T})$, where ϕ^2 is the minimum eigenvalue of the covariance matrix of contexts and T is the time horizon. This is the first regret bound of `LinTS` using ϕ^2 without the dimension of the context, d and the regret bound of the proposed algorithm is $\tilde{O}(d\sqrt{T})$ in many practical scenarios, improving the bound of `LinTS` by a factor of $\sqrt{d}$. A benefit of the proposed method is that it utilizes all the context data, chosen or not chosen, thus allowing to circumvent the technical definition of unsaturated arms used in theoretical analysis of `LinTS`. Empirical studies show the advantage of the proposed algorithm over `LinTS`.

`HyRan` is a novel bandit algorithm and show that our proposed algorithm establish the regret bound of $\tilde{O}(\sqrt{dT})$, which is optimal up to the logarithmic

factors. The novelty comes from the two modifications where the first is to utilize all contexts, both selected and unselected, and the second is to randomize the contribution to the estimator. These modifications render a novel decomposition of the cumulative regret into two main additive terms whose bounds can be derived by employing the structure of the compounding estimator. While previous algorithms such as `SupLinUCB` [Chu et al., 2011] have shown $\tilde{O}(\sqrt{dT})$ regret, exploiting independence via a phased algorithm, `HyRan` is the first to achieve $\tilde{O}(\sqrt{dT})$ regret keeping the practical advantage without resorting to generating independent samples. The numerical experiments show that the practical performance of our proposed algorithm is in line with the theoretical guarantees.

Keywords: Efficient linear contextual bandit algorithms, Improved regret bounds, Missing data, Randomization, Hybridization
Student Number : 2016-20263

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Doubly Robust Thompson Sampling with Linear Payoffs

1.1 Introduction

Contextual bandit has been popular in sequential decision tasks such as news article recommendation systems. In bandit problems, the learner sequentially pulls one arm among multiple arms and receives random rewards on each round of time. While not knowing the compensation mechanisms of rewards, the learner should make his/her decision to maximize the cumulative sum of rewards. In the course of gaining information about the compensation mechanisms through feedback, the learner should carefully balance between exploitation, pulling the best arm based on information accumulated so far, and exploration, pulling the arm that will assist in future choices, although it does not seem to be the best option at the moment. Therefore in the bandit problem, estimation or learning is an important element besides decision making.

A challenging aspect of estimation in the bandit problem is that a stochastic reward is observed only for the chosen arm. Consequently, only the context and reward pair of the chosen arm is used for estimation, which causes dependency

of the context data at the round on the past contexts and rewards. To handle this difficulty, I view bandit problems as missing data problems. The first step in handling missing data is to define full, observed, and missing data. In bandit settings, full data consist of rewards and contexts of all arms; observed data consist of full contexts for all arms and the reward for the chosen arm; missing data consist of the rewards for the arms that are not chosen. Typical estimation procedures require both rewards and contexts pairs to be observed, and the observed contexts from the unselected are discarded (see Table 1.1). The analysis based on the completely observed pairs only is called *complete record analysis*. Most stochastic bandit algorithms utilize estimates based on *complete record analysis*. Estimators from *complete record analysis* are known to be inefficient. In bandit setting, using the observed data whose probability of observation depends on previous rewards requires special theoretical treatment.

There are two main approaches to missing data: imputation and inverse probability weighting (IPW). Imputation is to fill in the predicted value of missing data from a specified model, and IPW is to use the observed records only but weight them by the inverse of the observation probability. The doubly robust (DR) method [Robins et al., 1994, Bang and Robins, 2005] is a combination of imputation and IPW tools. A review of missing data and DR methods is provided in Section 1.7.5. The robustness against model misspecification in missing data settings is insignificant in the bandit setting since the probability of observation or allocation to an arm is known. The merit of the DR method in the bandit setting is its ability to employ all the contexts including unselected arms.

We propose a novel multi-armed contextual bandit algorithm called Doubly Robust Thompson Sampling (`DRTS`) that applies the DR technique used in missing data literature to Thompson Sampling with linear contextual bandits (`LinTS`). The main thrust of `DRTS` is to utilize contexts information for all arms, not just chosen arms. By using the unselected, yet observed contexts,

Table 1.1: The shaded data are used in *complete record analysis* (left) and DR method (right) under multi-armed contextual bandit settings. The contexts, rewards and DR imputing values are denoted by X, Y, and Y^{DR}, respectively. The question mark refers to the missing reward of unchosen arms.

	$t=1$		$t=2$	
Arm 1	$X_{1,1}$	?	$X_{1,2}$	?
Arm 2	$X_{2,1}$	?	$X_{a_2}(2)$	$Y_{a_2}(2)$
Arm 3	$X_{a_1}(1)$	$Y_{a_1}(1)$	$X_{3,2}$	?
Arm 4	$X_{4,1}$	?	$X_{4,2}$	?

	$t=1$		$t=2$	
Arm 1	$X_{1,1}$	$Y_1^{DR}(1)$	$X_{1,2}$	$Y_1^{DR}(2)$
Arm 2	$X_{2,1}$	$Y_2^{DR}(1)$	$X_{a_2}(2)$	$Y_{a_2}^{DR}(2)$
Arm 3	$X_{a_1}(1)$	$Y_{a_1}^{DR}(1)$	$X_{3,2}$	$Y_3^{DR}(2)$
Arm 4	$X_{4,1}$	$Y_4^{DR}(1)$	$X_{4,2}$	$Y_4^{DR}(2)$

along with a novel algorithmic device, the proposed algorithm renders a unique regret decomposition which leads to a novel regret bound without resorting to the technical definition of unsaturated arms used by Agrawal and Goyal [2013]. Since categorizing the arms into saturated vs. unsaturated plays a critical role in costing extra $\sqrt{d}$, by circumventing it, we prove a $\tilde{O}(d\sqrt{T})$ bound of the cumulative regret in many practical occasions compared to $\tilde{O}(d^{3/2}\sqrt{T})$ shown in Agrawal and Goyal [2013].

The main contributions of this part of the thesis are as follows.

- We propose a novel contextual bandit algorithm that improves the cumulative regret bound of `LinTS` by a factor of $\sqrt{d}$ (Theorem 1.1) in many practical scenarios (Section 1.4.1). This improvement is attained mainly by defining a novel set called *super-unsaturated* arms, that is utilizable due to the proposed estimator and resampling technique adopted in the algorithm.

- We provide a novel estimation error bound of the proposed estimator (Theorem 1.3) which depends on the minimum eigenvalue of the covariance matrix of the contexts from all arms without d.

- We develop a novel dimension-free concentration inequality for sub-Gaussian

vector martingale (Lemma 1.4) and use it in deriving the regret bound in place of the self-normalized theorem by Abbasi-Yadkori et al. [2011].

- We develop a novel concentration inequality for the bounded matrix martingale (Lemma 1.6) which improves the existing result (Proposition 1.5) by removing the dependency on d in the bound. Lemma 1.6 also allows eliminating the forced sampling phases required in some bandit algorithms relying on Proposition 1.5 [Amani et al., 2019, Bastani and Bayati, 2020].

All missing proofs are in Section 1.7.3.

1.2 Related Works

Thompson Sampling [Thompson, 1933] has been extensively studied and shown solid performances in many applications (e.g. Chapelle and Li [2011]). Agrawal and Goyal [2013] is the first to prove theoretical bounds for `LinTS` and an alternative proof is given by Abeille et al. [2017]. Both papers show $\tilde{O}(d^{3/2}\sqrt{T})$ regret bound, which is known as the best regret bound for `LinTS`. Recently, Hamidi and Bayati [2020] points out that $\tilde{O}(d^{3/2}\sqrt{T})$ could be the best possible one can get when the estimator used by `LinTS` is employed. In this thesis, I improve this regret bound by a factor of $\sqrt{d}$ in many practical scenarios through a novel definition of super-unsaturated arms, which becomes utilizable due to the proposed estimator and resampling device implemented in the algorithm.

This work assumes the independence of the contexts from all arms across time rounds. Some notable works have used the assumption that the contexts are independently identically distributed (IID). Leveraging the IID assumption with a margin condition, Goldenshluger and Zeevi [2013] derives a two-armed linear contextual bandit algorithm with a regret upper bound of order $O(d^3 \log T)$. Bastani and Bayati [2020] has extended this algorithm to any num-

ber of arms and improves the regret bound to $O(d^2 \log^{\frac{3}{2}} d \cdot \log T)$. The margin condition states that the gap between the expected rewards of the optimal arm and the next best arm is nonzero with some constant probability. This condition is crucial in achieving a $O(\log T)$ regret bound instead of $\tilde{O}(\sqrt{T})$. In this thesis, we do not assume this margin condition, and focus on the dependence on the dimension of contexts d.

From a missing data point of view, most stochastic contextual bandit algorithms use the estimator from *complete record analysis* except Dimakopoulou et al. [2019] and Kim and Paik [2019]. Dimakopoulou et al. [2019] employs an IPW estimator that is based on the selected contexts alone. Dimakopoulou et al. [2019] proves a $\tilde{O}(d\sqrt{\epsilon^{-1}T^{1+\epsilon}N})$ regret bound for their algorithm which depends on the number of arms, N. Kim and Paik [2019] considers the high-dimensional settings with sparsity, utilizes a DR technique, and improves the regret bound in terms of the sparse dimension instead of the actual dimension of the context, d. Kim and Paik [2019] is different from the proposed algorithm in several aspects: the mode of exploration (ϵ-greedy vs. Thompson Sampling), the mode of regularization (Lasso vs. ridge regression); and the form of the estimator. A sharp distinction between the two estimators lies in that Kim and Paik [2019] aggregates contexts and rewards over the arms although they employ all the contexts. If we apply this aggregating estimator and DR-Lasso bandit algorithm to the low-dimensional setting, we obtain a regret bound of order $O(\frac{Nd}{\phi^2}\sqrt{T})$ when the contexts from the arms are independent. This bound is bigger than the novel bound by a factor of d and N. It is because the aggregated form of the estimator does not permit the novel regret decomposition derived in Section 1.4.2. The proposed estimator coupled with a novel algorithmic device renders the additive regret decomposition which in turn improves the order of the regret bound.

1.3 Proposed Estimator and Algorithm

1.3.1 Settings and Assumptions

We denote a d-dimensional context for the ith arm at round t by $X_{i,t} \in \mathbb{R}^d$, and the corresponding random reward by $Y_{i,t}$ for $i = 1, \ldots, N$. We assume $\mathbb{E}\left[Y_{i,t} | X_{i,t}\right] = X_{i,t}^T \beta$ for some unknown parameter $\beta \in \mathbb{R}^d$. At round t, the arm that the learner chooses is denoted by $a_t \in \{1, \ldots, N\}$, and the optimal arm by $a_t^* := \arg\max_{i=1,\ldots,N} \left\{X_{i,t}^T \beta\right\}$. Let $\texttt{regret}(t)$ be the difference between the expected reward of the chosen arm and the optimal arm at round t, i.e., $\texttt{regret}(t) := X_{a_t^*,t}^T \beta - X_{a_t,t}^T \beta$. The goal is to minimize the sum of regrets over T rounds, $R(T) := \sum_{t=1}^T \texttt{regret}(t)$. The total round T is finite but possibly unknown. We also make the following assumptions.

Assumption 1. Boundedness for scale-free regrets. For all $i = 1, \ldots, N$ and $t = 1, \ldots, T$, we have $\|X_{i,t}\|_2 \leq 1$ and $\|\beta\|_2 \leq 1$.

Assumption 2. Sub-Gaussian error. Let

$$\mathcal{H}_t := \bigcup_{\tau=1}^{t-1} \left[\{X_{i,\tau}\}_{i=1}^N \cup \{a_\tau\} \cup \{Y_{a_\tau,\tau}\}\right] \cup \{X_{i,t}\}_{i=1}^N,$$

be the set of observed data at round t. For each t and i, the error $\eta_{i,t} := Y_{i,t} - X_{i,t}^T \beta$ is conditionally zero-mean σ-sub-Gaussian for a fixed constant $\sigma \geq 0$, i.e, $\mathbb{E}\left[\eta_{i,t} | \mathcal{H}_t\right] = 0$ and $\mathbb{E}\left[\exp\left(\lambda \eta_{i,t}\right) | \mathcal{H}_t\right] \leq \exp(\lambda^2 \sigma^2 / 2)$, for all $\lambda \in \mathbb{R}$. Furthermore, the distribution of $\eta_{i,t}$ does not depend on the choice at round t, i.e. a_t.

Assumption 3. Independently distributed contexts. The stacked contexts vectors $\{X_{i,1}\}_{i=1}^N, \ldots, \{X_{i,T}\}_{i=1}^N \in \mathbb{R}^{dN}$ are independently distributed.

Assumption 4. Positive minimum eigenvalue of the average of covariance matrices. For each t, there exists a constant $\phi^2 > 0$ such that $\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N X_{i,t} X_{i,t}^T\right]\right) \geq \phi^2$.

Assumptions 1 and 2 are standard in stochastic bandit literature Agrawal

and Goyal [2013]. We point out that given round t, Assumption 3 allows that the contexts among different arms, $X_{1,t}, \ldots, X_{N,t}$ are correlated to each other. Assumption 3 is weaker than the assumption of IID, and the IID condition is considered by Goldenshluger and Zeevi [2013] and Bastani and Bayati [2020]. As Bastani and Bayati [2020] points out, the IID assumption is reasonable in some practical settings, including clinical trials, where health outcomes of patients are independent of those of other patients. Both Goldenshluger and Zeevi [2013] and Bastani and Bayati [2020] address the problem where the contexts are equal across all arms, i.e. $X(t) = X_{1,t} = \ldots = X_{N,t}$, while this thesis admits different contexts over all arms. Assumption 4 guarantees that the average of covariance matrices of contexts over the arms is well-behaved so that the inverse of the sample covariance matrix is bounded by the spectral norm. This assumption helps controlling the estimation error of β in linear regression models. Similar assumptions are adopted in existing works in the bandit setting [Goldenshluger and Zeevi, 2013, Li et al., 2017, Amani et al., 2019, Bastani and Bayati, 2020].

1.3.2 Doubly Robust Estimator

To describe the contextual bandit DR estimator, let $\pi_i(t) := \mathbb{P}\left(a_t = i | \mathcal{H}_t\right) > 0$ be the probability of selecting arm i at round t. We define a DR pseudo-reward as

$$Y_i^{DR}(t) = \left\{1 - \frac{\mathbb{I}\left(i = a_t\right)}{\pi_{i,t}}\right\} X_{i,t}^T \breve{\beta}_t + \frac{\mathbb{I}\left(i = a_t\right)}{\pi_{i,t}} Y_{a_t,t}, \tag{1.1}$$

for some $\breve{\beta}_t$ depending on $\mathcal{H}_t$. Background of missing data methods and derivation of the DR pseudo-reward is provided in Section 1.7.5. Now, we propose a new estimator $\widehat{\beta}_t$ with a regularization parameter λ_t as below:

$$\widehat{\beta}_t = \left(\sum_{\tau=1}^{t}\sum_{i=1}^{N} X_{i,\tau} X_{i,\tau}^T + \lambda_t I\right)^{-1} \left(\sum_{\tau=1}^{t}\sum_{i=1}^{N} X_{i,\tau} Y_i^{DR}(\tau)\right). \tag{1.2}$$

Harnessing the pseudo-rewards defined in (1.1), we can make use of all contexts rather than just selected contexts. The DR estimator by Kim and Paik [2019] utilizes all contexts but has a different form from (1.2). While Kim and Paik [2019] uses Lasso estimator with pseudo-rewards *aggregated* over all arms, we use ridge regression estimator with pseudo-rewards in (1.1) which are defined *separately* for each $i = 1, \ldots, N$. This seemingly small but important difference in forms paves a way in rendering the novel regret decomposition and improving the regret bound.

1.3.3 Algorithm

In this subsection, we describe the proposed algorithm, `DRTS` which adapts DR technique to `LinTS`. The `DRTS` is presented in Algorithm 1.1. Distinctive features of `DRTS` compared to `LinTS` include the novel estimator and the resampling technique. At each round $t \geq 1$, the algorithm samples $\tilde{\beta}_i(t)$ from the distribution $N(\widehat{\beta}_{t-1}, v^2 V_{t-1}^{-1})$ for each i independently. Let $\tilde{Y}_i(t) := X_{i,t}^T \tilde{\beta}_i(t)$ and $m_t := \arg\max_i \tilde{Y}_i(t)$. We set m_t as a candidate action and compute $\tilde{\pi}_{m_t}(t) := \mathbb{P}(\tilde{Y}_{m_t}(t) = \max_i \tilde{Y}_i(t) | \mathcal{H}_t)$. [1] If $\tilde{\pi}_{m_t}(t) > \gamma$, then the arm m_t is selected, i.e., $a_t = m_t$. Otherwise, the algorithm resamples $\tilde{\beta}_i(t)$ until it finds another arm satisfying $\tilde{\pi}_i(t) > \gamma$ up to a predetermined fixed value M_t. Section 1.7.1 describes issues related to M_t including a suitable choice of M_t.

The resampling step is incorporated to avoid small values of the probability of selection so that the pseudo-reward in (1.1) is numerically stable. A naive remedy to stabilize the pseudo-reward is to use $\max\{\pi_{i,t}, \gamma\}$, which fails to leading to $\tilde{O}(d\sqrt{T})$ regret bound since it induces bias and also cannot guarantee that the selected arm is in the super-unsaturated arms defined in (1.5) with high probability (For details, see Section 1.4.2). The resampling step

[1] This computation is known to be challenging but employing the independence among $\tilde{\beta}_1(t), \ldots, \tilde{\beta}_N(t)$, we derive an explicit form approximating $\tilde{\pi}_{m_t}(t)$ in supplementary materials Section 1.7.4

Algorithm 1.1 Doubly Robust Thompson Sampling for Linear Contextual Bandits (`DRTS`)

Input: Exploration parameter $v > 0$, Regularization parameter $\lambda > 0$, Selection probability threshold $\gamma \in [1/(N+1), 1/N)$, Imputation estimator $\breve{\beta}_u = f(\{X(\tau), Y_{a_\tau,\tau}\}_{\tau=1}^{u-1})$, Number of maximum possible resampling M_t.
Set $F_0 = 0$, $W_0 = 0$, $\widehat{\beta}_0 = 0$ and $V_0 = \lambda I$
for $t = 1$ **to** T **do**
  Observe contexts $\{X_{i,t}\}_{i=1}^N$.
  Sample $\tilde{\beta}_1(t), \ldots, \tilde{\beta}_N(t)$ from $N(\widehat{\beta}_{t-1}, v^2 V_{t-1}^{-1})$ independently. Compute $\tilde{Y}_i(t) = X_{i,t}^T \tilde{\beta}_i(t)$
  Observe a candidate action $m_t := \arg\max_i \tilde{Y}_i(t)$.
  Compute $\tilde{\pi}_{m_t}(t) := \mathbb{P}\left(\max_i \tilde{Y}_i(t) = \tilde{Y}_{m_t}(t) \middle| \mathcal{H}_t\right)$.
  for $l = 1$ **to** M_t **do**
    if $\tilde{\pi}_{m_t}(t) \leq \gamma$ **then**
      Sample another $\tilde{\beta}_1(t), \ldots, \tilde{\beta}_N(t)$, observe another m_t, and update $\tilde{\pi}_{m_t}(t)$.
    else
      Break.
    end if
  end for
  Set $a_t = m_t$, and play arm a_t.
  Observe reward $Y_{a_t,t}$ and compute $Y_i^{DR}(t)$
  $F_t = F_{t-1} + \sum_{i=1}^N X_{i,t} Y_i^{DR}(t)$; $W_t = W_{t-1} + \sum_{i=1}^N X_{i,t} X_{i,t}^T$; $V_t = W_t + \lambda\sqrt{t} I$
  $\widehat{\beta}_t = V_t^{-1} F_t$
  Update $\breve{\beta}_{t+1}$ for next round.
end for

implemented in the proposed algorithm is designed to solve these problems.

1.4 Theoretical Results

The theoretical results are organized as follows. In Section 1.4.1, we provide the main result, the cumulative regret bound of $\tilde{O}(\phi^{-2}\sqrt{T})$ of `DRTS`. The main thrust of deriving the regret bound is to define super-unsaturated arms. In Section 1.4.2 we introduce the definition of super-unsaturated arms and show how it admits a novel decomposition of the regret into two additive terms as in

(1.6). In Section 1.4.3 we bound each term of the decomposed regret bounds (1.6). The first term is the estimation error, and Theorem 1.3 finds its bound. In the course of proving Theorem 1.3, we need Lemma 1.4, which plays a similar role to the self-normalized theorem of Abbasi-Yadkori et al. [2011]. We conclude the section by presenting Lemma 1.6 and bound the second term of (1.6).

1.4.1 An Improved Regret Bound

Theorem 1.1 provides the regret bound of `DRTS` in terms of the minimum eigenvalue without d.

Theorem 1.1. *Suppose that Assumptions 1-4 hold. If $\breve{\beta}_t$ in Algorithm 1.1 satisfies $\|\breve{\beta}_t - \beta\|_2 \leq b$ for a constant $b > 0$, for all $t = 1, \ldots, T$, then with probability $1-2\delta$, the cumulative regret by time T for `DRTS` algorithm is bounded by*

$$R(T) \leq 2 + \frac{4C_{b,\sigma}}{\phi^2}\sqrt{T \log \frac{12T^2}{\delta}} + \frac{2\sqrt{2T}}{\phi\sqrt{N}}, \tag{1.3}$$

where $C_{b,\sigma}$ is a constant which depends only on b and σ.

The bound (1.3) has a rate of $O(\phi^{-2}\sqrt{T})$. The relationship between the dimension d and the minimum eigenvalue ϕ^2 can be shown by

$$d\phi^2 = \frac{d}{N}\lambda_{\min}\left(\mathbb{E}\sum_{i=1}^{N} X_{i,t}X_{i,t}^T\right) \leq \frac{1}{N}\mathbb{E}\sum_{i=1}^{N}\mathrm{Tr}\left(X_{i,t}X_{i,t}^T\right) = \frac{1}{N}\mathbb{E}\sum_{i=1}^{N}\|X_{i,t}\|_2^2 \leq 1.$$

This implies $\phi^{-2} \geq d$, [2] but there are many practical scenarios such that $\phi^{-2} = O(d)$ holds. Bastani et al. [2021] identifies such examples including the uniform distribution and truncated multivariate normal distributions. When the context has uniform distribution on the unit ball, $\phi^{-2} = d + 2$. When the

[2] Some previous works assume $\phi^{-2} = O(1)$ even when $\|X_{i,t}\|_2 \leq 1$ (e.g. Li et al. [2017]). As pointed out by Ding et al. [2021], this assumption is unrealistic and the reported regret bound should be multiplied by $O(d)$.

context has truncated multivariate normal distribution with mean 0 and covariance Σ, we can set $\phi^{-2} = (d+2)\exp(\frac{1}{2\lambda_{\min}(\Sigma)})$. For more examples, we refer to Bastani et al. [2021]. Furthermore, regardless of distributions, $\phi^{-2} = O(d)$ holds when the correlation structure has the row sum of off-diagonals independent of the dimension, for example, AR(1), tri-diagonal, block-diagonal matrices. In these scenarios, the regret bound in (1.3) becomes $\tilde{O}(d\sqrt{T})$. Compared to the previous bound of `LinTS` [Agrawal and Goyal, 2014, Abeille et al., 2017], we obtain a better regret bound by the factor of $\sqrt{d}$ for identified practical cases.

As for the imputation estimator $\check{\beta}_t$, we assume that $\|\check{\beta}_t - \beta\|_2 \leq b$, where b is an absolute constant. We suggest two cases which guarantee this assumption. First, if a biased estimator is used, we can rescale the estimator so that its l_2-norm is bounded by some constant $C > 0$. Then, $\|\check{\beta}_t - \beta\|_2 \leq \|\check{\beta}_t\|_2 + \|\beta\|_2 \leq C + 1$ and $b = C + 1$. Second, consistent estimators such as ridge estimator or the least squared estimator satisfy the condition since $\|\check{\beta}_t - \beta\|_2 = O(d\sqrt{\log t/t})$. The term d is canceled out when $t \geq t_d$, where t_d is the minimum integer that satisfies $\log t/t \leq d^{-2}$. In these two cases, we can find a constant b which satisfies the assumption on the imputation estimator $\check{\beta}_t$.

1.4.2 Super-unsaturated Arms and a Novel Regret Decomposition

The key element in deriving (1.3) is to decompose the regret into two additive terms as in (1.6). To allow such decomposition to be utilizable, we need to define a novel set of arms called super-unsaturated arms, which replaces the role of unsaturated arms in Agrawal and Goyal [2014]. The super-unsaturated arms are formulated so that the chosen arm is included in this set with high probability. For each i and t, let $\Delta_i(t) := X_{a_t^*,t}^T\beta - X_{i,t}^T\beta$. Define $A_t := \sum_{\tau=1}^t X_{a_\tau,\tau}X_{a_\tau,\tau}^T + \lambda I$ and $V_t := \sum_{\tau=1}^t \sum_{i=1}^N X_{i,\tau}X_{i,\tau}^T + \lambda_t I$. For the sake of

contrast, recall the definition of unsaturated arms by Agrawal and Goyal [2014] as

$$U_t := \left\{ i : \Delta_i(t) \leq g_t \left\| X_{i,t} \right\|_{A_{t-1}^{-1}} \right\}, \tag{1.4}$$

where $g_t := C\sqrt{d\log(t/\delta)}\min\{\sqrt{d}, \sqrt{\log N}\}$ for some constant $C > 0$. This g_t is constructed to ensure that there exists a positive lower bound for the probability that the selected arm is unsaturated. In place of (1.4), we define a set of super-unsaturated arms for each round t by

$$N_t := \left\{ i : \Delta_i(t) \leq 2 \left\| \widehat{\beta}_{t-1} - \beta \right\|_2 + \sqrt{\left\| X_{a_t^*,t} \right\|_{V_{t-1}^{-1}}^2 + \left\| X_{i,t} \right\|_{V_{t-1}^{-1}}^2} \right\}. \tag{1.5}$$

While $g_t \left\| X_{i,t} \right\|_{A_{t-1}^{-1}}$ in (1.4) is normalized with only selected contexts, the second term in the right hand side of (1.5) is normalized with all contexts including $X_{a_t^*,t}$, the contexts of the optimal arm. This bound of $\Delta_i(t)$ plays a crucial role in bounding the regret with a novel decomposition as in (1.6). The following Lemma shows a lower bound of the probability that the candidate arm is super-unsaturated.

Lemma 1.2. *For each t, let $m_t := \arg\max_i \tilde{Y}_i(t)$ and let N_t be the super-unsaturated arms defined in (1.5). For any given $\gamma \in [1/(N+1), 1/N)$, set $v = (2\log(N/(1-\gamma N)))^{-1/2}$. Then, $\mathbb{P}(m_t \in N_t | \mathcal{H}_t) \geq 1 - \gamma$.*

Lemma 1.2 directly contributes to the reduction of $\sqrt{d}$ in the hyperparameter v. In Agrawal and Goyal [2014], to prove a lower bound of $\mathbb{P}(a_t \in U_t | \mathcal{H}_t)$, it is required to set $v = \sqrt{9d\log(t/\delta)}$, with the order of $\sqrt{d}$. In contrast, Lemma 1.2 shows that v does not need to depend on d due to the definition of super-unsaturated arms in (1.5). In this way, we obtain a lower bound of $\mathbb{P}(m_t \in N_t | \mathcal{H}_t)$ without costing extra $\sqrt{d}$.

Using the lower bound, we can show that the resampling scheme allows the algorithm to choose the super-unsaturated arms with high probability. For all

$i \notin N_t$,

$$\tilde{\pi}_i(t) := \mathbb{P}\left(m_t = i \middle| \mathcal{H}_t\right) \leq \mathbb{P}\left(\cup_{j \notin N_t}\{m_t = j\} \middle| \mathcal{H}_t\right) = \mathbb{P}\left(m_t \notin N_t \middle| \mathcal{H}_t\right) \leq \gamma,$$

where the last inequality holds due to Lemma 1.2. Thus, in turn, if $\tilde{\pi}_i(t) > \gamma$, then $i \in N_t$. This means that $\{i : \tilde{\pi}_i(t) > \gamma\}$ is a subset of N_t and

$$\{a_t \in \{i : \tilde{\pi}_i(t) > \gamma\}\} \subset \{a_t \in N_t\}.$$

Hence, the probability of the event $\{a_t \in N_t\}$ is greater than the probability of sampling any arm which satisfies $\tilde{\pi}_i(t) > \gamma$. Therefore, with resampling, the event $\{a_t \in N_t\}$ occurs with high probability. (See Section 1.7.1 for details.)

When the algorithm chooses the arm from the super-unsaturated set, i.e., when $a_t \in N_t$ happens, (1.5) implies

$$\Delta_{a_t}(t) \leq 2\left\|\widehat{\beta}_{t-1} - \beta\right\|_2 + \sqrt{\left\|X_{a_t^*,t}\right\|^2_{V_{t-1}^{-1}} + \left\|X_{a_t,t}\right\|^2_{V_{t-1}^{-1}}}. \tag{1.6}$$

By definition, $\Delta_{a_t}(t) = \texttt{regret}(t)$ and the regret at round t can be expressed as the two additive terms, which presents a stark contrast with multiplicative decomposition of the regret in Agrawal and Goyal [2014]. In section 1.4.3 we show how each term can be bounded with separate rate.

1.4.3 Bounds for the Cumulative Regret

We first bound the leading term of (1.6) and introduce a novel estimation error bound free of d for the contextual bandit DR estimator.

Theorem 1.3. *(A dimension-free estimation error bound for the contextual bandit DR estimator.) Suppose Assumptions 1-4 hold. For each $t = 1, \ldots, T$, let $\breve{\beta}_t$ be any $\mathcal{H}_t$-measurable estimator satisfying $\|\breve{\beta}_t - \beta\|_2 \leq b$, for some constant $b > 0$. For each i and t, assume that $\pi_{i,t} > 0$ and that there ex-*

ists $\gamma \in [1/(N+1), 1/N)$ *such that* $\pi_{a_t,t} > \gamma$. *Given any* $\delta \in (0,1)$, *set* $\lambda_t = 4\sqrt{2}N\sqrt{t \log \frac{12\tau^2}{\delta}}$. *Then with probability at least* $1-\delta$, *the estimator* $\widehat{\beta}_t$ *in (1.2) satisfies*

$$\left\|\widehat{\beta}_t - \beta\right\|_2 \leq \frac{C_{b,\sigma}}{\phi^2\sqrt{t}}\sqrt{\log \frac{12t^2}{\delta}}, \tag{1.7}$$

for all $t = 1, \ldots, T$, *where the constant* $C_{b,\sigma}$ *which depends only on* b *and* σ.

In bandit literature, estimation error bounds typically include a term involving d which emerges from using the following two Lemmas: (i) the self-normalized bound for vector-valued martingales [Abbasi-Yadkori et al., 2011, Theorem 1] , and (ii) the concentration inequality for the covariance matrix [Tropp, 2015, Corollary 5.2]. Instead of using (i) and (ii), we develop the two dimension-free bounds in Lemmas 1.4 and 1.6, to replace (i) and (ii), respectively. With the two Lemmas, we eliminate the dependence on d and express the estimation error bound with ϕ^2 alone.

Lemma 1.4. *(A dimension-free bound for vector-valued martingales.) Let* $\{\mathcal{F}_\tau\}_{\tau=1}^t$ *be a filtration and* $\{\eta(\tau)\}_{\tau=1}^t$ *be a real-valued stochastic process such that* $\eta(\tau)$ *is* $\mathcal{F}_\tau$*-measurable. Let* $\{X(\tau)\}_{\tau=1}^t$ *be an* $\mathbb{R}^d$*-valued stochastic process where* $X(\tau)$ *is* $\mathcal{F}_{\tau-1}$*-measurable and* $\|X(\tau)\|_2 \leq 1$. *Assume that* $\{\eta(\tau)\}_{\tau=1}^t$ *are* σ*-sub-Gaussian as in Assumption 2. Then with probability at least* $1-\delta/t^2$, *there exists an absolute constant* $C > 0$ *such that*

$$\left\|\sum_{\tau=1}^t \eta(\tau)X(\tau)\right\|_2 \leq C\sigma\sqrt{t}\sqrt{\log \frac{4t^2}{\delta}}. \tag{1.8}$$

Compared to Theorem 1 of Abbasi-Yadkori et al. [2011], the bound (1.8) does not involve d, yielding a dimension-free bound for vector-valued martingales. However, the bound (1.8) has $\sqrt{t}$ term which comes from using $\|\cdot\|_2$ instead of the self-normalized norm $\|\cdot\|_{V_t^{-1}}$.

To complete the proof of Theorem 1.3, we need the following condition,

$$\lambda_{\min}(V_t) \geq ct, \tag{1.9}$$

for some constant $c > 0$. Li et al. [2017] points out that satisfying (1.9) is challenging. To overcome this difficulty, Amani et al. [2019] and Bastani and Bayati [2020] use an assumption on the covariance matrix of contexts and a concentration inequality for matrix to prove (1.9), described as follows.

Proposition 1.5. *Tropp [2015, Theorem 5.1.1] Let $P(1), \ldots, P(t) \in \mathbb{R}^{d\times d}$ be the symmetric matrices such that $\lambda_{\min}(P(\tau)) \geq 0$, $\lambda_{\max}(P(\tau)) \leq L$ and $\lambda_{\min}(\mathbb{E}[P(\tau)]) \geq \phi^2$, for all $\tau = 1, 2, \ldots, t$. Then,*

$$\mathbb{P}\left(\lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau)\right) \leq \frac{t\phi^2}{2}\right) \leq d\exp\left(-\frac{t\phi^2}{8L}\right). \tag{1.10}$$

To prove (1.9) using (1.10) with probability at least $1-\delta$, for $\delta \in (0,1)$, it requires $t \geq \frac{8L}{\phi^2}\log\frac{d}{\delta}$. Thus, one can use (1.10) only after $O(\phi^{-2}\log d)$ rounds. Due to this requirement, Bastani and Bayati [2020] implements the forced sampling techniques for $O\left(N^2 d^4 (\log d)^2\right)$ rounds, and Amani et al. [2019] forces to select arms randomly for $O\left(\phi^{-2}\log d\right)$ rounds. These mandatory exploration phase empirically prevents the algorithm choosing the optimal arm. An alternative form of matrix Chernoff inequality for adapted sequences is Theorem 3 in Tropp [2011], but the bound also has a multiplicative factor of d. Instead of applying Proposition 1.5 to prove (1.9), we utilize a novel dimension-free concentration inequality stated in the following Lemma.

Lemma 1.6. *(A dimension-free concentration bound for symmetric bounded matrices.) Let $\|A\|_F$ be a Frobenious norm of a matrix A. Let $\{P(\tau)\}_{\tau=1}^{t} \in \mathbb{R}^{d\times d}$ be the symmetric matrices adapted to a filtration $\{\mathcal{F}_\tau\}_{\tau=1}^{t}$. For each $\tau = 1, \ldots, t$, suppose that $\|P(\tau)\|_F \leq c$, for some $c > 0$ and $\lambda_{\min}\left(\mathbb{E}\left[P(\tau)|\,\mathcal{F}_{\tau-1}\right]\right) \geq \phi^2 > 0$, almost surely. For given any $\delta \in (0,1)$, set $\lambda_t \geq 4\sqrt{2}c\sqrt{t}\sqrt{\log\frac{4t^2}{\delta}}$.*

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

$$\lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau) + \lambda_t I\right) \geq \phi^2 t. \tag{1.11}$$

Lemma 1.6 shows that setting λ_t with $\sqrt{t}$ rate guarantees (1.9) for all $t \geq 1$. We incorporate λ_t stated in Lemma 1.6 in the estimator (1.2), and show in Section 1.5 that the DR estimator regularized with λ_t outperforms estimators from other contextual bandit algorithms in early rounds.

We obtain the bounds free of d in Lemmas 1.4 and 1.6 mainly by applying Lemma 2.3 in Lee et al. [2016] which states that any Hilbert space martingale can be reduced to $\mathbb{R}^2$. Thus, we can project the vector-valued (or the matrix) martingales to $\mathbb{R}^2$-martingales, and reduce the dimension from d (or d^2) to 2. Then we apply Azuma-Hoeffding inequality just twice, instead of d times. In this way, Lemma 1.6 provides a novel dimension-free bound for the covariance matrix.

Lemmas 1.4 and 1.6 can be applied to other works to improve the existing bounds. For example, using these Lemmas, the estimation error bound of Bastani and Bayati [2020] can be improved by a factor of $\log d$. Proposition EC.1 of Bastani and Bayati [2020] provides an estimation error bound for the ordinary least square estimator by using Proposition 1.5 and bounding all values of d coordinates. By applying Lemmas 1.4 and 1.6, one does not have to deal with each coordinate and eliminate dependence on d.

Using Lemma 1.6, we can bound the second term of the regret in (1.6) as follows. For $j = 1, \ldots, N$

$$\|X_{j,t}\|_{V_{t-1}^{-1}} \leq \|X_{j,t}\|_2 \sqrt{\left\|V_{t-1}^{-1}\right\|_2} \leq \lambda_{\min}\left(V_{t-1}\right)^{-1/2} \leq \frac{1}{\sqrt{\phi^2 N(t-1)}}. \tag{1.12}$$

Finally, we are ready to bound $\texttt{regret}(t)$ in (1.6).

Lemma 1.7. *Suppose the assumptions in Theorem 1.1 hold. Then with prob-*

ability at least $1 - 2\delta$*,*

$$\texttt{regret}(t) \leq \frac{2C_{b,\sigma}}{\phi^2\sqrt{t-1}}\sqrt{\log\frac{12t^2}{\delta}} + \frac{\sqrt{2}}{\phi\sqrt{N(t-1)}}, \tag{1.13}$$

for all $t = 2, \ldots, T$*.*

Proof. Since a_t is shown to be super-unsaturated with high probability, we can use (1.6) to have $\texttt{regret}(t) \leq 2\|\widehat{\beta}_{t-1} - \beta\|_2 + \sqrt{\|X_{a_t^*,t}\|^2_{V_{t-1}^{-1}} + \|X_{a_t,t}\|^2_{V_{t-1}^{-1}}}$, for all $t = 2, \ldots, T$. We see that the first term is bounded by Theorem 1.3, and the second term by (1.12). Note that to prove Theorem 1, Lemma 1.6 is invoked, and the event (1.11) of Lemma 1.6 is a subset of that in (1.7). Therefore (1.13) holds with probability at least $1 - 2\delta$ instead of $1 - 3\delta$. Details are given in Section 1.7.3. □

Lemma 1.7 shows that the regret at round t does not exceed a $O(\phi^{-2}t^{-1/2})$ bound when $a_t \in N_t$, which is guaranteed in the algorithm via resampling with high probability. This concludes the proof of Theorem 1.1.

1.5 Simulation Studies

In this section, we compare the performances of the three algorithms: (i) `LinTS` [Agrawal and Goyal, 2013], (ii) `BLTS` [Dimakopoulou et al., 2019], and (iii) the proposed `DRTS`. We use simulated data described as follows. The number of arms N is set to 10 or 20, and the dimension of contexts d is set to 20 or 30. For each element of the contexts $j = 1, \cdots, d$, we generate $[X_{1j}(t), \cdots, X_{Nj}(t)]$ from a normal distribution $\mathcal{N}(\mu_N, V_N)$ with mean $\mu_{10} = [-10, -8, \cdots, -2, 2, \cdots, 8, 10]^T$, or $\mu_{20} = [-20, -18, \cdots, -2, 2, \cdots, 18, 20]^T$, and the covariance matrix $V_N \in \mathbb{R}^{N\times N}$ has $V_N(i,i) = 1$ for every i and $V_N(i,k) = \rho$ for every $i \neq k$. We set $\rho = 0.5$ and truncate the sampled contexts to satisfy $\|X_i(t)\|_2 \leq 1$. To generate the stochastic rewards, we sam-

ple $\eta_i(t)$ independently from $\mathcal{N}(0,1)$. Each element of β follows a uniform distribution, $\mathcal{U}(-1/\sqrt{d}, 1/\sqrt{d})$.

All three algorithms have v as an input parameter which controls the variance of $\tilde{\beta}_i(t)$. `BLTS` and `DRTS` require a positive threshold γ which truncates the selection probability. We consider $v \in \{0.001, 0.01, 0.1, 1\}$ in all three algorithms, $\gamma \in \{0.01, 0.05, 0.1\}$ for `BLTS`, and set $\gamma = 1/(N+1)$ in `DRTS`. Then we report the minimum regrets among all combinations. The regularization parameter is $\lambda_t = \sqrt{t}$ in `DRTS` and $\lambda_t = 1$ in both `LinTS` and `BLTS`. To obtain an imputation estimator $\check{\beta}_t$ required in `DRTS`, we use ridge regression with $\{X_{a_\tau,\tau}, Y_{a_\tau,\tau}\}_{\tau=1}^{t-1}$, for each round t. Other implementation details are in Section 1.7.4.

Figure 1.1 shows the average of the cumulative regrets and the estimation error $\|\widehat{\beta}_t - \beta\|_2$ of the three algorithms based on 10 replications. The figures in the two left columns show the average cumulative regret according to the number of rounds with the best set of hyperparameters for each algorithm. The total rounds are $T = 20000$. The figures in the third columns show the average of the estimation error $\|\widehat{\beta}_t - \beta\|_2$. In the early stage, the estimation errors of `LinTS` and `BLTS` increase rapidly, while that of `DRTS` is stable. The stability of the DR estimator follows possibly by using full contexts and the regularization parameter $\lambda_t = \sqrt{t}$. This yields a large margin of estimation error among `LinTS`, `BLTS` and `DRTS`, especially when the dimension is large.

1.6 Conclusion

In this part of the thesis, we propose a novel algorithm for stochastic contextual linear bandits. Viewing the bandit problem as a missing data problem, we use the DR technique to employ all contexts including those that are not chosen. With the definition of super-unsaturated arms, we show a regret bound which only depends on the minimum eigenvalue of the sample covariance ma-

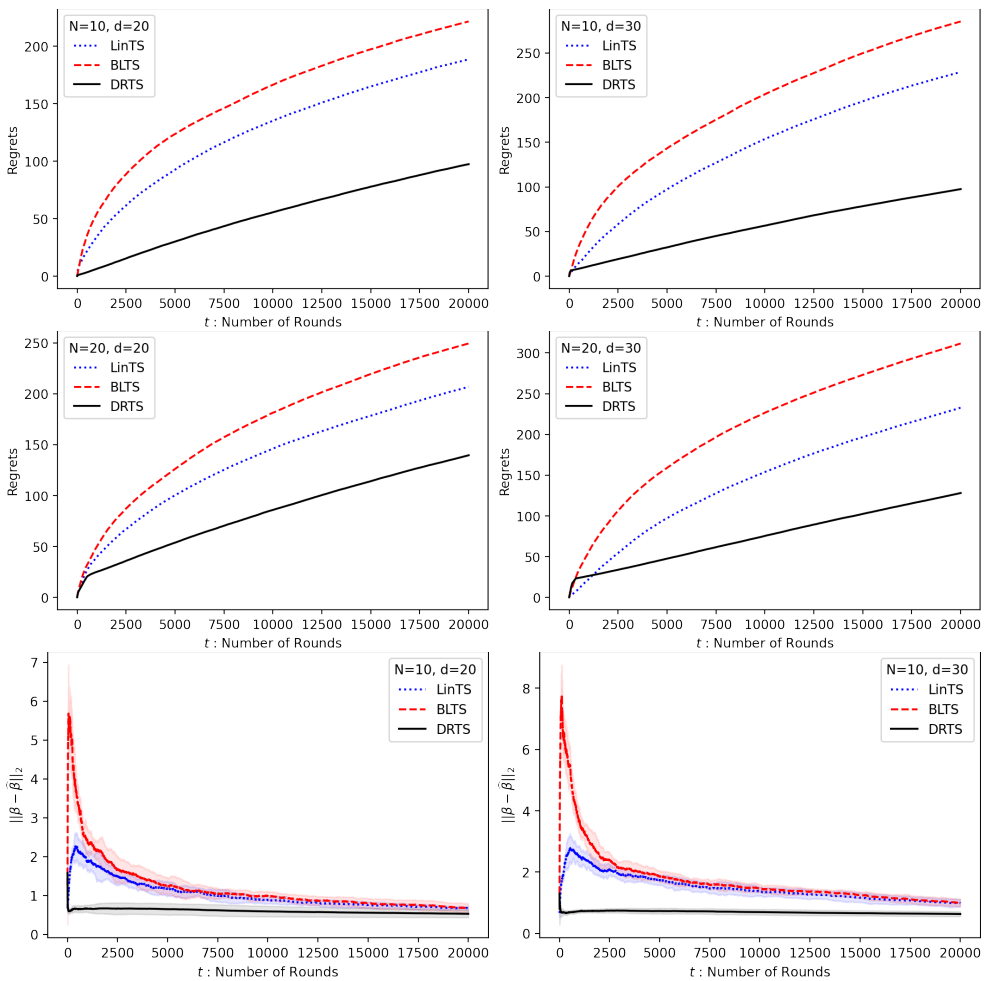

Figure 1.1: A Comparison of cumulative regrets and estimation errors of `LinTS`, `BLTS` and `DRTS`. Each line shows the averaged cumulative regrets (estimation errors, resp.) and the shaded area in the right two figures represents the standard deviations over 10 repeated experiments.

trices. This new bound has $\tilde{O}(d\sqrt{T})$ rate in many practical scenarios, which is improved by a factor of $\sqrt{d}$ compared to the previous `LinTS` regret bounds. Simulation studies show that the proposed algorithm performs better than other `LinTS` algorithms in a large dimension.

1.7 Appendix

1.7.1 Detailed Analysis of the Resampling

In this subsection, we give details about the issues which can be raised from the resampling in Algorithm 1.1.

1.7.1.1 Precise Definition of Action Selection

We give precise definition of the action at round t, a_t. For each round $t \geq 2$, given $\mathcal{H}_t$, let $a_t^{(1)}, a_t^{(2)}, \ldots, a_t^{(M_t)}$ to be maximum possible sequence of actions to be resampled. These actions are IID, with $\mathbb{P}\left(a_t^{(1)} = i \middle| \mathcal{H}_t\right) = \tilde{\pi}_i(t)$ for $i = 1, \ldots, N$. Define a subset of arms $\tilde{\Gamma}_t := \{i : \tilde{\pi}_i(t) > \gamma\}$ and a stopping time

$$\mathcal{T} := \inf\{m \geq 1 : a_t^{(m)} \in \tilde{\Gamma}_t\} \tag{1.14}$$

with respect to the filtration $\mathcal{F}_m := \mathcal{H}_t \cup \{a_t^{(1)}, \ldots, a_t^{(m)}\}$. Since the algorithm stops resampling when the candidate action is in $\tilde{\Gamma}_t$, the stopping time $\mathcal{T}$ is the actual number of resampling in algorithm. Thus we can write the action after resampling as $a_t := a_t^{(\min\{\mathcal{T}, M_t\})}$.

1.7.1.2 Computing the Probability of Selection

The probability of selection $\pi_{i,t} := \mathbb{P}\left(a_t = i | \mathcal{H}_t\right)$ is not the same as $\tilde{\pi}_i(t)$ due to resampling. This might cause the problem of computing $\pi_{i,t}$ which is essential to compute $Y_i^{DR}(t)$. However, with the precise definition of a_t, we can derive a closed form for $\pi_{i,t}$.

First, we consider two cases separately: (i) the case when the resampling succeeds and (ii) the case when the resampling fails and the maximum possible number of resampling runs out. In case (i), $a_t \in \tilde{\Gamma}_t$, and for any $i \in \tilde{\Gamma}_t$, we have

$$\begin{aligned}
\mathbb{P}\left(a_t = i \middle| \mathcal{H}_t\right) =& \mathbb{P}\left(\mathcal{T} \leq M_t,\ a_t^{(\mathcal{T})} = i \middle| \mathcal{H}_t\right) \\
=& \sum_{m=1}^{M_t} \mathbb{P}\left(\mathcal{T} = m,\ a_t^{(m)} = i \middle| \mathcal{H}_t\right) \\
=& \sum_{m=1}^{M_t} \mathbb{P}\left(a_t^{(m)} = i \middle| \mathcal{H}_t\right) \left(\prod_{j=0}^{m-1} \mathbb{P}\left(a_t^{(j)} \notin \tilde{\Gamma}_t \middle| \mathcal{H}_t\right)\right) \\
=& \tilde{\pi}_i(t) \sum_{m=1}^{M_t} \left(1 - \sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)\right)^{m-1} \\
=& \tilde{\pi}_i(t) \frac{1 - \left(1 - \sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)\right)^{M_t}}{\sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)}.
\end{aligned} \tag{1.15}$$

Now, for the case (ii) $a_t \notin \tilde{\Gamma}_t$, and for any $i \notin \tilde{\Gamma}_t$, we have

$$\begin{aligned}
\mathbb{P}\left(a_t = i \middle| \mathcal{H}_t\right) &= \mathbb{P}\left(\mathcal{T} > M_t, a_t^{(M_t)} = i \middle| \mathcal{H}_t\right) \\
&= \mathbb{P}\left(\bigcap_{m=1}^{M_t-1} \left\{a_t^{(m)} \notin \tilde{\Gamma}_t\right\}, a_t^{(M_t)} = i \middle| \mathcal{H}_t\right) \\
&= \left(1 - \sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)\right)^{M_t-1} \tilde{\pi}_i(t).
\end{aligned} \tag{1.16}$$

With (1.15) and (1.16), we can compute $\pi_{i,t}$ for all $i = 1, \ldots, N$.

1.7.1.3 The Number of Maximum Possible Resampling

The proposed algorithm attempts resampling up to M_t times to find an arm in $\{i : \tilde{\pi}_i(t) > \gamma\}$. The main point in selecting M_t is to bound the probability that the resampling fails in finding an arm whose selection probability exceeds

γ for some δ, i.e.,

$$\mathbb{P}(a_t \notin \{i : \tilde{\pi}_i(t) > \gamma\}) \leq \delta/t^2. \tag{1.17}$$

Intuitively, as M_t increases, we have more opportunities for resampling and the probability that the resampling fails in finding arms in $\{i : \tilde{\pi}_i(t) > \gamma\}$ decreases. Since $\gamma < 1/N$, there exists j such that $\tilde{\pi}_j(t) > \gamma$, and the probability that the resampling fails is less than $1 - \gamma$ in each resampling trial.

Specifically, we can achieve (1.17) by choosing M_t as a minimum integer that exceeds $\log \frac{t^2}{\delta} / \log \frac{1}{1-\gamma}$. For any given $\delta \in (0, 1)$, the event $\{a_t \in \tilde{\Gamma}_t\}$ occurs with probability at least $1 - \delta/t^2$. By (1.14), we have

$$\begin{aligned}\mathbb{P}\left(a_t \notin \tilde{\Gamma}_t \middle| \mathcal{H}_t\right) =& \mathbb{P}\left(\mathcal{T} > M_t | \mathcal{H}_t\right) = \mathbb{P}\left(\bigcap_{m=1}^{M_t} \left\{a_t^{(m)} \notin \tilde{\Gamma}_t\right\} \middle| \mathcal{H}_t\right) \\ =& \left(1 - \sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)\right)^{M_t}.\end{aligned}$$

Since $\gamma < 1/N$, there exists at least one arm in $\tilde{\Gamma}_t$, and thus

$$\mathbb{P}\left(a_t \notin \tilde{\Gamma}_t \middle| \mathcal{H}_t\right) \leq (1 - \gamma)^{M_t}.$$

If we set M_t as a minimum integer that exceeds $\left(\log \frac{t^2}{\delta}\right)\left(\log \frac{1}{1-\gamma}\right)^{-1}$ then (1.17) holds. Thus, by choosing M_t for each round that satisfies (1.17), the algorithm finds an arm j such that $\tilde{\pi}_j(t) > \gamma$ in all rounds with high probability.

Selecting an arm from the set $\{i : \tilde{\pi}_i(t) > \gamma\}$ with high probability is crucial in achieving the regret bound of order $\tilde{O}(\phi^{-2}\sqrt{T})$ for two reasons. First, it guarantees that the arm is super-unsaturated and the novel regret decomposition (1.6) holds to achieve the novel regret bound. Let N_t be the set of super-unsaturated arm defined in (1.5). With Lemma 1.2, we prove that if

$\tilde{\pi}_i(t) > \gamma$ then $i \in N_t$, which implies $\tilde{\Gamma}_t \subseteq N_t$, and thus

$$\mathbb{P}\left(a_t \in N_t \middle| \mathcal{H}_t\right) \geq \mathbb{P}\left(a_t \in \tilde{\Gamma}_t \middle| \mathcal{H}_t\right).$$

Thus we can conclude that a_t is super-unsaturated with probability at least $1 - \delta/t^2$ with M_t defined in Section 1.7.1. Second, the inverse probability, $\pi_{a_t}(t)^{-1}$ is bounded by γ^{-1} which appears in $Y_i^{DR}(t)$ and the proof of Theorem 1.3. From (1.15) we can deduce $\pi_{a_t}(t) \geq \tilde{\pi}_{a_t}(t) > \gamma$, for $a_t \in \tilde{\Gamma}_t$. This shows that the assumptions regarding $\pi_{a_t,t}$ in Theorem 1.3 hold.

1.7.2 Technical Lemmas

Lemma 1.8. *[Wainwright, 2019, Theorem 2.19] (Bernstein Concentration) Let $\{D_k, \mathfrak{S}_k\}_{k=1}^{\infty}$ be a martingale difference sequence and suppose D_k is σ-sub-Gaussian in an adapted sense, i.e. for all $\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda D_k} \middle| \mathfrak{S}_{k-1}\right] \leq e^{\lambda^2\sigma^2/2}$ almost surely. Then for all $x \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} D_k\right| \geq x\right) \leq 2\exp\left(-\frac{x^2}{2n\sigma^2}\right).$$

Lemma 1.9. *[Azuma, 1967] (Azuma-Hoeffding inequality) If a super-martingale $(Y_t; t \geq 0)$ corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t, for all $t = 1, \ldots, T$, then for any $a \geq 0$,*

$$\mathbb{P}\left(Y_T - Y_0 \geq a\right) \leq e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}.$$

Lemma 1.10. *[Lee et al., 2016, Lemma 2.3] Let $\{N_t\}$ be a martingale on a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Then there exists a $\mathbb{R}^2$-valued martingale $\{P_t\}$ such that for any time $t \geq 0$, $\|P_t\|_2 = \|N_t\|_{\mathcal{H}}$ and $\|P_{t+1} - P_t\|_2 = \|N_{t+1} - N_t\|_{\mathcal{H}}$.*

Lemma 1.11. *[Chung and Lu, 2006, Lemma 1, Theorem 32] For a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_T$, suppose each random variable X_t is $\mathcal{F}_t$-measurable*

martingale, for $0 \leq t \leq T$. *Let* B_t *denote the bad set associated with the following admissible condition:*

$$|X_t - X_{t-1}| \leq c_t,$$

for $1 \leq t \leq T$, *where* $c_1, \ldots, c_n$ *are non-negative numbers. Then there exists a collection of random variables* $Y_0, \ldots, Y_T$ *such that* Y_t *is* $\mathcal{F}_t$*-measurable martingale such that*

$$|Y_t - Y_{t-1}| \leq c_t,$$

and $\{\omega : Y_t(\omega) \neq X_t(\omega)\} \subset B_t$, *for* $0 \leq t \leq T$.

Lemma 1.12. *Suppose a random variable* X *satisfies* $\mathbb{E}[X] = 0$, *and let* η *be an* σ*-sub-Gaussian random variable. If* $|X| \leq |\eta|$ *almost surely, then* X *is* $C\sigma$*-sub-Gaussian for some absolute constant* $C > 0$.

Proof. By Proposition 2.5.2 in Vershynin [2018], there exists an absolute constant $C_1 > 0$ such that

$$\mathbb{E} \exp\left(\lambda^2 \eta^2\right) \leq \exp\left(\frac{\lambda^2 C_1^2 \sigma^2}{2}\right), \quad \forall \lambda \in \left[-\frac{\sqrt{2}}{C_1 \sigma}, \frac{\sqrt{2}}{C_1 \sigma}\right].$$

Since $|X| \leq |\eta|$ almost surely,

$$\mathbb{E} \exp\left(\lambda^2 X^2\right) \leq \exp\left(\frac{\lambda^2 C_1^2 \sigma^2}{2}\right), \quad \forall \lambda \in \left[-\frac{\sqrt{2}}{C_1 \sigma}, \frac{\sqrt{2}}{C_1 \sigma}\right].$$

Since $\mathbb{E}[X] = 0$, by Proposition 2.5.2 in Vershynin [2018], there exists an absolute constant $C_2 > 0$ such that

$$\mathbb{E} \exp\left(\lambda X\right) \leq \exp\left(\frac{\lambda^2 C_1^2 C_2^2 \sigma^2}{2}\right), \quad \forall \lambda \in \mathbb{R}.$$

Setting $C = C_1 C_2$ completes the proof. □

1.7.3 Proofs of Theoretical Results

1.7.3.1 Proof Theorem 1.1

In subsection 1.7.1, we prove that $a_t \in \tilde{\Gamma}_t$ with probability at least $1 - \delta/t^2$, for all $t \geq 2$. Thus, for any $x > 0$,

$$\begin{aligned}
\mathbb{P}\left(R(T) > x\right) \leq & \mathbb{P}\left(R(T) > x, \bigcap_{t=2}^{T} \left\{a_t \in \tilde{\Gamma}_t\right\}\right) + \mathbb{P}\left(\bigcup_{t=2}^{T} \left\{a_t \notin \tilde{\Gamma}_t\right\}\right) \\
\leq & \mathbb{P}\left(R(T) > x, \bigcap_{t=2}^{T} \left\{a_t \in \tilde{\Gamma}_t\right\}\right) + \delta \\
\leq & \mathbb{P}\left(2 + \sum_{t=2}^{T} \texttt{regret}(t) > x, \bigcap_{t=2}^{T} \left\{a_t \in \tilde{\Gamma}_t\right\}\right) + \delta
\end{aligned}$$

The last inequality holds by Assumption 1. Since $\tilde{\Gamma}_t$ is a subset of N_t and by (1.6),

$$\begin{aligned}
&\mathbb{P}\left(R(T) > x\right) \\
&\leq \mathbb{P}\left(2 + \sum_{t=2}^{T} \left\{2\left\|\widehat{\beta}_{t-1} - \beta\right\|_2 + \sqrt{\left\|X_{a_t^*,t}\right\|^2_{V_{t-1}^{-1}} + \left\|X_{a_t,t}\right\|^2_{V_{t-1}^{-1}}}\right\} > x, \bigcap_{t=2}^{T} \left\{a_t \in \tilde{\Gamma}_t\right\}\right) \\
&+ \delta.
\end{aligned} \tag{1.18}$$

To bound the term $\left\|\widehat{\beta}_t - \beta\right\|_2$ for all $t = 1, \ldots, T-1$, we use Theorem 1.3. Before that, we need to verify whether the two assumptions on $\pi_{i,t}$ in Theorem 1.3 hold.

First, we show that $\pi_{a_t,t} > \gamma$. When $t = 1$, we have $\tilde{\pi}_i(1) = 1/N$ for all i. Since $\gamma < 1/N$, we do not need resampling and thus $\pi_{i,t} = \tilde{\pi}_i(t) > \gamma$. When $t \geq 2$, $a_t \in \tilde{\Gamma}_t$ is already concerned in (1.18), and thus $\tilde{\pi}_{a_t}(t) > \gamma$. From (1.15), we can deduce that $\pi_{i,t} > \tilde{\pi}_i(t)$ for all $i \in \tilde{\Gamma}_t$, and thus $\pi_{a_t,t} > \gamma$.

Now, we prove that $\pi_{i,t} > 0$ for all i and t. The case of $t = 1$ is already

proved above. When $t \geq 2$, from (1.15), we have

$$\pi_{i,t} := \mathbb{P}\left(a_t = i \middle| \mathcal{H}_t\right) = \tilde{\pi}_i(t) \sum_{m=1}^{M_t} \left(1 - \sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)\right)^{m-1} > \tilde{\pi}_i(t) > \gamma,$$

for all $i \in \tilde{\Gamma}_t$. If there exists an arm $i \notin \tilde{\Gamma}_t$, from (1.16),

$$\pi_{i,t} = \left(1 - \sum_{i \in \tilde{\Gamma}_t} \tilde{\pi}_i(t)\right)^{M_t - 1} \tilde{\pi}_i(t).$$

The first term is positive since there exists an arm $i \notin \tilde{\Gamma}_t$. The second term is also positive since the distribution of $\tilde{\beta}_i(t)$ has support $\mathbb{R}^d$, which implies that

$$\tilde{\pi}_i(t) := \mathbb{P}\left(X_{i,t}^T \tilde{\beta}_i(t) = \max_j X_{j,t}^T \tilde{\beta}_j(t) \middle| \mathcal{H}_t\right) > 0,$$

for all i. Thus, $\pi_{i,t} > 0$ for all i and t. This implies that the two assumptions on $\pi_{i,t}$ in Theorem 1.3 hold.

Now we can use Theorem 1.3 and Lemma 1.6 to have

$$\begin{aligned} \left\|\widehat{\beta}_{t-1} - \beta\right\|_2 &\leq \frac{C_{b,\sigma}}{\phi^2 \sqrt{t-1}} \sqrt{\log \frac{12(t-1)^2}{\delta}}, \\ \sqrt{\left\|X_{a_t^*,t}\right\|_{V_{t-1}^{-1}}^2 + \left\|X_{a_t,t}\right\|_{V_{t-1}^{-1}}^2} &\leq \frac{1}{\phi \sqrt{N(t-1)}}, \end{aligned}$$

for all $t = 2, \ldots, T$ with probability at least $1 - \delta$. Thus, setting

$$x = 2 + \frac{4C_{b,\sigma}}{\phi^2} \sqrt{T \log \frac{12T^2}{\delta}} + \frac{2\sqrt{T}}{\phi \sqrt{N}}$$

in (1.18) proves the result.

1.7.3.2 Proof of Lemma 1.2

Proof. First, we bring attention to the fact that the optimal arm a_t^* is in N_t by definition. Suppose that the estimated reward of the optimal arm, $\tilde{Y}_{a_t^*}(t)$ is greater than $\tilde{Y}_j(t)$ for all $j \notin N_t$. In this case, any arm $j \notin N_t$ cannot be the $m_t := \arg\max_i \tilde{Y}_i(t)$. Then we have

$$\begin{aligned}\mathbb{P}\left(m_t \in N_t \middle| \mathcal{H}_t\right) &\geq \mathbb{P}\left(\tilde{Y}_{a_t^*}(t) > \tilde{Y}_j(t), \forall j \notin N_t \middle| \mathcal{H}_t\right) \\ &= \mathbb{P}\left(Z_j(t) > \{X_{j,t} - X_{a_t^*,t}\}^T \widehat{\beta}_{t-1}, \forall j \notin N_t \middle| \mathcal{H}_t\right),\end{aligned}$$

where $Z_j(t) := \tilde{Y}_{a_t^*}(t) - \tilde{Y}_j(t) - \{X_{a_t^*,t} - X_{j,t}\}^T \widehat{\beta}_{t-1}$. Note that $Z_j(t)$ is a Gaussian random variable with mean 0 and variance $v^2(\|X_{a_t^*,t}\|^2_{V_{t-1}^{-1}} + \|X_{j,t}\|^2_{V_{t-1}^{-1}})$ given $\mathcal{H}_t$. For all $j \notin N_t$,

$$\begin{aligned}\{X_{j,t} - X_{a_t^*,t}\}^T \widehat{\beta}_{t-1} &= \{X_{j,t} - X_{a_t^*,t}\}^T \{\widehat{\beta}_{t-1} - \beta\} - \Delta_j(t) \\ &\leq 2\left\|\widehat{\beta}_t - \beta\right\|_2 - \Delta_j(t) \leq -\sqrt{\|X_{a_t^*,t}\|^2_{V_{t-1}^{-1}} + \|X_{j,t}\|^2_{V_{t-1}^{-1}}}.\end{aligned}$$

The last inequality is due to $j \notin N_t$. Thus, we can conclude that

$$\begin{aligned}\mathbb{P}\left(m_t \in N_t \middle| \mathcal{H}_t\right) \geq &\mathbb{P}\left(\frac{Z_j(t)}{v\sqrt{\|X_{a_t^*,t}\|^2_{V_{t-1}^{-1}} + \|X_{j,t}\|^2_{V_{t-1}^{-1}}}} > -\frac{1}{v}, \forall j \notin N_t \middle| \mathcal{H}_t\right) \\ := &\mathbb{P}\left(Y_j > -v^{-1}, \forall j \neq N_t \middle| \mathcal{H}_t\right).\end{aligned}$$

Using the fact that

$$Y_j := \frac{Z_j(t)}{v\sqrt{\|X_{a_t^*,t}\|^2_{V_{t-1}^{-1}} + \|X_{j,t}\|^2_{V_{t-1}^{-1}}}}$$

is a standard Gaussian random variable given $\mathcal{H}_t$, we have

$$\mathbb{P}\left(Y_j \leq -v^{-1} \middle| \mathcal{H}_t\right) \leq \exp\left(-\frac{1}{2v^2}\right).$$

Setting $v = \{2\log(N/(1-\gamma N))\}^{-1/2}$ gives

$$\mathbb{P}\left(Y_j \leq -v^{-1} \middle| \mathcal{H}_t\right) = \exp\left(-\log\left(N/(1-\gamma N)\right)\right) = \frac{1-\gamma N}{N}.$$

Thus,

$$\begin{aligned}
\mathbb{P}\left(m_t \in N_t \middle| \mathcal{H}_t\right) \geq & 1 - \mathbb{P}\left(Y_j \leq -v^{-1}, \exists j \neq N_t \middle| \mathcal{H}_t\right) \\
\geq & 1 - \sum_{j \neq N_t} \mathbb{P}\left(Y_j < -v^{-1} \middle| \mathcal{H}_t\right) \\
\geq & 1 - (1-\gamma N) \\
= & \gamma N \\
\geq & 1 - \gamma.
\end{aligned}$$

The last inequality holds due to $\gamma \geq 1/(N+1)$. □

1.7.3.3 Proof of Theorem 1.3

Proof. Fix $t = 1, \ldots, T$ and let $V_t := \sum_{\tau=1}^{t} \sum_{i=1}^{N} X_{i,\tau} X_{i,\tau}^T + \lambda_t I$. For each i and τ, let $\widehat{\eta}_{i,\tau}^{DR} = Y_i^{DR}(\tau) - X_{i,\tau}^T \beta$. Then

$$\widehat{\beta}_t = \beta + V_t^{-1}\left(-\lambda_t \beta + \sum_{\tau=1}^{t} \sum_{i=1}^{N} \widehat{\eta}_{i,\tau}^{DR} X_{i,\tau}\right).$$

To bound $\left\|\widehat{\beta}_t - \beta\right\|_2$,

$$\begin{aligned}\left\|\widehat{\beta}_t - \beta\right\|_2 &= \left\|V_t^{-1}\left(-\lambda_t\beta + \sum_{\tau=1}^{t}\sum_{i=1}^{N}\widehat{\eta}_{i,\tau}^{DR}X_{i,\tau}\right)\right\|_2 \\ &\leq \left\|V_t^{-1}\right\|_2 \left\|\left(-\lambda_t\beta + \sum_{\tau=1}^{t}\sum_{i=1}^{N}\widehat{\eta}_{i,\tau}^{DR}X_{i,\tau}\right)\right\|_2 \\ &= \left\{\lambda_{\min}\left(V_t\right)\right\}^{-1}\left\|\left(-\lambda_t\beta + \sum_{\tau=1}^{t}\sum_{i=1}^{N}\widehat{\eta}_{i,\tau}^{DR}X_{i,\tau}\right)\right\|_2.\end{aligned}$$

By Assumption 1, $\|\beta\|_2 \leq 1$. Using triangle inequality,

$$\left\|\widehat{\beta}_t - \beta\right\|_2 \leq \left\{\lambda_{\min}\left(V_t\right)\right\}^{-1}\lambda_t + \left\{\lambda_{\min}\left(V_t\right)\right\}^{-1}\left\|\sum_{\tau=1}^{t}\sum_{i=1}^{N}\widehat{\eta}_{i,\tau}^{DR}X_{i,\tau}\right\|_2. \tag{1.19}$$

We will bound the first term in (1.19). Let $\mathrm{Tr}(A)$ be the trace of a matrix A. By the definition of the Frobenious norm, for $\tau = 1, \ldots, t$, and for $i = 1, \ldots, N$,

$$\left\|\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T\right\|_F \leq \sum_{i=1}^{N}\sqrt{\mathrm{Tr}\left(X_{i,\tau}X_{i,\tau}^T X_{i,\tau}X_{i,\tau}^T\right)} \leq N.$$

By Assumptions 3 and 4, $\left\{\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T\right\}_{\tau=1}^{t}$ are independent random variables such that $\mathbb{E}\left[\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T\right] \geq N\phi^2 > 0$. Let $\delta \in (0,1)$ be given. By Lemma 1.6, if we set $\lambda_t = 4\sqrt{2}N\sqrt{t\log\frac{12t^2}{\delta}}$,

$$\left\{\lambda_{\min}\left(V_t\right)\right\}^{-1} < \frac{1}{\phi^2 N t},$$

holds with probability at least $1-\delta/(3t^2)$. Thus, the first term can be bounded by

$$\left\{\lambda_{\min}\left(V_t\right)\right\}^{-1}\lambda_t \leq \frac{4\sqrt{\log\frac{12t^2}{\delta}}}{\sqrt{t}\phi^2}. \tag{1.20}$$

Now we will bound the second term in (1.19). Let $U_i(\tau) := X_{i,\tau}X_{i,\tau}^T(\breve{\beta}_\tau - \beta)$.

Then we can decompose $\widehat{\eta}_{i,\tau}^{DR} X_{i,\tau}$ as,

$$\begin{aligned}\widehat{\eta}_{i,\tau}^{DR} X_{i,\tau} =& U_i(\tau) + \frac{\mathbb{I}\left(a_\tau = i\right)}{\pi_{i,\tau}} \left(Y_{i,\tau} - X_{i,\tau}^T \breve{\beta}_\tau\right) X_{i,\tau} \\ =& \left(1 - \frac{\mathbb{I}\left(a_\tau = i\right)}{\pi_{i,\tau}}\right) U_i(\tau) + \frac{\mathbb{I}\left(a_\tau = i\right)}{\pi_{i,\tau}} \eta_{i,\tau} X_{i,\tau} \\ :=& D_i(\tau) + E_i(\tau).\end{aligned} \tag{1.21}$$

Let $D_\tau := \sum_{i=1}^N D_i(\tau)$. Since $U_i(\tau)$ is $\mathcal{H}_\tau$-measurable, the conditional expectation of D_τ is

$$\begin{aligned}\mathbb{E}\left[D_\tau \middle| \mathcal{H}_\tau\right] &= \mathbb{E}\left[\sum_{i=1}^N D_i(\tau) \middle| \mathcal{H}_\tau\right] = \sum_{i=1}^N \mathbb{E}\left[\left(1 - \frac{\mathbb{I}\left(a_\tau = i\right)}{\pi_{i,\tau}}\right) \middle| \mathcal{H}_\tau\right] U_i(\tau) \\ &= \sum_{i=1}^N \left(1 - \frac{\pi_{i,\tau}}{\pi_{i,\tau}}\right) U_i(\tau) = 0\end{aligned}$$

Thus, $\{\sum_{u=1}^\tau D_\tau\}_{\tau=1}^t$ is a martingale sequence on $\left(\mathbb{R}^d, \|\cdot\|_2\right)$ with respect to $\mathcal{H}_\tau$, with

$$\begin{aligned}\|D_\tau\|_2 \leq& \sum_{i=1}^N \left|1 - \frac{\mathbb{I}\left(a_\tau = i\right)}{\pi_{i,\tau}}\right| \|U_i(\tau)\|_2 \\ \leq& \sum_{i=1}^N \left|1 - \frac{\mathbb{I}\left(a_\tau = i\right)}{\pi_{i,\tau}}\right| \left\|\breve{\beta}_\tau - \beta\right\|_2 \\ \leq& \left(N + \pi_{a_\tau,\tau}^{-1}\right) b.\end{aligned}$$

By Lemma 1.10, since $\left(\mathbb{R}^d, \|\cdot\|_2\right)$ is a Hilbert space, there exists a martingale sequence $\{P_\tau\}_{\tau=1}^t = \left\{\left(P_\tau^{(1)}, P_\tau^{(2)}\right)^T\right\}_{\tau=1}^t$ on $\mathbb{R}^2$ such that

$$\left\|\sum_{u=1}^\tau D_u\right\|_2 = \|P_\tau\|_2, \quad \|D_\tau\|_2 = \|P_\tau - P_{\tau-1}\|_2 \tag{1.22}$$

and $P_0 = 0$, for any $\tau = 1, \ldots, t$. Since $\left\|\breve{\beta}_\tau - \beta\right\|_2 \leq b$, for $r = 1, 2$

$$\left|P_\tau^{(r)} - P_{\tau-1}^{(r)}\right| \leq \|P_\tau - P_{\tau-1}\|_2 = \|D_\tau\|_2 \leq \left(N + \pi_{a_\tau,\tau}^{-1}\right) b.$$

By Lemma 1.11, there exists a martingale sequence $\left\{N_\tau^{(r)}\right\}_{\tau=1}^t$ such that $\left|N_\tau^{(r)} - N_{\tau-1}^{(r)}\right| \leq (N + \gamma^{-1})b$, for all $\tau = 1, \ldots, t$ and

$$\left\{N_t^{(r)} \neq P_t^{(r)}\right\} \subset \bigcup_{\tau=1}^t \left\{\left|P_\tau^{(r)} - P_{\tau-1}^{(r)}\right| > (N + \gamma^{-1})b\right\} \subset \bigcup_{\tau=1}^t \left\{\pi_{a_\tau,\tau} \leq \gamma\right\}. \tag{1.23}$$

Thus, by (1.22) and (1.23), for any $x > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left\|\sum_{u=1}^t D_u\right\|_2 > x,\ \bigcap_{\tau=1}^T \left\{\pi_{a_\tau,\tau} > \gamma\right\}\right) =& \mathbb{P}\left(\|P_t\|_2 \geq x,\ \bigcap_{\tau=1}^T \left\{\pi_{a_\tau,\tau} > \gamma\right\}\right) \\
\leq& \mathbb{P}\left(\sum_{r=1}^2 \left|P_t^{(r)}\right| \geq x,\ \bigcap_{\tau=1}^t \left\{\pi_{a_\tau,\tau} > \gamma\right\}\right) \\
\leq& \sum_{r=1}^2 \mathbb{P}\left(\left|P_t^{(r)}\right| \geq \frac{x}{2},\ \bigcap_{\tau=1}^t \left\{\pi_{a_\tau,\tau} > \gamma\right\}\right) \\
\leq& \sum_{r=1}^2 \mathbb{P}\left(\left|P_t^{(r)}\right| \geq \frac{x}{2},\ N_t^{(r)} = P_t^{(r)}\right) \\
\leq& \sum_{r=1}^2 \mathbb{P}\left(\left|N_t^{(r)}\right| \geq \frac{x}{2}\right).
\end{aligned}$$

Since $N_\tau^{(r)}$ has bounded differences, we can apply Lemma 1.9 to have

$$\sum_{r=1}^2 \mathbb{P}\left(\left|N_t^{(r)}\right| \geq \frac{x}{2}\right) \leq 4\exp\left(-\frac{x^2}{8tb^2\left(N + \gamma^{-1}\right)^2}\right).$$

Thus, with probability at least $1 - \delta/(3t^2)$,

$$\left\| \sum_{\tau=1}^{t} D_\tau \right\|_2 \leq 2\sqrt{2}b(N + \gamma^{-1})\sqrt{\log \frac{12t^2}{\delta}} \tag{1.24}$$

holds with the event $\bigcap_{t=1}^{T} \{\pi_{a_t,t} > \gamma\}$.

Now we will bound the $E_i(\tau)$ term in (1.21). Under the event $\bigcap_{t=1}^{T} \{\pi_{a_t,t} > \gamma\}$, we have

$$\sum_{\tau=1}^{t} \sum_{i=1}^{N} E_i(\tau) = \sum_{\tau=1}^{t} \frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}} X_{a_\tau,\tau} = \sum_{\tau=1}^{t} \frac{\mathbb{I}\left(\pi_{a_t,t} > \gamma\right) \eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}} X_{a_\tau,\tau}$$

For each $\tau \geq 1$, define a filtration $\mathcal{F}_{\tau-1} := \mathcal{H}_\tau \cup \{a_\tau\}$. Then $X_{a_\tau,\tau}$ is $\mathcal{F}_{\tau-1}$-measurable. By Assumption 2, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda \frac{\mathbb{I}\left(\pi_{a_t,t} > \gamma\right) \eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}} \right) \middle| \mathcal{F}_{\tau-1} \right] \leq \exp\left(\frac{\lambda^2 \mathbb{I}\left(\pi_{a_t,t} > \gamma\right) \sigma^2}{2\pi_{a_\tau,\tau}^2} \right) \leq \exp\left(\frac{\lambda^2 \sigma^2}{2\gamma^2} \right),$$

almost surely. Since $\|X_{a_\tau,\tau}\|_2 \leq 1$, by Lemma 1.4, there exists an absolute constant $C > 0$ such that, with probability at least $1 - \delta/(3t^2)$,

$$\left\| \sum_{\tau=1}^{t} \sum_{i=1}^{N} E_i(\tau) \right\|_2 \leq 2C\sigma\gamma^{-1}\sqrt{t}\sqrt{\log \frac{12t^2}{\delta}}. \tag{1.25}$$

Thus, with (1.20), (1.24), and (1.25), under the event $\bigcap_{t=1}^{T} \{\pi_{a_t,t} > \gamma\}$, we

have

$$
\begin{aligned}
\left\|\widehat{\beta}_t - \beta\right\|_2 \leq & \frac{4\sqrt{\log \frac{12t^2}{\delta}}}{\sqrt{t}\phi^2} \\
& + \frac{1}{\phi^2 N t}\left(4\left(N+\gamma^{-1}\right) b\sqrt{t}\sqrt{\log\frac{12t^2}{\delta}} + 2C\sigma\gamma^{-1}\sqrt{t}\sqrt{\log\frac{12t^2}{\delta}}\right) \\
\leq & \frac{4+4b+\gamma^{-1}N^{-1}\left(4b+2C\sigma\right)}{\phi^2\sqrt{t}}\sqrt{\log\frac{12t^2}{\delta}} \\
\leq & \frac{4+4b+2\left(4b+2C\sigma\right)}{\phi^2\sqrt{t}}\sqrt{\log\frac{12t^2}{\delta}} \\
:= & \frac{C_{b,\sigma}}{\phi^2\sqrt{t}}\sqrt{\log\frac{12t^2}{\delta}},
\end{aligned}
\tag{1.26}
$$

with probability at least $1-\delta/t^2$. Since (1.26) holds for all $t = 1, \ldots, T$,

$$
\begin{aligned}
& \leq \mathbb{P}\left(\bigcup_{t=1}^{T}\left\{\left\|\widehat{\beta}_t - \beta\right\|_2 > \frac{C_{b,\sigma}}{\phi^2\sqrt{t}}\sqrt{\log\frac{12t^2}{\delta}}\right\}, \bigcap_{t=1}^{T}\left\{\pi_{a_t,t} > \gamma\right\}\right) \\
& \leq \mathbb{P}\left(\bigcup_{t=1}^{T}\left\{\left\|\widehat{\beta}_t - \beta\right\|_2 > \frac{C_{b,\sigma}}{\phi^2\sqrt{t}}\sqrt{\log\frac{12t^2}{\delta}}\right\}, \bigcap_{t=1}^{T}\left\{\pi_{a_t,t} > \gamma\right\}\right) \\
& \leq \sum_{t=1}^{T}\mathbb{P}\left(\left\|\widehat{\beta}_t - \beta\right\|_2 > \frac{C_{b,\sigma}}{\phi^2\sqrt{t}}\sqrt{\log\frac{12t^2}{\delta}}, \bigcap_{t=1}^{T}\left\{\pi_{a_t,t} > \gamma\right\}\right) \\
& \leq \delta.
\end{aligned}
$$

□

1.7.3.4 Proof of Lemma 1.4

Proof. Fix a $t \geq 1$. Since for each $\tau = 1, \ldots, t$, $\mathbb{E}\left[\eta(\tau)\middle|\mathcal{F}_{\tau-1}\right] = 0$ and $X(\tau)$ is $\mathcal{F}_{\tau-1}$-measurable, the stochastic process,

$$
\left\{\sum_{\tau=1}^{u}\eta(\tau)X(\tau)\right\}_{u=1}^{t} \tag{1.27}
$$

is a $\mathbb{R}^d$-martingale. Since $(\mathbb{R}^d, \|\cdot\|_2)$ is a Hilbert space, by Lemma 1.10, there exists a $\mathbb{R}^2$-martingale $\{M_u\}_{u=1}^t$ such that

$$\left\|\sum_{\tau=1}^{u} \eta(\tau)X(\tau)\right\|_2 = \|M_u\|_2\,,\ \|\eta(u)X(u)\|_2 = \|M_u - M_{u-1}\|_2\,, \tag{1.28}$$

and $M_0 = 0$. Set $M_u = (M_1(u), M_2(u))^T$. Then for each $i = 1, 2$, and $u \geq 2$, by the assumption $\|X(u)\|_2 \leq 1$,

$$\begin{aligned} |M_i(u) - M_i(u-1)| &\leq \|M_u - M_{u-1}\|_2 \\ &= \|\eta(u)X(u)\|_2 \\ &\leq |\eta(u)|\,. \end{aligned}$$

By Lemma 1.12, $M_i(u) - M_i(u-1)$ is $C\sigma$-sub-Gaussian for some constant $C > 0$. By Lemma 1.9, for $x > 0$,

$$\begin{aligned} \mathbb{P}\left(|M_i(t)| > x\right) =& \mathbb{P}\left(\left|\sum_{u=1}^{t} M_i(u) - M_i(u-1)\right| > x\right) \\ \leq& 2\exp\left(-\frac{x^2}{2tC^2\sigma^2}\right), \end{aligned}$$

for each $i = 1, 2$. Thus, with probability $1 - \delta/(2t^2)$,

$$M_i(t)^2 \leq 2tC^2\sigma^2 \log\frac{4t^2}{\delta}.$$

In summary, with probability at least $1 - \delta/t^2$,

$$\left\|\sum_{\tau=1}^{t} \eta(\tau)X(\tau)\right\|_2 = \sqrt{M_1(t)^2 + M_2(t)^2} \leq 2C\sigma\sqrt{t}\sqrt{\log\frac{4t^2}{\delta}}.$$

□

1.7.3.5 Proof of Lemma 1.6

Proof. For each $\tau = 1, \ldots, t$, let $\Sigma_\tau = \mathbb{E}\left[P(\tau) | \mathcal{F}_{\tau-1}\right]$. Since $P(\tau)$ and Σ_τ are symmetric matrices,

$$\begin{aligned}
\lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau) + \lambda_t I\right) =& \lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau)\right) + \lambda_t \\
=& \lambda_{\min}\left(\sum_{\tau=1}^{t} \{P(\tau) - \Sigma_\tau\} + \sum_{\tau=1}^{t} \Sigma_\tau\right) + \lambda_t \\
\geq& \lambda_{\min}\left(\sum_{\tau=1}^{t} \{P(\tau) - \Sigma_\tau\}\right) + \sum_{\tau=1}^{t} \lambda_{\min}\left(\Sigma_\tau\right) + \lambda_t \\
\geq& \lambda_{\min}\left(\sum_{\tau=1}^{t} \{P(\tau) - \Sigma_\tau\}\right) + \phi^2 t + \lambda_t.
\end{aligned}$$

The last inequality uses the fact that $\lambda_{\min}\left(\Sigma_\tau\right) \geq \phi^2$ for all τ.

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau) + \lambda_t I\right) \leq \phi^2 t\right) \leq& \mathbb{P}\left(\lambda_{\min}\left(\sum_{\tau=1}^{t} \{P(\tau) - \Sigma_\tau\}\right) + \lambda_t \leq 0\right) \\
=& \mathbb{P}\left(\lambda_{\max}\left(\sum_{\tau=1}^{t} \{\Sigma_\tau - P(\tau)\}\right) \geq \lambda_t\right) \\
\leq& \mathbb{P}\left(\left\|\sum_{\tau=1}^{t} \{\Sigma_\tau - P(\tau)\}\right\|_F \geq \lambda_t\right).
\end{aligned} \tag{1.29}$$

Set $S_u = \sum_{\tau=1}^{u} \{\Sigma_\tau - P(\tau)\}$. Then $\{S_u\}_{u=1}^{t}$ can be regarded as a martingale sequence on $\mathbb{R}^{d\times d}$ with respect to $\{P(\tau)\}_{\tau=1}^{t}$. Note that $\left(\mathbb{R}^{d\times d}, \|\cdot\|_F\right)$ is a Hilbert space. By Lemma 1.10, there exists a martingale sequence $\left\{D_u = (D_1(u), D_2(u))^T\right\}_{u=1}^{t}$ on $\mathbb{R}^2$ such that

$$\|S_u\|_F = \sqrt{D_1(u)^2 + D_2(u)^2}, \quad \|M_u - \Sigma_u\|_F = \|D_u - D_{u-1}\|_2, \tag{1.30}$$

for any $u \geq 1$, and $D_0 = 0$. Then, for any $i = 1, 2$,

$$|D_i(u) - D_i(u-1)|^2 \leq \|D_u - D_{u-1}\|_2^2 = \|P(u) - \Sigma_u\|_F^2$$

Since $\|P(u) - \Sigma_u\|_F \leq 2c$, we can apply Lemma 1.9 for $D_1(\tau)$, and $D_2(\tau)$, respectively. For any $i = 1, 2$, and for any $x > 0$,

$$\mathbb{P}\left(|D_i(t)| \geq x\right) \leq 2\exp\left(-\frac{x^2}{8c^2 t}\right).$$

From (1.29) and (1.30),

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau) + \lambda_t I\right) \leq \phi^2 t\right) &\leq \mathbb{P}\left(\|S_t\|_F \geq \lambda_t\right) \\
&= \mathbb{P}\left(\sqrt{D_1(t)^2 + D_2(t)^2} \geq \lambda_t\right) \\
&\leq \mathbb{P}\left(|D_1(t)| + |D_2(t)| \geq \lambda_t\right) \\
&\leq \mathbb{P}\left(|D_1(t)| \geq \frac{\lambda_t}{2}\right) + \mathbb{P}\left(|D_2(t)| \geq \frac{\lambda_t}{2}\right) \\
&\leq 4\exp\left(-\frac{\lambda_t^2}{32c^2 t}\right).
\end{aligned}$$

Thus, for any $\delta \in (0, 1)$, if $\lambda_t \geq 4\sqrt{2}c\sqrt{t}\sqrt{\log\frac{4t^2}{\delta}}$, then with probability at least $1 - \delta$,

$$\lambda_{\min}\left(\sum_{\tau=1}^{t} P(\tau) + \lambda_t I\right) > \phi^2 t.$$

□

1.7.4 Implementation Details

1.7.4.1 Efficient Calculation of the Sampling Probability

In the proposed algorithm, we use quasi-Monte Carlo estimation to calculate the sampling probability, $\tilde{\pi}_i(t)$. At round t, for each $i = 1, \ldots, N$, define

$Z_i = \frac{X_{i,t}^T(\tilde{\beta}_i(t)-\widehat{\beta}_{t-1})}{v\|X_{i,t}\|_{V_t^{-1}}}$. Then, $Z_1, \ldots, Z_N$ are IID standard Gaussian random variables. For each $i = 1, \ldots, N$,

$$\begin{aligned}\tilde{\pi}_i(t) =& \mathbb{P}\left(X_{i,t}^T\tilde{\beta}_i(t) \geq X_{j,t}^T\tilde{\beta}_j(t), \forall j \neq i \middle| \mathcal{H}_t\right) \\ =& \mathbb{P}\left(\frac{\|X_{i,t}\|_{V_t^{-1}}}{\|X_{j,t}\|_{V_t^{-1}}} Z_i \geq Z_j + \frac{(X_{j,t}-X_{i,t})^T\widehat{\beta}_{t-1}}{v\|X_{j,t}\|_{V_t^{-1}}}, \forall j \neq i \middle| \mathcal{H}_t\right)\end{aligned}$$

let f and F be the density and the distribution function of the standard Gaussian random variables, respectively. Since Z_i, and $\{Z_j\}_{j\neq i}$ are independent, the selection probability can be written as,

$$\tilde{\pi}_i(t) = \int \prod_{j\neq i} F\left(\frac{\|X_{i,t}\|_{V_t^{-1}}}{\|X_{j,t}\|_{V_t^{-1}}} z + \frac{(X_{i,t}-X_{j,t})^T\widehat{\beta}_{t-1}}{v\|X_{j,t}\|_{V_t^{-1}}}\right) f(z)dz.$$

This can be estimated by,

$$\frac{1}{M}\sum_{m=1}^{M} F \prod_{j\neq i}\left(\frac{\|X_{i,t}\|_{V_t^{-1}}}{\|X_{j,t}\|_{V_t^{-1}}} Z^{(m)} + \frac{(X_{i,t}-X_{j,t})^T\widehat{\beta}_{t-1}}{v\|X_{j,t}\|_{V_t^{-1}}}\right), \tag{1.31}$$

where $Z^{(m)}$ is the standard Gaussian random variables.

In this way, we can compute $\tilde{\pi}_i(t)$ without sampling $\tilde{\beta}_i(t)$ $M \times N$ times from $N(\widehat{\beta}_{t-1}, v_t I)$. The error of the quasi Monte Carlo method is bounded by $O\left(\frac{(\log M)^s}{M}\right)$, where s is the dimension of the domain of function to integrate. If we sample $\tilde{\beta}_i(t)$ $M \times N$ times, it gives $O\left(\frac{(\log M)^{N-1}}{M}\right)$ error. In contrast, using (1.31) reduces the error to $O\left(\frac{\log M}{M}\right)$.

In simulation studies, we use `sobol_seq` module in Python 3 to generate the quasi-Monte Carlo samples. The number of samples is $M = 200$ in `BLTS` and `DRTS`. We plot the estimator of $\tilde{\pi}_i(t)$ using $m = 1, \ldots, 200$ quasi-Monte Carlo samples, and observe that it converges within the small errors.

1.7.5 A Review of Approaches to Missing Data and Doubly-robust Method

In this section, we review approaches to missing data and the doubly-robust method used in the proposed method. First, we provide the approaches from a purely missing data point of view and how the doubly-robust method is motivated. In the second section, we show the procedures applying the doubly-robust method in bandit settings.

1.7.5.1 Doubly-robust Method in Missing Data

There are two main approaches to missing data: imputation and inverse probability weighting (IPW). Imputation is to fill in the predicted value of missing data from a specified model, and IPW is to use the observed records only but weight them by the inverse of the observation probability. The doubly-robust method can be viewed as a combination of the two.

For illustrative purposes, consider the problem of estimating the marginal mean of $Y \in \mathbb{R}$, $\mathbb{E}(Y) =: \mu$. Denoting $(Y_i - \mu)$ by $U_i(\mu)$, when all data are observed,

$$U(\mu) = \sum_{i=1}^{n} U_i(\mu) = \sum_{i=1}^{n} (Y_i - \mu) = 0,$$

gives an unbiased estimator of μ, $\sum_{i=1}^{n} Y_i/n$, and $U(\mu)$ is called an unbiased estimating function since $\mathbb{E}[U(\mu)] = 0$. Let δ_i be the observation indicator which takes value 1 if Y_i is observed, 0, otherwise. Suppose there are auxiliary variables, $X_i \in \mathbb{R}^d$, and X_i's are observed for all i. Also denote the probability of observation by $P(\delta_i = 1|X_i) =: \pi_i$. We assume $P(\delta_i = 1|Y_i, X_i) = P(\delta_i = 1|X_i)$, that is, the observation indicator is independent of Y_i. This is called *missing at random* mechanism. This assumption is required for the doubly robust method to be valid. Using the observed values only, the estimating

equation for the observed data

$$U_o(\mu) = \sum_{i=1}^{n} \delta_i U_i(\mu) = \sum_{i=1}^{n} \delta_i (Y_i - \mu) = 0,$$

gives $\frac{\sum_{i=1}^{n} \delta_i Y_i}{\sum_{i=1}^{n} \delta_i}$ as an estimator for μ. This estimator may be biased since $\mathbb{E}U_o(\mu) \neq 0$.

The two main approaches modify the observed estimating function employing two new quantities, $\mathbb{E}(Y_i|X_i)$ and π_i. These two quantities are usually unknown and we need to specify models. Therefore the two approaches require assumptions for auxiliary models: the imputation model, $\mathbb{E}(Y_i|X_i;\beta)$, and the model for observation probability, $\pi_i(\phi)$. The validity of each approach depends on the correct specification of the auxiliary model assumptions. The qualifier 'auxiliary' comes from the fact that these models are not needed when there is no missing data. In IPW, one constructs an unbiased estimating equation by amplifying the observed record according to the inverse of the observation probability as follows:

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\phi)} U_i(\mu) = \sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\phi)} (Y_i - \mu).$$

If $\pi(\phi)$ is correctly specified, i.e., $\pi = \pi(\phi)$, $\mathbb{E}(\sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\phi)} U_i(\mu)) = 0$, hence the resulting IPW estimator is valid. In the imputation method, we replace missing Y_i with $\mathbb{E}(Y_i|X_i;\beta)$ and the estimator is the solution of $U^{IMP}(\mu,\beta) = 0$ where

$$\begin{aligned} U^{IMP}(\mu,\beta) &= \sum_{i=1}^{n} \left[\delta_i U_i(\mu) + (1-\delta_i)\mathbb{E}(U_i(\mu)|X_i;\beta)\right] \\ &= \sum_{i=1}^{n} \left[\mathbb{E}(Y_i|X_i;\beta) + \delta_i\{Y_i - \mathbb{E}(Y_i|X_i;\beta)\} - \mu\right]. \end{aligned}$$

The doubly robust (DR) method [Robins et al., 1994, Bang and Robins, 2005] was initially motivated by attempting to improve the efficiency of the

IPW method. Note that we can construct an auxiliary unbiased estimating function $(\frac{\delta_i}{\pi_i(\phi)} - 1)$. Geometrically we can reduce the norm of the estimating function $\frac{\delta_i}{\pi_i(\phi)} U_i(\mu)$ by subtracting the projection on to the nuisance tangent space formed from $(\frac{\delta_i}{\pi_i(\phi)} - 1)$. The nuisance tangent space is the closed linear span of $B(\frac{\delta_i}{\pi_i(\phi)} - 1)$ for some $B \in \mathbb{R}^d$, and the projection onto the nuisance tangent space is

$$\sum_{i=1}^{n} \frac{\delta_i - \pi_i(\phi)}{\pi_i(\phi)} \mathbb{E}(U_i | X_i; \beta).$$

After subtraction, the DR estimating function has a form

$$\begin{aligned} U^{DR}(\mu, \beta, \phi) &= \sum_{i=1}^{n} \left[\frac{\delta_i}{\pi_i(\phi)} U_i(\mu) + (1 - \frac{\delta_i}{\pi_i(\phi)}) \mathbb{E}(U_i | X_i; \beta) \right] \\ &= \sum_{i=1}^{n} \left[\mathbb{E}(U_i | X_i; \beta) + \frac{\delta_i}{\pi_i(\phi)} \{ U_i(\mu) - \mathbb{E}(U_i(\mu) | X_i; \beta) \} \right]. \end{aligned}$$

Note that when you replace δ_i in $U^{IMP}(\mu)$ with $\frac{\delta_i}{\pi_i(\phi)}$, you obtain $U^{DR}(\mu)$. The DR method requires both auxiliary models. However, its validity is guaranteed when *either* of the models is correct. To verify, if the imputation model is correctly specified, i.e., $\mathbb{E}[U_i(\mu) - \mathbb{E}(U_i(\mu) | X_i; \beta) | X_i] = 0$, we have

$$\begin{aligned} \mathbb{E}\{U^{DR}(\mu, \beta, \phi)\} =& \mathbb{E} \sum_{i=1}^{n} \left[\mathbb{E}(U_i | X_i) - \frac{\delta_i}{\pi_i(\phi)} \{ U_i(\mu) - \mathbb{E}(U_i(\mu) | X_i) \} \right] \\ =& \sum_{i=1}^{n} \mathbb{E}\mathbb{E}(U_i | X_i) = 0, \end{aligned}$$

even if the π model is misspecified, i.e., $\pi_i(\phi) \neq \pi_i$. If the observation model

is correctly specified, $\pi_i(\phi) = \pi_i$, then $\mathbb{E}(1 - \frac{\delta_i}{\pi_i}|X_i) = 0$, and

$$\begin{aligned}\mathbb{E}\{U^{DR}(\mu, \beta, \phi)\} &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\delta_i}{\pi_i} U_i(\mu) + \left\{(1 - \frac{\delta_i}{\pi_i})\mathbb{E}(U_i|X_i; \beta)\right\}\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\delta_i}{\pi_i} U_i(\mu)\right] = 0,\end{aligned}$$

even if the imputation model is misspecified, i.e., $\mathbb{E}[U_i(\mu)|X_i] \neq \mathbb{E}[U_i(\mu)|X_i; \beta)]$. Therefore when *either* of the models is correct, $U^{DR}(\mu)$ is unbiased and with other technical conditions, the estimator can be shown to be consistent. That is why the qualifier *doubly robust* is adopted. The construction of the DR estimating function is possible because we have two unbiased estimating functions.

1.7.5.2 Application to Bandit Settings

In bandit settings, the missingness is controlled since the learner selects the arm. Therefore, the probability of observation or selection is known and the DR estimator is guaranteed to be valid although the imputation model for missing reward is incorrectly specified. The merit of the DR estimator in the bandit setting is that we can utilize the observed contexts from selected or unselected arms. Below we describe the DR method in the contextual bandit setting.

Let $\pi_i(t) := \mathbb{P}(a_t = i|\mathcal{H}_t)$ be the probability of selecting arm i at round t. As defined in the manuscript, the DR pseudo-reward is

$$Y_i^{DR}(t) = \left\{1 - \frac{\mathbb{I}\,(i = a_t)}{\pi_{i,t}}\right\} X_{i,t}^T \breve{\beta}_t + \frac{\mathbb{I}\,(i = a_t)}{\pi_{i,t}} Y_{a_t,t}, \tag{1.32}$$

for some $\breve{\beta}_t$ depending on $\mathcal{H}_t$. The pseudo-reward (1.32) comes from the following procedures. First we construct an unbiased estimating function also

known as the IPW score,

$$\sum_{\tau=1}^{t}\sum_{i=1}^{N}\frac{\mathbb{I}\left(i=a_\tau\right)}{\pi_{i,\tau}}X_{i,\tau}\left(Y_{i,\tau}-X_{i,\tau}^T\beta\right), \tag{1.33}$$

where only the pairs $(X_{i,t}, Y_{i,t})$ from the selected arms are contributed according the weight of the inverse of $\pi_{i,t}$. Setting this score equal to 0 and solving β gives the estimator used in Dimakopoulou et al. [2019]. Now we can subtract the projection on the nuisance tangent space from (1.33). The nuisance tangent space is the closed linear span of $B(\frac{\mathbb{I}(i=a(t)}{\pi_i(t)}-1)$ for some $B\in\mathbb{R}^d$, and the projection onto the nuisance tangent space is

$$\sum_{\tau=1}^{t}\sum_{i=1}^{N}\frac{\mathbb{I}\left(i=a_\tau\right)-\pi_{i,\tau}}{\pi_{i,\tau}}X_{i,\tau}\left(E(Y_i(\tau)|\mathcal{H}_\tau)-X_{i,\tau}^T\beta\right).$$

When the projection is subtracted from the (1.33) after replacing $E(Y_i(t)|\mathcal{H}_t)$ with $X_{i,t}^T\breve{\beta}_t$, the IPW score becomes the efficient score,

$$\sum_{\tau=1}^{t}\sum_{i=1}^{N}X_{i,\tau}\left(Y_i^{DR}(\tau)-X_{i,\tau}^T\beta\right). \tag{1.34}$$

Any $\breve{\beta}_t$ that depends on $\mathcal{H}_t$ serves the purpose of imputation. Due to the doubly robustness property, $X_{i,t}^T\breve{\beta}_t$ does not have to be an unbiased estimator of $E(Y_i(t)|\mathcal{H}_t)$. We recommend setting $\breve{\beta}_t$ as the ridge regression estimator based on the selected arms only. The expression (1.34) resembles the score when the rewards for all arms were observed, if $Y_{i,t}$ is replaced with $Y_i^{DR}(t)$.

The proposed estimator $\widehat{\beta}_t$ is a solution of (1.34) with a regularization parameter λ_t:

$$\widehat{\beta}_t=\left(\sum_{\tau=1}^{t}\sum_{i=1}^{N}X_{i,\tau}X_{i,\tau}^T+\lambda_t I\right)^{-1}\left(\sum_{\tau=1}^{t}\sum_{i=1}^{N}X_{i,\tau}Y_i^{DR}(\tau)\right).$$

Harnessing the pseudo-rewards defined in (1.32), we can make use of all con-

texts rather than just selected contexts. The use of all contexts instead of $X_{a_t,t}$ induces the improvement in the regret bound of the proposed algorithm. Kim and Paik [2019] also suggests DR estimator, but it uses Lasso estimator from the following pseudo-reward

$$Y_i^{DR}(t) = \bar{X}(t)^T \hat{\beta}(t-1) + \frac{1}{N} \frac{Y_{a(t)}(t) - b_{a(t)}(t)^T \hat{\beta}(t-1)}{\pi_{a(t)}(t)},$$

where $\bar{X}(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t)$. This estimator is of an aggregated form. As described in the text, the estimator using the aggregated pseudo-reward does not permit the regret decomposition as equation (1.6) in the thesis.

Chapter 2

Near-optimal Algorithm for Linear Contextual Bandits with Compounding Estimator

2.1 Introduction

The multi-armed bandit (MAB) is a sequential decision making problem where a learner repeatedly chooses an arm and receives a reward as partial feedback associated with the only selected arm. The goal of the learner is to maximize cumulative rewards over a horizon of length T by suitably balancing exploitation and exploration. The *Linear contextual bandit* is a general version of the MAB problem, where d-dimensional context vectors are given for each of the arms and the expected rewards for each arm is a linear function of the corresponding context vector.

There are a family of algorithms that utilize the principle of *optimism in the face of uncertainty* (OFU) [Lai and Robbins, 1985]. These algorithms for the linear contextual bandit have been widely used in practice (e.g., news recommendation in Li et al. [2010]) and extensively analyzed [Auer, 2002, Dani

et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Chu et al., 2011, Abbasi-Yadkori et al., 2011]. Some of the most widely used algorithms in this family are `LinUCB` Li et al. [2010] and `OFUL` Abbasi-Yadkori et al. [2011] due to their practicality and performance guarantees. The best known regret bound for these algorithms is $\tilde{O}(d\sqrt{T})$, where $\tilde{O}$ stands for big-O notation up to logarithmic factors.

The improved regret bound of $\tilde{O}(\sqrt{dT})$ has been shown for `SupLinUCB` Chu et al. [2011] with a matching lower bound $\Omega(\sqrt{dT})$, hence provably optimal up to logarithmic factors. `SupLinUCB` and its variants (e.g., Li et al., 2017) improved the regret bound by $\sqrt{d}$ factor exploiting independence of samples via a phased bandit technique proposed by Auer [2002]. Despite their provable optimality, `SupLinUCB` and other algorithms based on the framework of Auer [2002] have been known to be impractical due to the lack of adaptiveness, resulting in performing excessive random sampling, and computational inefficiency. Furthermore, the regret bound of `SupLinUCB` has $(\log N)^{3/2}$ dependence, where N is the total number of arms. Therefore, if N is exponentially large in d (which often arises in practice, e.g., large-scale recommender systems), then the regret bound would be sub-optimal. Hence, whether N-independent $\tilde{O}(\sqrt{dT})$ regret is achievable has been an open problem. Moreover, the question of whether $\tilde{O}(\sqrt{dT})$ regret is attainable by a more practical algorithm than the algorithms based on the framework of Auer [2002] has remained open.

A tighter upper bound of `SupLinUCB` than that of `LinUCB` (and `OFUL`) stems from utilizing phases by handling computation separately for each phase. In phased algorithms, the arms in the same phase are chosen without making use of the rewards in the same phase. This independence of samples allows to apply a tight confidence bound, improving the regret bound by $\sqrt{d}$ factor. On the other hand, this operation should be handled for each arm, which costs polylogarithmic dependence on N by invoking the union bound over the arms at the expense of improving $\sqrt{d}$. In UCB algorithms, the estimate is adaptive

in a sense that the update is made in every round, and the independence argument cannot be utilized. Instead, self-normalized theorem Abbasi-Yadkori et al. [2011] helps avoid the dependence on N.

We propose a novel bandit algorithm. The proposed algorithm achieves $\tilde{O}(\sqrt{dT})$ as in `SupLinUCB`, yet without resorting to independence and without dependence on N. The proposed algorithm has two notable features: the first is to utilize the contexts of all arms both selected and unselected for parameter estimation, and the second is to randomly perturb the contribution to the estimator. Intuitively, new randomization on the estimator elevates the level of exploration, but more importantly, this randomness creates nuisance tangent space (See Section 2.4.1) essential to form the compounding estimator that uses all contexts. These two features allow a novel additive decomposition of the regret which can be bounded using the self-normalized norm of the compounding estimator.

The main contributions are as follows:

- We propose a novel algorithm, Hybridization by Randomization algorithm (`HyRan`), for a linear contextual bandit. The proposed algorithm adopts the compounding estimator utilizing contexts from all arms both selected and unselected and the random perturbation of the principle of *optimism in the face of uncertainty* for arm selection.

- We establish that `HyRan` achieves $\tilde{O}(\sqrt{dT})$ regret upper bound without any dependence on N. To the best of my knowledge, this is the first N-independent $\tilde{O}(\sqrt{dT})$ regret bound for the linear contextual bandit. For the analysis, we utilize a novel decomposition of the cumulative regret into two main additive terms whose bounds can be derived by employing the structure of the compound estimator. This allows us to establish the faster rate of $\tilde{O}(\sqrt{dT})$ regret without incurring dependence on N.

- We show the estimation error bound for the self-normalized compound

estimator (Theorem 1.3), which may be of independent interest.

- We evaluate `HyRan` on numerical experiments and show that the practical performance of the proposed algorithm is in line with the theoretical guarantees.

All missing proofs are in supplementary materials.

2.2 Related Works

The linear contextual bandit problem was first introduced by Abe and Long [1999]. UCB algorithms for the linear contextual bandit have been proposed and analyzed by [Auer, 2002, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Chu et al., 2011, Abbasi-Yadkori et al., 2011] and their follow-up works. To our knowledge, the best regret upper-bound is $\tilde{O}(\sqrt{dT})$ established for `SupLinUCB`, an UCB-based algorithm proposed by Chu et al. [2011] adapting the IID sample generation technique in Auer [2002].

The rewards for the unselected arms are not observed, hence, missing. Recently some bandit literature has framed the bandit setting as a missing data problem, and employed missing data methodologies [Dimakopoulou et al., 2019, Kim and Paik, 2019, Kim et al., 2021]. Dimakopoulou et al. [2019] employs an *inverse probability weighting* (IPW) estimator using the selected contexts alone and proves a $\tilde{O}(d\sqrt{\epsilon^{-1}T^{1+\epsilon}N})$ regret bound for Thompson sampling which depends on the number of arms, N. The *doubly robust* (DR) method [Robins et al., 1994, Bang and Robins, 2005] is adopted in Kim and Paik [2019] with Lasso penalty for high-dimensional settings with sparsity and the regret bound is shown to be improved in terms of the sparse dimension instead of d. Recently in Kim et al. [2021], a modified Thompson Sampling employing the DR method is proposed and provided $\tilde{O}(d\sqrt{T})$ bound. The authors improve the bound by using contexts of all arms including the unselected

ones which paves a way to circumvent the technical definition of unsaturated arms.However, using contexts of all arms by the DR method requires non-zero probability of selection for all arms, which limits the application to Thompson Sampling with Gaussian prior. In this thesis, we provide another way of using contexts of all arms which can be applied to the algorithm where there is an arm whose selection probability is zero. By using all contexts we circumvent the inequalities which induces additional $\sqrt{d}$ or $\log N$ by developing a novel decomposition of the regret.

2.3 Linear Contextual Bandit Problem

In each round $t \in [T] := \{1, \ldots, T\}$, the learner observes a set of arms $[N] := \{1, ..., N\}$ with their corresponding context vectors $\{X_{i,t} \in \mathbb{R}^d \mid i \in [N]\}$. Then, the learner chooses an arm $a_t \in [N]$ and receives a random reward $Y_{t,:} = Y_{a_t,t}$ for the chosen arm. For all $t \in [T]$ and $i \in [N]$, we assume the linear reward model, i.e.,

$$Y_{i,t} = X_{i,t}^T \beta + \eta_{i,t},$$

where $\beta \in \mathbb{R}^d$ is an *unknown* parameter and $\eta_{i,t} \in \mathbb{R}$ is an independent noise. Let $\mathcal{H}_t$ be the history at round t that contains contexts $\{X_{i,\tau}\}_{i=1,\tau=1}^{N,t}$, chosen arms $\{a_\tau\}_{\tau=1}^{t-1}$ and the corresponding rewards $\{Y_{\tau,a_\tau}\}_{\tau=1}^{t-1}$. For each t and i, the noise $\eta_{i,t}$ is zero-mean conditioned on $\mathcal{H}_t$, i.e, $\mathbb{E}\left[\eta_{i,t} | \mathcal{H}_t\right] = 0$. The optimal arm at round t is defined as $a_t^* := \arg\max_{i \in [N]} \left\{X_{i,t}^T \beta\right\}$. Let $\texttt{regret}(t)$ be the difference between the expected rewards of the chosen arm and the optimal arm at round t.

$$\texttt{regret}(t) := X_{a_t^*,t}^T \beta - X_{a_t,t}^T \beta.$$

The goal is to minimize the sum of regrets over T rounds, $R(T) := \sum_{t=1}^T \texttt{regret}(t)$. The time horizon T is finite but possibly unknown.

2.4 Proposed methods

2.4.1 Compounding Estimator

We introduce the novel estimator with simple randomization technique. For each round $t \in [T]$ and given $p \in (0,1)$, we define $\tilde{a}_t$ as a random variable sampled from $[N]$ with probability

$$\begin{aligned} \pi_{a_t,t} &:= \mathbb{P}\left(\tilde{a}_t = a_t \middle| \mathcal{F}_t\right) = p, \\ \pi_{j,t} &:= \mathbb{P}\left(\tilde{a}_t = j \middle| \mathcal{F}_t\right) = \frac{1-p}{N-1}, \forall j \neq a_t, \end{aligned} \tag{2.1}$$

where $\mathcal{F}_t := \mathcal{H}_t \cup \{a_t\} \cup \{\tilde{a}_1, \ldots, \tilde{a}_{t-1}\}$. Since this $\tilde{a}_t$ is random given $\mathcal{F}_t$, we can construct a random variable $\mathbb{I}\left(\tilde{a}_t = i\right)/\pi_{i,t}$ whose conditional mean given $\mathcal{F}_t$ is one. Using these random variables we can construct the inverse probability weighting estimating equation,

$$\sum_{\tau=1}^{t} \sum_{i=1}^{N} \frac{\mathbb{I}\left(\tilde{a}_\tau = i\right)}{\pi_{i,\tau}} X_{i,\tau} \left(Y_{i,t} - X_{i,\tau}^T \beta\right) = 0. \tag{2.2}$$

The semi-parametric theory Bickel et al. [1993] suggests subtracting the projection onto the nuisance tangent space from (2.2) to improve efficiency. Using the fact that the conditional mean of $\left(\mathbb{I}\left(a_t = i\right) - \pi_{i,t}\right)$ given $\mathcal{F}_t$ is zero, we can define a nuisance tangent space by the closed linear span of $B(\mathbb{I}\left(\tilde{a}_t = 1\right) - \pi_{1,t}\}, \ldots, \mathbb{I}\left(\tilde{a}_t = N\right) - \pi_{N,t}\})^T$ for some $B \in \mathbb{R}^{d \times N}$. Now the projection onto the nuisance tangent space is

$$\sum_{\tau=1}^{t} \sum_{i=1}^{N} \frac{\mathbb{I}\left(\tilde{a}_t = i\right) - \pi_{i,\tau}}{\pi_{i,\tau}} X_{i,\tau} \left(\mathbb{E}\left[Y_{i,\tau} \middle| \mathcal{F}_\tau\right] - X_{i,\tau}^T \beta\right),$$

where we replace $\mathbb{E}\left[Y_{i,\tau} \middle| \mathcal{F}_\tau\right]$ with $X_{i,\tau}^T \breve{\beta}_\tau$ for some estimator $\breve{\beta}_\tau$. Subtracting the projection from the IPW score and rearranging it, we obtain the efficient

score,

$$\sum_{\tau=1}^{t}\sum_{i=1}^{N} X_{i,\tau}\left(\tilde{Y}_{i,\tau} - X_{i,\tau}^{T}\beta\right), \tag{2.3}$$

where the pseudo reward $\tilde{Y}_{i,\tau}$ is defined as

$$\tilde{Y}_{i,\tau} = \left\{1 - \frac{\mathbb{I}\left(\tilde{a}_{\tau} = i\right)}{\pi_{i,\tau}}\right\} X_{i,\tau}^{T}\breve{\beta}_{\tau} + \frac{\mathbb{I}\left(\tilde{a}_{\tau} = i\right)}{\pi_{i,\tau}} Y_{\tilde{a}_t,t}. \tag{2.4}$$

The equation (2.3) uses $\tilde{Y}_{i,\tau}$ instead of $Y_{i,\tau}$ in the original score to estimate β as if all rewards were observed. Using the pseudo rewards (2.4), we can use all contexts rather than just selected contexts. However, we cannot compute (2.4) since $Y_{\tilde{a}_\tau,\tau}$ is missing when $\tilde{a}_\tau \neq a_t$. To handle this problem, we use the efficient score (2.3) not in all rounds $[t]$ but in some subset of rounds $\Psi_t \subseteq [t]$, where $Y_{\tilde{a}_\tau,\tau}$ is observed and (2.4) is computable. With this subsample set of rounds Ψ_t we can define the compound score equation

$$\begin{aligned}\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\left(\tilde{Y}_{i,\tau} - X_{i,\tau}^{T}\beta\right) \\ + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}\left(Y_{a_\tau,\tau} - X_{a_\tau,\tau}^{T}\beta\right) = 0.\end{aligned} \tag{2.5}$$

The proposed estimator is the solution of (2.5) which can be written as

$$\begin{aligned}\widehat{\beta}_t := &\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^{T} + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}X_{a_\tau,\tau}^{T} + \lambda_t I\right)^{-1} \\ &\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\tilde{Y}_{i,\tau} + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}Y_{\tau,a_\tau}\right).\end{aligned} \tag{2.6}$$

This is a hybrid form of using all contexts and using selected contexts, and the contribution is set by the random variable $\tilde{a}_\tau$. This contribution of using contexts of all arms is crucial in achieving the regret bound of $\tilde{O}(\sqrt{dT})$ for the

proposed algorithm.

2.4.2 HyRan Algorithm

The proposed algorithm, `HyRan`, is presented in Algorithm 2.1. At each round t, the algorithm computes $X_{i,t}^T \widehat{\beta}_{t-1}$ for each arm $i \in [N]$ based on (1.2) and finds the arm with the maximum value, a_t. After pulling an arm a_t and observing the reward for the selected arm, `HyRan` samples $\tilde{a}_t$ and determine whether the round t is included in the subset Ψ_t. When $\tilde{a}_t$ is equal to a_t, we can observe the reward $Y_{\tilde{a}_t,t}$ and compute (2.4). Therefore we include the round t in Ψ_t. Using this subset Ψ_t, `HyRan` updates $\widehat{\beta}_t$ as in (1.2). In order to compute $\widehat{\beta}_t$, the imputation estimator $\breve{\beta}_t$ needs to be specified to determine the pseudo reward in (2.4). We recommend the following form

$$\begin{aligned}\breve{\beta}_t := &\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}X_{a_\tau,\tau}^T + \gamma_t I\right)^{-1} \\ &\left\{\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\left(\left\{1-\frac{\mathbb{I}\left(\tilde{a}_\tau = i\right)}{\pi_{i,\tau}}\right\} X_{i,\tau}^T \widehat{\beta}_{t-1}^{ridge} + \frac{\mathbb{I}\left(\tilde{a}_\tau = i\right)}{\pi_{i,\tau}} Y_{\tilde{a}_t,t}\right)\right. \\ &\left. + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau} Y_{\tau,a_\tau}\right\}, \end{aligned} \tag{2.7}$$

for some $\gamma_t > 0$ and $\widehat{\beta}_t^{ridge}$ is a ridge estimator using pairs of selected contexts and corresponding rewards until round t. We can also use another estimator such that $\left\|\breve{\beta}_t - \beta\right\|_2 \leq N^{-1}$ holds after some explorations. Examples are the ridge estimator $\widehat{\beta}_{t-1}^{ridge}$, and the estimator used in round $t-1$, $\widehat{\beta}_{t-1}$. Since $\breve{\beta}_t$ is multiplied with mean zero random variable in (2.4) the unbiasedness of the estimator (1.2) does not depend on the choice of $\breve{\beta}_t$.

Algorithm 2.1 Hybridization by Randomization Algorithm for Linear Contextual Bandits (`HyRan`)

INPUT: A regularization parameter $\lambda_t > 0$, a sub-sampling parameter $p \in (0, 1)$.
Initialize $V_0 = I_d$, $f_0 = \mathbf{0}$.
for $t = 1$ **to** T **do**
  Observe contexts $\{X_{i,t}\}_{i=1}^N$.
  Estimate $\widehat{\beta}_{t-1} = (V_{t-1} + \lambda_t I_d)^{-1} f_{t-1}$.
  Play $a_t = \arg\max_i X_{i,t}^T \widehat{\beta}_{t-1}$, and observe reward Y_{t,a_t}.
  Set $\pi_{a_t,t} := p$ and $\pi_{j,t} := \frac{1-p}{N-1}$ for $j \neq a_t$.
  Sample $\tilde{a}_t$ from the categorical distribution with probability $\pi_{i,t}$.
  if $\tilde{a}_t = a_t$ **then**
    Update $V_t = V_{t-1} + \sum_{i=1}^N X_{i,t} X_{i,t}^T$
    Update $\breve{\beta}_t$ and $f_t = f_{t-1} + \sum_{i=1}^N X_{i,t} \tilde{Y}_{i,\tau}$
  else
    Update $V_t = V_{t-1} + X_{a_t,t} X_{a_t,t}^T$.
    Update $f_t = f_{t-1} + X_{a_t,t} Y_{a_t,t}$.
  end if
end for

2.5 Main Results

In this section, we present the main theoretical results: the regret bound for `HyRan` (Theorem 2.1) and the estimation error bound of the proposed compounding estimator (Theorem 2.4). We first provide the assumptions used throughout the analysis.

Assumption 1 [Boundedness] For all $i \in [N]$ and $t \in [T]$, $\|X_{i,t}\|_2 \leq 1$ and $\|\beta\|_2 \leq 1$.

Assumption 2 [Sub-Gaussian noise] For each t and i, the noise $\eta_{i,t}$ is conditionally σ-sub-Gaussian for a fixed constant $\sigma \geq 0$, i.e, $\mathbb{E}[\eta_{i,t} | \mathcal{H}_t] = 0$ and $\mathbb{E}[\exp(\lambda \eta_{i,t}) | \mathcal{H}_t] \leq \exp(\lambda^2 \sigma^2 / 2)$, for all $\lambda \in \mathbb{R}$.

Assumption 3 [Context sets IID across rounds] The context sets

$$\{X_{i,1}\}_{i=1}^N, \{X_{i,2}\}_{i=1}^N, \ldots, \{X_{i,T}\}_{i=1}^N$$

are distributed independently from some unknown distribution P_X supported on $\mathbb{R}^{d\times N}$.

Assumption 4 [Positive definiteness of the covariance of the contexts] For all $t \in [T]$, there exists a constant $\phi^2 > 0$ such that

$$\lambda_{\min}\left(\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_{i,t}X_{i,t}\right]\right) \geq \phi^2$$

.

Discussion of the assumptions. Assumptions 1 and 2 are standard in the stochastic contextual bandit literature (see e.g. Agrawal and Goyal [2013]). As for Assumptions 3, we emphasize that the IID assumption is on a context set across the time horizon, not on the individual context vectors. Hence, we allow context vectors to be correlated in a given round. Similar IID assumptions on context sets are used in the contextual bandit literature [Goldenshluger and Zeevi, 2013, Li et al., 2017, Kim and Paik, 2019, Bastani and Bayati, 2020]. Assumption 4 is essential to efficiently solve the linear regression problem. Previous literature imposes this assumption on the eigenvalue of the covariance matrix [Goldenshluger and Zeevi, 2013, Li et al., 2017].

2.5.1 Regret Bound of `HyRan`

Under the assumptions above, we present the following regret bound for the `HyRan` algorithm.

Theorem 2.1. *Suppose Assumptions 1-4 hold and*

$$T \geq \mathcal{E} = \max\left\{\frac{1}{p}\log\frac{T}{\delta}, C_{p,\sigma}N^2\phi^{-4}\log\frac{2T}{\delta}\right\},$$

where $C_{p,\sigma}$ is a constant depending only on p and σ. Set $\lambda_t := d\log\frac{4t^2}{\delta}$. Then

the total regret by time T for ***HyRan*** *is bounded by*

$$R(T) \leq 2\mathcal{E} + 4D_{p,\sigma}\sqrt{2T\log\frac{1}{\delta}} + 3\delta D_{p,\sigma} + \frac{\left(16\sqrt{2}+8\right)D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}}, \quad (2.8)$$

with probability at least $1-8\delta$*, where* $D_{p,\sigma} := 1 + C_{p,\sigma}$*.*

There are other works [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Lattimore and Szepesvári, 2020] which proves the lower bound of $\Omega(d\sqrt{T})$. However, those bounds assumes that N is infinity and the contexts does not follows the Assumption 3 or Assumption 4. In Section 2.5.3 we provide the lower bound under the Assumptions 1-4.

2.5.2 Regret Decomposition

In the analysis of `LinUCB` and `OFUL`, an instantaneous regret is controlled by using

$$a_t, \widehat{\beta}_{\text{ucb}} = \arg\max_{i\in[N],\widehat{\beta}\in\mathcal{C}_t} X_{i,t}^T\widehat{\beta}$$

where $\mathcal{C}_t$ is a high-probability confidence ellipsoid. Then, $\texttt{regret}(t)$ is typically decomposed as

$$\texttt{regret}(t) \leq \left\|\widehat{\beta}_{ucb} - \beta\right\|_{Z_t} \|X_{a_t,t}\|_{Z_t^{-1}}, \quad (2.9)$$

where $Z_t := \sum_{\tau=1}^{t} X_{a_\tau,\tau}X_{a_\tau,\tau}^T + \lambda I$. Each of two terms on the right hand side in (2.9) has a $\sqrt{d}$ factor. In particular, $\sqrt{d}$ factor in the first term comes from the radius of $\mathcal{C}_t$. Hence, this results in $O(d)$ regret when combined.

In contrast, we decompose the regret into additive terms using the definition of max-residual given in Lemma 2.2. This new decomposition allows for non-OFU based analysis, hence exploration parameter γ need not be the radius of $\mathcal{C}_t$.

Lemma 2.2. *Define a max-residual function as*

$$\Delta_{\widehat{\beta}}(x) := \max_{i\in[N]} \left| x_i^T \left(\widehat{\beta} - \beta \right) \right|, \tag{2.10}$$

where $x = (x_1, \ldots, x_N) \in \mathbb{R}^{d\times N}$. *For each* $t \in [T]$, *let* $\mathcal{X}_t := (X_{1,t}, \ldots, X_{N,t})$ *and define a filtration* $\mathcal{G}_t := \bigcup_{\tau=1}^{t} \left\{ \mathcal{X}_\tau, \widehat{\beta}_\tau \right\}$. *Then for* $t \geq 1$,

$$\begin{aligned} \texttt{regret}(t+1) &\leq 2\left\{ \Delta_{\widehat{\beta}_t}(\mathcal{X}_{t+1}) - \mathbb{E}\left[\Delta_{\widehat{\beta}_t}(\mathcal{X}_{t+1}) \middle| \mathcal{G}_t \right] \right\} \\ &+ 2\left\{ \mathbb{E}\left[\Delta_{\widehat{\beta}_t}(\mathcal{X}_{t+1}) \middle| \mathcal{G}_t \right] - \frac{1}{|\Psi_t|} \sum_{\tau\in\Psi_t} \Delta_{\widehat{\beta}_t}(\mathcal{X}_\tau) \right\} \\ &+ \frac{2}{\sqrt{|\Psi_t|}} \left\| \widehat{\beta}_t - \beta \right\|_{V_t}, \end{aligned} \tag{2.11}$$

where

$$V_t := \sum_{\tau\in\Psi_t} \sum_{i=1}^{N} X_{i,\tau} X_{i,\tau}^T + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau} X_{a_\tau,\tau}^T + \lambda_t I.$$

A proof sketch is given below. The decomposition of the expected regret given in (2.11) is insightful in that the regret from suboptimal arm selections is incurred due to poor estimate, thus can be bounded by the quantities involving the maximum residual. Many bandit algorithms that induces $\tilde{O}(\sqrt{dT})$ regret bound (e.g. `SuplinUCB`) bounds the maximum residual with the union of $N \times T$ probability inequalities, and this gives $\log N$ term in the regret bound. But in Lemma 2.2, we use the fact that the maximum is bounded by a sum, the sum of residual can be shown to be bounded by the self-normalized bound for the estimator (1.2). In this way we can use only T probability inequalities and eliminate the N independence on the main term of the regret bound. We emphasize that the decomposition yields the self-normalized bound for the estimator, not any other estimator, and the estimator is self-normalized using all contexts of both selected and unselected arms.

Proof. By the definition of a_t, we have

$$\begin{aligned}
\texttt{regret}(t+1) &= \left(X_{a^*_{t+1},t+1} - X_{a_{t+1},t+1}\right)^T \left(\beta - \widehat{\beta}_t\right) \\
&\quad + \left(X_{a^*_{t+1},t+1} - X_{a_{t+1},t+1}\right)^T \widehat{\beta}_t \\
&\leq \left(X_{a^*_{t+1},t+1} - X_{a_{t+1},t+1}\right)^T \left(\beta - \widehat{\beta}_t\right) \\
&\leq 2 \max_{i\in[N]} \left| X_{i,t+1}^T \left(\widehat{\beta}_t - \beta\right)\right|,
\end{aligned}$$

which concludes $\texttt{regret}(t+1) \leq 2\Delta_{\widehat{\beta}_t}(\mathcal{X}_{t+1})$. Now to prove (2.11), we need

$$\frac{1}{|\Psi_t|} \sum_{\tau\in\Psi_t} \Delta_{\widehat{\beta}_t}(\mathcal{X}_\tau) \leq \frac{1}{\sqrt{|\Psi_t|}} \left\|\widehat{\beta}_t - \beta\right\|_{V_t},$$

which is proved by using the Cauchy-Schwartz inequality as

$$\begin{aligned}
\sum_{\tau\in\Psi_t} \Delta_{\widehat{\beta}_t}(\mathcal{X}_\tau) &\leq \sqrt{|\Psi_t|}\sqrt{\sum_{\tau\in\Psi_t} \left\{\Delta_{\widehat{\beta}_t}(\mathcal{X}_\tau)\right\}^2} \\
&= \sqrt{|\Psi_t|}\sqrt{\sum_{\tau\in\Psi_t} \max_{i\in[N]} \left\{X_{i,\tau}^T \left(\widehat{\beta}_t - \beta\right)\right\}^2} \\
&\leq \sqrt{|\Psi_t|}\sqrt{\sum_{\tau\in\Psi_t} \sum_{i=1}^N \left\{X_{i,\tau}^T \left(\widehat{\beta}_t - \beta\right)\right\}^2} \\
&\leq \sqrt{|\Psi_t|}\sqrt{\left(\widehat{\beta}_t - \beta\right)^T V_t \left(\widehat{\beta}_t - \beta\right)},
\end{aligned}$$

where the last inequality holds using the fact that $V_t \succeq \sum_{\tau\in\Psi_t} \sum_{i=1}^N X_{i,\tau} X_{i,\tau}^T$. □

The first term in (2.11) can be bounded using with Azuma's inequality (Lemma 1.9). Now, we bound the second and third terms in (2.11) using Lemma 2.3 and Theorem 2.4, respectively. Lemma 2.3 adopts the empirical theories on the distribution of the contexts.

Lemma 2.3. *Suppose Assumptions 1-4 hold. For each $t \in [T]$, and $L > 0$,*

conditioned on Ψ_t, with probability at least $1-\delta/T$,

$$\sup_{\|\beta_1-\beta\|\le L}\left|\mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_{t+1}\right)|\,\mathcal{G}_t\right]-\frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}\left(\mathcal{X}_\tau\right)\right|$$
$$\le\frac{3L\delta}{2T}+4L\sqrt{\frac{1}{|\Psi_t|}}\sqrt{d\log\frac{2T}{\delta}}.$$

In the following theorem, we present the self-normalized bound for the compound estimator which allows us to bound the last term in (2.11).

Theorem 2.4. *Suppose Assumptions 1-4 hold. Let $\widehat{\beta}_t$ be the estimator defined in (1.2), and let $p\in(0,1)$ be a constant used in (2.1). For all $t\in[T]$, let Ψ_t be a subset of $[t]$ determined by Algorithm 2.1. Then with probability at least $1-6\delta$,*

$$\left\|\widehat{\beta}_t-\beta\right\|_{V_t}\le\sqrt{\lambda_t}+\left(\frac{4\sqrt{2}}{1-p}+\frac{\sigma}{p}\right)\sqrt{d\log\frac{4t^2}{\delta}},\tag{2.12}$$

for all $t\ge\max\left\{\frac{1}{p}\log\frac{T}{\delta},C_{p,\sigma}N^2\phi^{-4}\log\frac{2T}{\delta}\right\}$, where $C_{p,\sigma}>0$ is a constant depending only on p and σ.

Theorem 2.4 is a self-normalized bound for the compound estimator, which is a crucial element in the regret analysis. Compared to the widely-used self-normalization bound (Theorem 2 in Abbasi-Yadkori et al. [2011]) in the contextual bandit literature, the estimation error bound (2.12) is self-normalized by the covariance matrix constructed by the contexts of all arms, not just selected contexts. This difference enables us to take advantage of the new decomposition of the regret in (2.11), which derives a $\tilde{O}(\sqrt{dT})$ regret bound.

To use the bound (2.12), we need an estimator for $\breve{\beta}_t$ whose estimation error is smaller than N^{-1}. This estimator can be obtained by using data from at least $O(N^2\phi^{-4}\log T)$ number of rounds, which is tolerable for the exploration.

The last concern regarding the regret bound is the size of Ψ_t. To obtain the $O(\sqrt{dT\log T})$ regret bound, we need to make sure that the number of rounds

in $\Psi_t = \Omega(t)$. In the following Lemma, we show that the size of the selected subset Ψ_t is $\Omega(t)$ with high probability.

Lemma 2.5. *Let Ψ_t be a subset of $[t]$ determined by the Algorithm 2.1 at round t. For any $\epsilon \in (0,1)$, with probability at least $1-\delta$,*

$$|\Psi_t| \geq \epsilon p t, \tag{2.13}$$

for all $t \geq \frac{1}{2p(1-\epsilon)} \log \frac{T}{\delta}$.

The proof is based on lower Chernoff bounds. With (2.13), we guarantee the rate of the regret bound is sub-linear with respect to the total round T.

2.5.3 A Matching Lower Bound

Theorem 2.6. *Assume $2 \leq d \leq N < \infty$ and $T \geq d/4$. Then there exists a distribution of contexts, $\mathcal{P}_X$, a distribution of noise, $\eta_{i,t}$ and β, which satisfies Assumptions 1-4 and for any bandit algorithms that selects a_t,*

$$\mathbb{E}_\beta R(T) \geq \frac{1}{8}\sqrt{dT}. \tag{2.14}$$

The lower bound (2.14) matches with that of the upper bound (2.8) up to the logarithm factor. Chu et al. [2011] prove a lower bound $\Omega(\sqrt{dT})$ for the linear contextual bandits with finite number of arms. But the contexts in the bound does not hold Assumptions 3 and 4, and cannot be directly applied. We call for proving a novel lower bound which can be applied to the setting with Assumptions 1-4.

2.6 Numerical Experiments

In this section, we compare the performances of the four linear contextual bandit algorithms: `SuplinUCB` [Chu et al., 2011], `LinUCB` [Li et al., 2010], `LinTS`

[Agrawal and Goyal, 2013], and the proposed method, `HyRan`. For simulation, the number of arms N is set to 10 or 20, and the dimension of contexts d is set to 5, 10 and 20. Let $X_{i,t}^{(1)}, \ldots, X_{i,t}^{(d)}$ be the d elements of a context $X_{i,t}$. For $j = 1, \ldots, d-1$, we independently generate $(X_{1,t}^{(j)}, \cdots, X_{N,t}^{(j)})$ from a normal distribution $\mathcal{N}(\mu_N, V_N)$ with mean $\mu_{10} = (-10, -8, \cdots, -2, 2, \cdots 8, -10)^T$, or $\mu_{20} = (-20, -18, \cdots, -2, 2, \cdots, 18, 20)^T$. To impose correlation among each arms the covariance matrix $V_N \in \mathbb{R}^{N \times N}$ is set as $V(i,i) = 1$ for every i and $V(i,k) = 0.5$ for every $i \neq k$. Then, for each arm $i \in [N]$, we select a generated element $X_{i,t}^{(j)}$ randomly and append it to the last element, i.e. $X_{i,t}^{(d)}$ is the same as one of $X_{i,t}^{(1)}, \ldots, X_{i,t}^{(d-1)}$ This setting is to impose a severe multicollinearity on each contexts. Finally, we truncated the sampled contexts to satisfy $\|X_{i,t}\|_2 \leq 1$. To generate the stochastic rewards, we sample $\eta_{i,t}$ independently from $\mathcal{N}(0,1)$. Each element of β is sampled from a uniform distribution, $\mathcal{U}(-1/\sqrt{d}, 1/\sqrt{d})$ at the beginning of each instance and stays fixed during a single instance of the experiments. About the set of hyperparameters, `LinTS` and `LinUCB` v and α $\{0.001, 0.01, 0.1, 1\}$, respectively. In `HyRan` we set $\lambda_t := d \log(t+1)^2$ to be consistent with the theoretical results and p to be in $\{0.5, 0.65, 0.8, 0.95\}$. We optimize over these hyperparameters and report the best performance for each algorithm.

Figure **??** shows the average of the cumulative regrets over the horizon length $T = 30000$ with 10 repeated experiments. The experimental results demonstrate that `HyRan` performs better than the benchmarks in all of the cases and show more evident superior performances as the context dimension increases. To our knowledge, `HyRan` is the first algorithm with $\tilde{O}(\sqrt{dT})$ regret that has competitive empirical performances. The previously known algorithms with $\tilde{O}(\sqrt{dT})$ regret (e.g., `SupLinUCB` in Chu et al. 2011 and `SupCB-GLM` in Li et al. 2017) tend to be impractical. Hence, `HyRan` is a provably efficient and practical method.

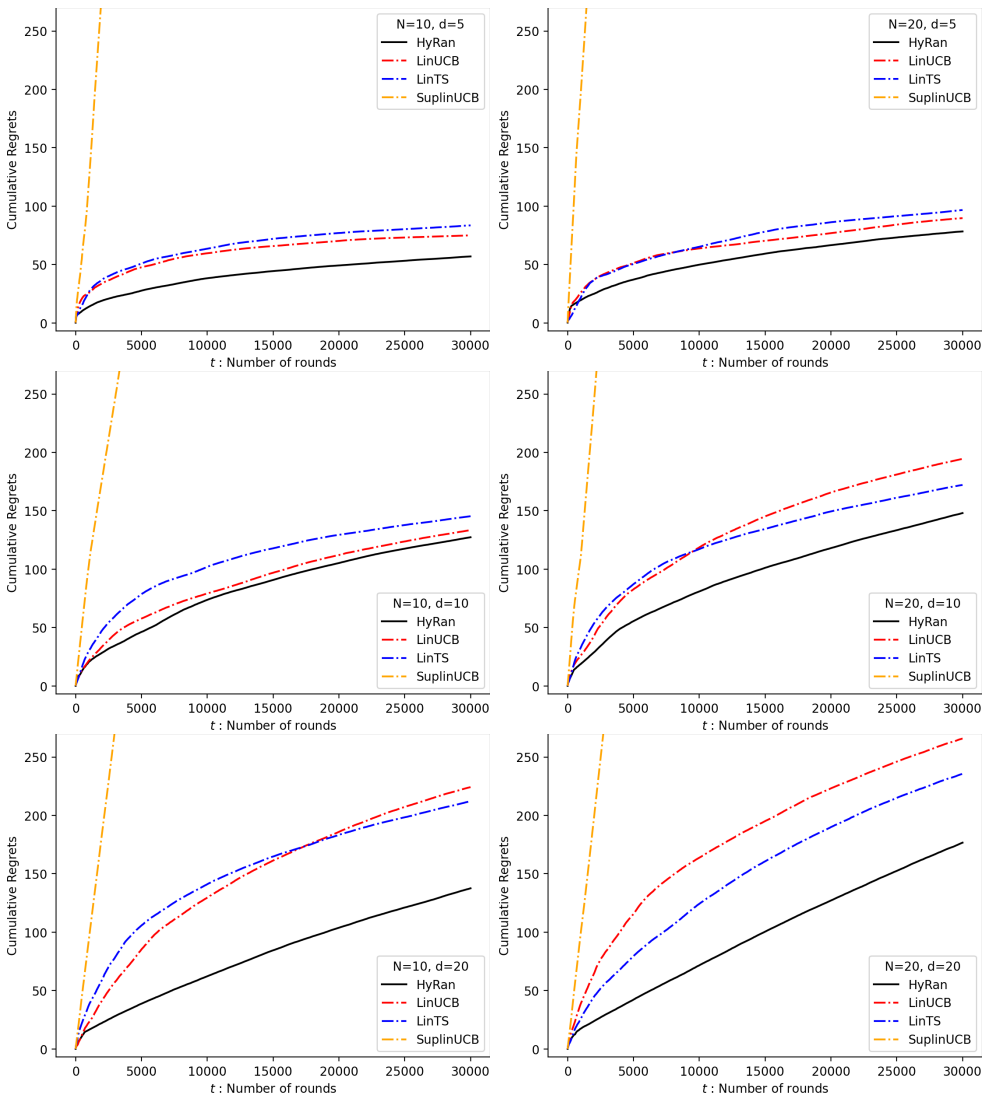

Figure 2.1: A Comparison of cumulative regrets of `SuplinUCB`, `TS`, `LinUCB` and `HyRan`. Each line shows the averaged cumulative regrets over 10 repeated experiments. The scale of the axis of cumulative regrets is fixed for comparison as d increases.

2.7 Appendix

2.7.1 Technical Lemmas

Lemma 2.7. *Lee et al. [2016, Lemma 2.3] Let $\{N_t\}$ be a martingale on a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. Then there exists a $\mathbb{R}^2$-valued martingale $\{M_t\}$ such that for any time $t \geq 0$, $\|M_t\|_2 = \|N_t\|_{\mathcal{H}}$ and $\|M_{t+1} - M_t\|_2 = \|N_{t+1} - N_t\|_{\mathcal{H}}$.*

Lemma 2.8. *(Azuma-Hoeffding) If a super-martingale $(Y_t; t \geq 0)$ corresponding to filtration $\mathcal{F}_t$, satisfies $|Y_t - Y_{t-1}| \leq c_t$ for some constant c_t, for all $t = 1, \ldots, T$, then for any $a \geq 0$,*

$$\mathbb{P}\left(Y_T - Y_0 \geq a\right) \leq e^{-\frac{a^2}{2\sum_{t=1}^{T} c_t^2}}.$$

2.7.2 Proof of Theorem 2.1

Proof. For each $t \in [T]$, define the event

$$\begin{aligned}
A_t &:= \left\{ |\Psi_t| > \frac{1}{2} pt \right\}, \\
B_t &:= \left\{ \left\| \widehat{\beta}_t - \beta \right\|_{V_t} \leq \sqrt{\lambda_t} + \left(\frac{4\sqrt{2}}{1-p} + \frac{\sigma}{p} \right) \sqrt{d \log \frac{4t^2}{\delta}} \right\}, \\
C_t &:= \left\{ \left\| \widehat{\beta}_t - \beta \right\|_2 \leq 1 + \frac{4\sqrt{2}}{1-p} + \frac{\sigma}{p} := D_{p,\sigma} \right\}.
\end{aligned}$$

Set $\mathcal{E} := \max\left\{ \frac{1}{p} \log \frac{T}{\delta}, C_{p,\sigma} N^2 \phi^{-4} \log \frac{2T}{\delta} \right\}$, where $C_{p,\sigma}$ is defined in (2.25). While proving Theorem 2.4, Lemma 2.5 and 2.9 is used and the event B_t

requires A_t. Under the event B_t, setting $\lambda_t = d \log \frac{4t^2}{\delta}$ gives

$$\begin{aligned}
\left\|\widehat{\beta}_t - \beta\right\|_2 \leq & \sqrt{\left(\widehat{\beta}_t - \beta\right)^T V_t^{\frac{1}{2}} V_t^{-1} V_t^{\frac{1}{2}} \left(\widehat{\beta}_t - \beta\right)} \\
\leq & \sqrt{\lambda_{\max}\left(V_t^{-1}\right)} \left\|\widehat{\beta}_t - \beta\right\|_{V_t} \\
\leq & \lambda_t^{-\frac{1}{2}} \left(\sqrt{\lambda_t} + \left(\frac{4\sqrt{2}}{1-p} + \frac{\sigma}{p}\right) \sqrt{d \log \frac{4t^2}{\delta}}\right) \\
\leq & D_{p,\sigma},
\end{aligned}$$

which implies C_t. Thus by Theorem 2.4 we have

$$\mathbb{P}\left(\bigcap_{t \geq \mathcal{E}} \{A_t \cap B_t \cap C_t\}\right) \geq 1 - 6\delta. \tag{2.15}$$

By Lemma 2.2, for each $t \geq \mathcal{E}$,

$$\begin{aligned}
\texttt{regret}(t) \leq & 2\left\{\Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_t) - \mathbb{E}\left[\Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_t) \middle| \mathcal{G}_{t-1}\right]\right\} \\
& + 2\left\{\mathbb{E}\left[\Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_t) \middle| \mathcal{G}_{t-1}\right] - \frac{1}{|\Psi_{t-1}|} \sum_{\tau \in \Psi_{t-1}} \Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_\tau)\right\} \\
& + \frac{2}{\sqrt{|\Psi_{t-1}|}} \left\|\beta - \widehat{\beta}_{t-1}\right\|_{V_{t-1}}.
\end{aligned}$$

Let

$$\begin{aligned}
R_1(t) :=& 2\left\{\Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_t) - \mathbb{E}\left[\Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_t) \middle| \mathcal{G}_{t-1}\right]\right\}, \\
R_2(t) :=& 2\left\{\mathbb{E}\left[\Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_t) \middle| \mathcal{G}_{t-1}\right] - \frac{1}{|\Psi_{t-1}|} \sum_{\tau \in \Psi_{t-1}} \Delta_{\widehat{\beta}_{t-1}}(\mathcal{X}_\tau)\right\}, \\
R_3(t) :=& \frac{2}{\sqrt{|\Psi_{t-1}|}} \left\|\beta - \widehat{\beta}_{t-1}\right\|_{V_{t-1}}.
\end{aligned} \tag{2.16}$$

Let us bound $R_1(t)$. Since the event C_t is $\mathcal{G}_t$-measurable for all $t \in [T]$, we have

$$R_1(t)\mathbb{I}\left(C_{t-1}\right) = 2\left\{\Delta_{\widehat{\beta}_{t-1}}\left(\mathcal{X}_t\right)\mathbb{I}\left(C_{t-1}\right) - \mathbb{E}\left[\Delta_{\widehat{\beta}_{t-1}}\left(\mathcal{X}_t\right)\mathbb{I}\left(C_{t-1}\right)\middle|\,\mathcal{G}_{t-1}\right]\right\}.$$

By Assumption 1,

$$\begin{aligned}
\Delta_{\widehat{\beta}_{t-1}}\left(\mathcal{X}_t\right)\mathbb{I}\left(C_{t-1}\right) :=& \max_{i\in[N]}\left|X_{i,t}^T\left(\widehat{\beta}_{t-1}-\beta\right)\right|\mathbb{I}\left(C_{t-1}\right)\\
\leq& \max_{i\in[N]}\left\|X_{i,t}\right\|_2\left\|\widehat{\beta}_{t-1}-\beta\right\|_2\mathbb{I}\left(C_{t-1}\right)\\
\leq& \left\|\widehat{\beta}_{t-1}-\beta\right\|_2\mathbb{I}\left(C_{t-1}\right)\\
\leq& D_{p,\sigma}.
\end{aligned}$$

Thus, $|R_1(t)\mathbb{I}\left(C_{t-1}\right)| \leq 4D_{p,\sigma}$. Since $R_1(t)\mathbb{I}\left(C_{t-1}\right)$ is $\mathcal{G}_t$-measurable and

$$\mathbb{E}\left[R_1(t)\mathbb{I}\left(C_{t-1}\right)|\,\mathcal{G}_{t-1}\right] = 0,$$

we can use Lemma 2.8 to have

$$\sum_{t>\mathcal{E}} R_1(t)\mathbb{I}\left(C_{t-1}\right) \leq 4D_{p,\sigma}\sqrt{2T\log\frac{1}{\delta}}, \tag{2.17}$$

with probability at least $1-\delta$.

Now we bound $R_2(t)$. By Lemma 2.3 with probability at least $1-\delta/T$,

$$
\begin{aligned}
&R_2(t)\mathbb{I}\left(A_{t-1}\cap C_{t-1}\right)\\
&\leq 2\mathbb{I}\left(A_{t-1}\right)\sup_{\|\beta_1-\beta\|_2\leq D_{p,\sigma}}\left|\mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right)|\,\mathcal{G}_{t-1}\right]-\frac{1}{|\Psi_{t-1}|}\sum_{\tau\in\Psi_{t-1}}\Delta_{\beta_1}\left(\mathcal{X}_\tau\right)\right|\\
&\leq\left(\frac{3\delta D_{p,\sigma}}{T}+8D_{p,\sigma}\sqrt{\frac{1}{|\Psi_{t-1}|}}\sqrt{d\log\frac{2T}{\delta}}\right)\mathbb{I}\left(A_{t-1}\right)\\
&\leq\frac{3\delta D_{p,\sigma}}{T}+8D_{p,\sigma}\sqrt{\frac{2}{pt}}\sqrt{d\log\frac{2T}{\delta}}.
\end{aligned}
$$

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

$$
\sum_{t>\mathcal{E}}R_2(t)\mathbb{I}\left(A_{t-1}\cap C_{t-1}\right)\leq 3\delta D_{p,\sigma}+\frac{16\sqrt{2}D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}}. \tag{2.18}
$$

To bound $R_3(t)$,

$$
\begin{aligned}
R_3(t)\mathbb{I}\left(A_{t-1}\cap B_{t-1}\right)\leq&\frac{2\sqrt{2}}{\sqrt{pt}}\left(1+\frac{4C}{1-p}+\frac{\sigma}{p}\right)\sqrt{d\log\frac{4t^2}{\delta}}\\
=&\frac{2\sqrt{2}}{\sqrt{pt}}D_{p,\sigma}\sqrt{d\log\frac{4t^2}{\delta}}.
\end{aligned}
$$

and

$$
\sum_{t>\mathcal{E}}R_3(t)\mathbb{I}\left(A_{t-1}\cap B_{t-1}\right)\leq\frac{8D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}}. \tag{2.19}
$$

□

Now for any $x > 2\mathcal{E}$,

$$
\begin{aligned}
&\mathbb{P}\left(R(T) > x\right) \\
&\leq \mathbb{P}\left(2\mathcal{E} + \sum_{t>\mathcal{E}} \texttt{regret}(t) > x\right) \\
&= \mathbb{P}\left(2\mathcal{E} + \sum_{t>\mathcal{E}} R_1(t) + R_2(t) + R_3(t) > x\right) \\
&\leq \mathbb{P}\left(2\mathcal{E} + \sum_{t>\mathcal{E}} R_1(t)\mathbb{I}\left(C_{t-1}\right) + R_2(t)\mathbb{I}\left(A_{t-1} \cap C_{t-1}\right) + R_3(t)\mathbb{I}\left(A_{t-1} \cap B_{t-1}\right) > x\right) \\
&+ \mathbb{P}\left(\bigcup_{t\geq\mathcal{E}} \{A_t^c \cup B_t^c \cup C_t^c\}\right) \\
&\leq \mathbb{P}\left(2\mathcal{E} + \sum_{t>\mathcal{E}} R_1(t)\mathbb{I}\left(C_{t-1}\right) + R_2(t)\mathbb{I}\left(A_{t-1} \cap C_{t-1}\right) + R_3(t)\mathbb{I}\left(A_{t-1} \cap B_{t-1}\right) > x\right) \\
&+ 6\delta.
\end{aligned}
$$

Setting

$$
x = 2\mathcal{E} + 4D_{p,\sigma}\sqrt{2T\log\frac{1}{\delta}} + 3\delta D_{p,\sigma} + \frac{16\sqrt{2}D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}} + \frac{8D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}},
$$

gives

$$
\begin{aligned}
\mathbb{P}\left(R(T) > x\right) \leq& 6\delta + \mathbb{P}\left(\sum_{t>\mathcal{E}} R_1(t)\mathbb{I}\left(C_{t-1}\right) > 4D_{p,\sigma}\sqrt{2T\log\frac{1}{\delta}}\right) \\
&+ \mathbb{P}\left(\sum_{t>\mathcal{E}} R_2(t)\mathbb{I}\left(A_{t-1} \cap C_{t-1}\right) > 3\delta D_{p,\sigma} + \frac{16\sqrt{2}D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}}\right) \\
&+ \mathbb{P}\left(\sum_{t>\mathcal{E}} R_3(t)\mathbb{I}\left(A_{t-1} \cap C_{t-1}\right) > \frac{8D_{p,\sigma}}{\sqrt{p}}\sqrt{dT\log\frac{2T}{\delta}}\right) \\
\leq& 8\delta,
\end{aligned}
$$

where the inequality holds due to (2.16), (2.17), (2.18) and (2.19).

2.7.3 Proof of Lemma 2.3

Proof. Let us fix $t \in [T]$ and $\Psi_t \subseteq [t]$. By Assumption 3, $\mathcal{X}_t$ is independent with $\mathcal{G}_{t-1}$. Thus,

$$\mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right) \middle| \mathcal{G}_{t-1}\right] = \mathbb{E}_X\left[\Delta_{\beta_1}\left(X\right)\right],$$

where $X \in \mathbb{R}^{d\times N}$ arises from P_X (Assumption 3). For any $x > 0$ and $\theta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\|\beta_1-\beta\|_2\le L}\left|\mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right)\middle|\mathcal{G}_{t-1}\right] - \frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}\left(\mathcal{X}_t\right)\right| > x \middle| \Psi_t\right)\\
&\le \exp\left(-\theta x\right)\mathbb{E}\left[\exp\left(\theta\sup_{\|\beta_1-\beta\|_2\le L}\left|\mathbb{E}_X\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right)\right] - \frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}\left(\mathcal{X}_t\right)\right|\right)\middle|\Psi_t\right].
\end{aligned}$$

Let $\tau_1 \le \tau_2, \ldots \le \tau_{|\Psi_t|}$ be an ordered round in Ψ_t. Then by Assumption 3, $\mathcal{X}_{\tau_1}, \ldots, \mathcal{X}_{\tau_{|\Psi_t|}}$ are IID random variables and we can use the symmetrization Lemma [van der Vaart and Wellner, 1996, Lemma 2.3.1] to have

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\theta\sup_{\|\beta_1-\beta\|_2\le L}\left|\mathbb{E}_X\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right)\right] - \frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}\left(\mathcal{X}_t\right)\right|\right)\right]\\
&\le \mathbb{E}\left[\exp\left(2\theta\sup_{\|\beta_1-\beta\|_2\le L}\left|\frac{1}{|\Psi_t|}\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\beta_1}\left(\mathcal{X}_{\tau_n}\right)\right|\right)\right],
\end{aligned} \tag{2.20}$$

where $\xi_1, \ldots, \xi_{|\Psi_t|}$ are independent Rademacher random variables. For any $\epsilon > 0$ let $\tilde{\beta}_1, \ldots, \tilde{\beta}_{\Theta(\epsilon)}$ be the ϵ-cover of $\mathcal{B} := \left\{\beta_1 \in \mathbb{R}^d : \|\beta_1 - \beta\|_2 \le L\right\}$. By the definition of ϵ-cover, for each $\beta_1 \in \mathcal{B}$, there exists $\tilde{\beta}_j$ such that $\left\|\tilde{\beta}_j - \beta_1\right\|_2 \le \epsilon$. Thus,

$$\begin{aligned}
\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\beta_1}\left(\mathcal{X}_{\tau_n}\right)\right| &\le \left|\sum_{n=1}^{|\Psi_t|}\xi_n\left\{\Delta_{\beta_1}\left(\mathcal{X}_{\tau_n}\right) - \Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right\}\right| + \left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right|\\
&\le \sum_{n=1}^{|\Psi_t|}\left|\Delta_{\beta_1}\left(\mathcal{X}_{\tau_n}\right) - \Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right| + \left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right|.
\end{aligned}$$

By the definition of $\Delta_{\beta_1}(\mathcal{X}_{\tau_n})$ and Assumption 1,

$$\begin{aligned}
\left|\Delta_{\beta_1}(\mathcal{X}_{\tau_n}) - \Delta_{\tilde{\beta}_j}(\mathcal{X}_{\tau_n})\right| &= \left|\max_i \left|X_{i,\tau_n}^T(\beta-\beta_1)\right| - \max_i \left|X_{i,\tau_n}^T\left(\beta-\tilde{\beta}_j\right)\right|\right| \\
&\leq \max_i \left|\left|X_{i,\tau_n}^T(\beta-\beta_1)\right| - \left|X_{i,\tau_n}^T\left(\beta-\tilde{\beta}_j\right)\right|\right| \\
&\leq \max_i \left|X_{i,\tau_n}^T\left(\beta_1-\tilde{\beta}_j\right)\right| \\
&\leq \max_i \|X_{i,\tau_n}\|_2 \left\|\beta_1-\tilde{\beta}_j\right\|_2 \\
&\leq \epsilon.
\end{aligned}$$

Thus,

$$\sup_{\|\beta_1-\beta\|_2\leq L}\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\beta_1}(\mathcal{X}_{\tau_n})\right| \leq |\Psi_t|\,\epsilon + \sup_{j=1,\ldots,\Theta(\epsilon)}\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}(\mathcal{X}_{\tau_n})\right|.$$

Plugging in (2.20) gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\|\beta_1-\beta\|_2\leq L}\left|\mathbb{E}\left[\Delta_{\beta_1}(\mathcal{X}_t)\middle|\mathcal{G}_{t-1}\right] - \frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}(\mathcal{X}_\tau)\right| > x\middle|\Psi_t\right) \\
&\leq \exp(-\theta x+\theta\epsilon)\,\mathbb{E}\left[\exp\left(\frac{2\theta}{|\Psi_t|}\sup_{j=1,\ldots,\Theta(\epsilon)}\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}(\mathcal{X}_{\tau_n})\right|\right)\middle|\Psi_t\right] \\
&\leq \exp(-\theta x+\theta\epsilon)\sum_{j=1}^{\Theta(\epsilon)}\mathbb{E}\left[\exp\left(\frac{2\theta}{|\Psi_t|}\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}(\mathcal{X}_{\tau_n})\right|\right)\middle|\Psi_t\right].
\end{aligned}$$

Since for each $j=1,\ldots,\Theta(\epsilon)$,

$$\left|\Delta_{\tilde{\beta}_j}(\mathcal{X}_{\tau_n})\right| \leq \max_i \|X_{i,\tau_n}\|_2 \left\|\beta-\tilde{\beta}_j\right\|_2 \leq L,$$

holds, by Hoeffding's Lemma,

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{2\theta}{|\Psi_t|}\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right|\right)\middle|\,\Psi_t\right] \\
&= \mathbb{E}\mathbb{E}\left[\exp\left(\frac{2\theta}{|\Psi_t|}\left|\sum_{n=1}^{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right|\right)\middle|\,\left\{X(\tau_n)\right\}_{n=1}^{|\Psi_t|},\Psi_t\right] \\
&= \mathbb{E}\prod_{n=1}^{|\Psi_t|}\mathbb{E}\left[\exp\left(\frac{2\theta}{|\Psi_t|}\xi_n\Delta_{\tilde{\beta}_j}\left(\mathcal{X}_{\tau_n}\right)\right)\middle|\,\left\{X(\tau_n)\right\}_{n=1}^{|\Psi_t|},\Psi_t\right] \\
&\leq \exp\left(\frac{2\theta^2L^2}{|\Psi_t|}\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\|\beta_1-\beta\|_2\leq L}\left|\mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right)\middle|\,\mathcal{G}_{t-1}\right]-\frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}\left(\mathcal{X}_\tau\right)\right|>x\middle|\,\Psi_t\right) \\
&\leq \exp\left(-\theta x+\theta\epsilon\right)2\Theta(\epsilon)\exp\left(\frac{2\theta^2L^2}{|\Psi_t|}\right) \\
&= 2\Theta(\epsilon)\exp\left\{-\theta\left(x-\epsilon\right)+\frac{2\theta^2L^2}{|\Psi_t|}\right\}.
\end{aligned}
$$

Minimizing with respect to $\theta>0$ gives,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{\|\beta_1-\beta\|_2\leq L}\left|\mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right)\middle|\,\mathcal{G}_{t-1}\right]-\frac{1}{|\Psi_t|}\sum_{\tau\in\Psi_t}\Delta_{\beta_1}\left(\mathcal{X}_\tau\right)\right|>x\middle|\,\Psi_t\right) \\
&\leq 2\Theta(\epsilon)\exp\left\{-\frac{|\Psi_t|\left(x-\epsilon\right)^2}{8L^2}\right\}.
\end{aligned}
$$

The covering number of $\mathcal{B}$ is bounded by $\Theta(\epsilon) \leq (\frac{3L}{\epsilon})^d$. Thus, with probability at least $1 - \delta/T$,

$$\begin{aligned}
&\sup_{\|\beta_1 - \beta\|_2 \leq L} \left| \mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right) | \, \mathcal{G}_{t-1}\right] - \frac{1}{|\Psi_t|} \sum_{\tau \in \Psi_t} \Delta_{\beta_1}\left(\mathcal{X}_\tau\right) \right| \\
&\leq \epsilon + L\sqrt{\frac{8}{|\Psi_t|}}\sqrt{\log \frac{2\Theta(\epsilon)T}{\delta}} \\
&\leq \epsilon + L\sqrt{\frac{8}{|\Psi_t|}}\sqrt{d \log \frac{3L}{\epsilon} + \log \frac{2T}{\delta}}.
\end{aligned}$$

Setting $\epsilon = 3L\delta/(2T)$ gives,

$$\begin{aligned}
&\sup_{\|\beta_1 - \beta\|_2 \leq L} \left| \mathbb{E}\left[\Delta_{\beta_1}\left(\mathcal{X}_t\right) | \, \mathcal{G}_{t-1}\right] - \frac{1}{|\Psi_t|} \sum_{\tau \in \Psi_t} \Delta_{\beta_1}\left(\mathcal{X}_\tau\right) \right| \\
&\leq \frac{3L\delta}{2T} + L\sqrt{\frac{8}{|\Psi_t|}}\sqrt{d \log \frac{2T}{\delta} + \log \frac{2T}{\delta}} \\
&\leq \frac{3L\delta}{2T} + 4L\sqrt{\frac{1}{|\Psi_t|}}\sqrt{d \log \frac{2T}{\delta}}.
\end{aligned}$$

□

2.7.4 Proof of Theorem 1.3

To prove Theorem 1.3, we need to prove the following bound for the imputation estimator $\breve{\beta}_t$ which is used in $\tilde{Y}_{i,t}$ and $\widehat{\beta}_t$.

Lemma 2.9. *Suppose the Assumptions 1-4 hold. Then for $\breve{\beta}_t$ computed in Algorithm 2.1, with probability at least $1 - \delta$,*

$$\left\| \breve{\beta}_t - \beta \right\|_2 \leq \frac{1}{N}, \tag{2.21}$$

holds for $t \geq \max\left\{\frac{1}{p}\log\frac{4T}{\delta}, C_{p,\sigma}N^2\phi^{-4}\log\frac{8T}{\delta}\right\}$.

Proof. Fix t and set $\gamma_t := 4\sqrt{2}N\sqrt{|\Psi_t| \log \frac{4t^2}{\delta}}$, and $W_t := \sum_{\tau \in \Psi_t} \sum_{i=1}^N X_{i,\tau} X_{i,\tau}^T + \sum_{\tau \notin \Psi_t} X_{a_\tau,\tau} X_{a_\tau,\tau} + \gamma_t I$. Then by definition of $\breve{\beta}_t$, we have

$$\begin{aligned}
\left\| \breve{\beta}_t - \beta \right\|_2 &= \left\| W_t^{-1} \left(\sum_{\tau \in \Psi_t} \sum_{i=1}^N X_{i,\tau} \tilde{Y}_{i,\tau} + \sum_{\tau \notin \Psi_t} X_{a_\tau,\tau} Y_{i,\tau} - W_t \beta \right) \right\|_2 \\
&\leq \left\| W_t^{-1} \right\|_2 \left\{ \left\| \sum_{\tau \in \Psi_t} \sum_{i=1}^N X_{i,\tau} \left(\tilde{Y}_{i,\tau} - X_{i,\tau}^T \beta \right) + \sum_{\tau \notin \Psi_t} X_{a_\tau,\tau} \eta_{a_\tau,\tau} \right\|_2 + \gamma_t \left\| \beta \right\|_2 \right\} \\
&\leq \lambda_{\min}\left(W_t\right)^{-1} \left\{ \left\| \sum_{\tau \in \Psi_t} \sum_{i=1}^N X_{i,\tau} \left(\tilde{Y}_{i,\tau} - X_{i,\tau}^T \beta \right) + \sum_{\tau \notin \Psi_t} X_{a_\tau,\tau} \eta_{a_\tau,\tau} \right\|_2 + \gamma_t \right\}.
\end{aligned} \tag{2.22}$$

For the minimum eigenvalue term, we have

$$\lambda_{\min}\left(W_t\right) \geq \lambda_{\min} \left(\sum_{\tau \in \Psi_t} \sum_{i=1}^N X_{i,\tau} X_{i,\tau}^T + \gamma_t I_d \right).$$

Let $\tau_1 < \tau_2 < \cdots < \tau_{|\Psi_t|}$ be the ordered rounds in Ψ_t. Since $\left\| \sum_{i=1}^N X_{i,\tau} X_{i,\tau}^T \right\|_F \leq N$ and

$$\lambda_{\min} \left(\mathbb{E} \left[\sum_{i=1}^N X_{i,\tau_k} X_{i,\tau_k}^T \middle| \mathcal{X}_{\tau_1}, \ldots, \mathcal{X}_{\tau_{k-1}} \right] \right) = \lambda_{\min} \left(\mathbb{E} \left[\sum_{i=1}^N X_{i,\tau_k} X_{i,\tau_k}^T \right] \right) \geq N\phi^2,$$

we can use Lemma 1.6 to have

$$\lambda_{\min}\left(W_t\right) \geq \lambda_{\min} \left(\sum_{\tau \in \Psi_t} \sum_{i=1}^N X_{i,\tau} X_{i,\tau}^T + \gamma_t I_d \right) \geq |\Psi_t| \, N\phi^2, \tag{2.23}$$

with probability at least $1-\frac{\delta}{t^2}$. By definition of $\tilde{Y}_{i,\tau}$, we have

$$\begin{aligned}\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\left(\tilde{Y}_{i,\tau}-X_{i,\tau}^T\beta\right) &= \sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)X_{i,\tau}X_{i,\tau}^T\left(\widehat{\beta}_{t-1}^{ridge}-\beta\right)\\ &\quad+\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\eta_{i,\tau}\\ &=\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}\left(\widehat{\beta}_{t-1}^{ridge}-\beta\right)\\ &\quad+\sum_{\tau\in\Psi_t}\frac{\eta_{\tilde{a}_\tau,\tau}}{\pi_{\tilde{a}_\tau,\tau}}X_{\tilde{a}_\tau,\tau},\end{aligned}$$

where $\boldsymbol{X}_{i,\tau}=X_{i,\tau}X_{i,\tau}^T$. Plugging this and (2.23) in (2.22) gives,

$$\begin{aligned}\left\|\breve{\beta}_t-\beta\right\|_2 \leq &\frac{1}{|\Psi_t|\,N\phi^2}\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}\left(\widehat{\beta}_{t-1}^{ridge}-\beta\right)\right\|_2\\ &+\frac{1}{|\Psi_t|\,N\phi^2}\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{\tilde{a}_\tau,\tau}}{\pi_{\tilde{a}_\tau,\tau}}X_{\tilde{a}_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_2+\frac{4\sqrt{2\log\frac{4t^2}{\delta}}}{\phi^2\sqrt{|\Psi_t|}}.\end{aligned} \tag{2.24}$$

For the first term,

$$\begin{aligned}&\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}\left(\widehat{\beta}_{t-1}^{ridge}-\beta\right)\right\|_2\\ &\leq\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}\right\|_2\left\|\widehat{\beta}_{t-1}^{ridge}-\beta\right\|_2\\ &\leq 2\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}\right\|_F.\end{aligned}$$

Define the filtration as $\mathcal{G}_0=\Psi_t$ and $\mathcal{G}_\tau=\mathcal{G}_{\tau-1}\cup\{\mathcal{X}_\tau,\tilde{a}_\tau,a_\tau\}$ for $\tau\in[t]$. This filtration refers to the case where the subset of rounds for using contexts from all arms is observed first and $\tilde{a}_1,\ldots,\tilde{a}_t$ is observed later. In Algorithm

2.1, $\tilde{a}_1, \ldots, \tilde{a}_t$ is observed first to determine Ψ_t. But in theoretical analysis, we define a novel filtration $\mathcal{G}_0, \ldots, \mathcal{G}_t$ to obtain a suitable bound by using the martingale method [Kontorovich and Ramanan, 2008]. Set

$$M := \sum_{\tau \in \Psi_t} \sum_{i=1}^{N} \left(1 - \frac{\mathbb{I}\left(\tilde{a}_\tau = i\right)}{\pi_{i,\tau}}\right) \boldsymbol{X}_{i,\tau},$$

and define $M_\tau = \mathbb{E}\left[M \middle| \mathcal{G}_\tau\right]$. Then $\{M_\tau\}_{\tau=0}^t$ is a $\mathbb{R}^{d\times d}$-valued martingale sequence since

$$\mathbb{E}\left[M_\tau \middle| \mathcal{G}_{\tau-1}\right] = \mathbb{E}\left[\mathbb{E}\left[M \middle| \mathcal{G}_\tau\right] \middle| \mathcal{G}_{\tau-1}\right] = \mathbb{E}\left[M \middle| \mathcal{G}_{\tau-1}\right] = M_{\tau-1}.$$

By Lemma 2.7, we can find a $\mathbb{R}^2$-valued martingale sequence $\{N_\tau\}_{\tau=0}^t$ such that $N_0 = (0,0)^T$ and

$$\|M_\tau\|_F = \|N_\tau\|_2, \quad \|M_\tau - M_{\tau-1}\|_F = \|N_\tau - N_{\tau-1}\|_2,$$

for all $\tau \in [t]$. Set $N_\tau = (N_\tau^{(1)}, N_\tau^{(2)})^T$. Then for each $r = 1, 2$ and $\tau \in [t]$,

$$\begin{aligned}
\left|N_\tau^{(r)} - N_{\tau-1}^{(r)}\right| &\leq \|N_\tau - N_{\tau-1}\|_2 \\
&= \|M_\tau - M_{\tau-1}\|_F \\
&= \|\mathbb{E}\left[M \middle| \mathcal{G}_\tau\right] - \mathbb{E}\left[M \middle| \mathcal{G}_{\tau-1}\right]\|_F \\
&= \begin{cases} \left\|\sum_{i=1}^N \left(1 - \frac{\mathbb{I}(\tilde{a}_\tau = i)}{\pi_{i,\tau}}\right) \boldsymbol{X}_{i,\tau}\right\|_F & \tau \in \Psi_t \\ 0 & \tau \notin \Psi_t \end{cases} \\
&\leq \begin{cases} \left\|\sum_{i=1}^N \boldsymbol{X}_{i,\tau}\right\|_F + \left\|\frac{1}{\pi_{\tilde{a}_\tau,\tau}} \boldsymbol{X}_{i,\tau}\right\|_F & \tau \in \Psi_t \\ 0 & \tau \notin \Psi_t \end{cases} \\
&\leq \begin{cases} N\left(\frac{2-p}{1-p}\right) & \tau \in \Psi_t \\ 0 & \tau \notin \Psi_t \end{cases},
\end{aligned}$$

holds almost surely. The third equality holds since for any $\tau \in [t]$,

$$\mathbb{E}\left[\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_u=i\right)}{\pi_{i,u}}\right)\boldsymbol{X}_{i,u}\middle|\mathcal{G}_\tau\right]=0,\ \forall u>\tau,$$

$$\mathbb{E}\left[M\middle|\mathcal{G}_\tau\right]=\sum_{u\in\Psi_t,u\leq\tau}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}.$$

Using Lemma 2.8, for $x>0$ and $r=1,2$,

$$\mathbb{P}\left(\left|N_\tau^{(r)}\right|>x\middle|\mathcal{G}_0\right)\leq 2\exp\left(-\frac{x^2}{2N^2\left|\Psi_t\right|\left(\frac{2-p}{1-p}\right)^2}\right),$$

which implies that

$$\mathbb{P}\left(\left|N_\tau^{(r)}\right|>N\left(\frac{2-p}{1-p}\right)\sqrt{2\left|\Psi_t\right|\log\frac{4t^2}{\delta}}\middle|\mathcal{G}_0\right)\leq\frac{\delta}{2t^2}.$$

Since

$$\left\|M\right\|_F=\left\|M_t\right\|_F=\left\|N_t\right\|_2\leq\left|N_t^{(1)}\right|+\left|N_t^{(2)}\right|,$$

we have

$$\mathbb{P}\left(\left\|M\right\|_F>2N\left(\frac{2-p}{1-p}\right)\sqrt{2\left|\Psi_t\right|\log\frac{4t^2}{\delta}}\middle|\Psi_t\right)\leq\frac{\delta}{t^2},$$

for any subset $\Psi_t\subseteq[t]$. Thus, we conclude that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}\left(\breve{\beta}_{t-1}-\beta\right)\right\|_2>4N\left(\frac{2-p}{1-p}\right)\sqrt{2\left|\Psi_t\right|\log\frac{4t^2}{\delta}}\right)\\
\leq&\mathbb{P}\left(2\left\|M\right\|_F>4N\left(\frac{2-p}{1-p}\right)\sqrt{2\left|\Psi_t\right|\log\frac{4t^2}{\delta}}\right)\\
\leq&\mathbb{E}\mathbb{P}\left(2\left\|M\right\|_F>4N\left(\frac{2-p}{1-p}\right)\sqrt{2\left|\Psi_t\right|\log\frac{4t^2}{\delta}}\middle|\Psi_t\right)\\
\leq&\frac{\delta}{t^2}.
\end{aligned}$$

Now for the second term in (2.24), we have for any $x > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{\tilde{a}_\tau,\tau}}{\pi_{\tilde{a}_\tau,\tau}}X_{\tilde{a}_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_2>x\right)\\
\leq&\mathbb{P}\left(\left\{\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{\tilde{a}_\tau,\tau}}{\pi_{\tilde{a}_\tau,\tau}}X_{\tilde{a}_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_2>x\right\}\bigcap\left\{\bigcap_{\tau\in\Psi_t}\{\tilde{a}_\tau=a_\tau\}\right\}\right)\\
&+\mathbb{P}\left(\bigcup_{\tau\in\Psi_t}\{\tilde{a}_\tau\neq a_\tau\}\right)\\
\leq&\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}}X_{a_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_2>x\right).
\end{aligned}
$$

Since $\pi_{a_\tau,\tau}=p$, we observe that $\frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}}$ and $\eta_{a_\tau,\tau}$ are $\frac{\sigma}{p}$-sub-Gaussian. Using Lemma 1.4, we have

$$
\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}}X_{a_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_2>\frac{C\sigma}{p}\sqrt{t}\sqrt{\log\frac{4t^2}{\delta}}\right)\leq\frac{\delta}{t^2},
$$

for some absolute constant $C>0$.

Now from (2.24), with probability $1-\frac{3\delta}{t^2}$, we have

$$
\left\|\breve{\beta}_t-\beta\right\|_2\leq\frac{1}{|\Psi_t|\,N\phi^2}\left\{4N\left(\frac{2-p}{1-p}\right)\sqrt{2\,|\Psi_t|\log\frac{4t^2}{\delta}}+\frac{C\sigma}{p}\sqrt{t}\sqrt{\log\frac{4t^2}{\delta}}\right\}+\frac{4\sqrt{2\log\frac{4t^2}{\delta}}}{\phi^2\sqrt{|\Psi_t|}}.
$$

By Lemma (2.5), $|\Psi_t|\geq\frac{p}{2}t$ for all $t\geq\frac{1}{p}\log\frac{T}{\delta}$, with probability at least $1-\delta$. Then we have

$$
\begin{aligned}
\left\|\breve{\beta}_t-\beta\right\|_2\leq&\frac{1}{\phi^2\sqrt{t}}\left\{\frac{8(2-p)}{(1-p)\sqrt{p}}+\frac{\sqrt{2}C\sigma}{p^2N}+\frac{8}{\sqrt{p}}\right\}\sqrt{2\log\frac{4t^2}{\delta}}\\
\leq&\frac{2}{\phi^2\sqrt{t}}\left\{\frac{8(2-p)}{(1-p)\sqrt{p}}+\frac{\sqrt{2}C\sigma}{p^2}+\frac{8}{\sqrt{p}}\right\}\sqrt{\log\frac{2T}{\delta}}.
\end{aligned}
$$

Set

$$C_{p,\sigma} := \frac{8(2-p)}{(1-p)\sqrt{p}} + \frac{\sqrt{2}C\sigma}{p^2} + \frac{8}{\sqrt{p}}. \tag{2.25}$$

Then for all $t \geq \max\left\{\frac{1}{p}\log\frac{T}{\delta}, C_{p,\sigma}N^2\phi^{-4}\log\frac{2T}{\delta}\right\}$, we have

$$\left\|\breve{\beta}_t - \beta\right\|_2 \leq \frac{1}{N},$$

with probability at least $1-4\delta$. □

Now we are ready to prove Theorem 2.4.

Proof. By the definition of $\widehat{\beta}_t$ in (1.2),

$$\begin{aligned}
\left\|\widehat{\beta}_t - \beta\right\|_{V_t} &= \left\| V_t^{-1}\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\tilde{Y}_{i,\tau} + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}Y_{a_\tau,\tau}V_t\beta\right)\right\|_{V_t} \\
&= \left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\tilde{Y}_{i,\tau} + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}Y_{a_\tau,\tau}V_t\beta\right\|_{V_t^{-1}} \\
&= \left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}\left(\tilde{Y}_{i,\tau} - X_{i,\tau}^T\beta\right) + \sum_{\tau\notin\Psi_t} X_{a_\tau,\tau}\left(Y_{a_\tau,\tau} - X_{a_\tau,\tau}^T\beta\right) - \lambda_t\beta\right\|_{V_t^{-1}}.
\end{aligned}$$

Set $\tilde{\eta}_{i,\tau} := \tilde{Y}_{i,\tau} - X_{i,\tau}^T\beta$. Since $Y_{a_\tau,\tau} = X_{a_\tau,\tau}^T\beta + \eta_{a_\tau,\tau}$, we have

$$\begin{aligned}
\left\|\widehat{\beta}_t - \beta\right\|_{V_t} &= \left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\tilde{\eta}_{i,\tau}X_{i,\tau} + \sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau} - \lambda_t\beta\right\|_{V_t^{-1}} \\
&\leq \left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\tilde{\eta}_{i,\tau}X_{i,\tau} + \sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_{V_t^{-1}} + \|\lambda_t\beta\|_{V_t^{-1}}. \qquad (2.26)
\end{aligned}$$

For the last term, we have

$$\|\lambda_t\beta\|_{V_t^{-1}} \leq \sqrt{\lambda_{\max}\left(V_t^{-1}\right)}\,\|\lambda_t\beta\|_2 \leq \sqrt{\lambda_t}\,\|\beta\|_2 \leq \sqrt{\lambda_t}, \tag{2.27}$$

where the last inequality holds due to Assumption 1. For the first term, we use the decomposition,

$$\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\tilde{\eta}_{i,\tau}X_{i,\tau} = \sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)X_{i,\tau}X_{i,\tau}^T(\breve{\beta}_t-\beta) + \sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\eta_{i,\tau}X_{i,\tau},$$

to have

$$\begin{aligned}
&\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\tilde{\eta}_{i,\tau}X_{i,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_{V_t^{-1}}\\
&\leq\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)X_{i,\tau}X_{i,\tau}^T(\breve{\beta}_t-\beta)\right\|_{V_t^{-1}}\\
&+\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\eta_{i,\tau}X_{i,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_{V_t^{-1}}. \qquad (2.28)
\end{aligned}$$

Let $\boldsymbol{X}_{i,\tau} := X_{i,\tau}X_{i,\tau}^T$. For the first term, we can use Lemma 2.9 to have

$$\begin{aligned}
&\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)\boldsymbol{X}_{i,\tau}(\breve{\beta}_t-\beta)\right\|_{V_t^{-1}}\\
&=\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)V_t^{-\frac{1}{2}}\boldsymbol{X}_{i,\tau}(\breve{\beta}_t-\beta)\right\|_2\\
&\leq\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)V_t^{-\frac{1}{2}}\boldsymbol{X}_{i,\tau}\right\|_2\left\|\breve{\beta}_t-\beta\right\|_2\\
&\leq\frac{1}{N}\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)V_t^{-\frac{1}{2}}\boldsymbol{X}_{i,\tau}\right\|_F.
\end{aligned}$$

With similar technique in the proof of Lemma 2.9, define the filtration as

$\mathcal{G}_0 = \Psi_t \cup \{\mathcal{X}_1, \ldots, \mathcal{X}_t\}$ and $\mathcal{G}_\tau = \mathcal{G}_{\tau-1} \cup \{\tilde{a}_\tau, a_\tau\}$ for $\tau \in [t]$. Set

$$M := \sum_{\tau \in \Psi_t} \sum_{i=1}^{N} \left(1 - \frac{\mathbb{I}\left(\tilde{a}_\tau = i\right)}{\pi_{i,\tau}}\right) V_t^{-\frac{1}{2}} \boldsymbol{X}_{i,\tau},$$

and define $M_\tau = \mathbb{E}\left[M \middle| \mathcal{G}_\tau\right]$. Then $\{M_\tau\}_{\tau=0}^t$ is a $\mathbb{R}^{d\times d}$-valued martingale sequence. Since for any $\tau \in [t]$, the contexts $\mathcal{X}_{\tau+1}, \ldots, \mathcal{X}_t$ are independent of $\tilde{a}_\tau$ and

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} \left(1 - \frac{\mathbb{I}\left(\tilde{a}_u = i\right)}{\pi_{i,u}}\right) V_t^{-\frac{1}{2}} \boldsymbol{X}_{i,u} \middle| \mathcal{G}_\tau\right] =& V_t^{-\frac{1}{2}} \sum_{i=1}^{N} \mathbb{E}\left[1 - \frac{\mathbb{I}\left(\tilde{a}_u = i\right)}{\pi_{i,u}} \middle| \mathcal{G}_\tau\right] \boldsymbol{X}_{i,u} \\ =& 0,\end{aligned}$$

for all $u > \tau$. This leads to

$$\mathbb{E}\left[M \middle| \mathcal{G}_\tau\right] = \sum_{u \in \Psi_t, u \le \tau} \sum_{i=1}^{N} \left(1 - \frac{\mathbb{I}\left(\tilde{a}_\tau = i\right)}{\pi_{i,\tau}}\right) V_t^{-\frac{1}{2}} \boldsymbol{X}_{i,\tau}.$$

By Lemma 2.7, we can find a $\mathbb{R}^2$-valued martingale sequence $\{N_\tau\}_{\tau=0}^t$ such that $N_0 = (0,0)^T$ and

$$\|M_\tau\|_F = \|N_\tau\|_2, \quad \|M_\tau - M_{\tau-1}\|_F = \|N_\tau - N_{\tau-1}\|_2,$$

for all $\tau \in [t]$. Set $N_\tau = (N_\tau^{(1)}, N_\tau^{(2)})^T$. Then for each $r = 1, 2$ and $\tau \in [t]$,

$$\begin{aligned}
\left| N_\tau^{(r)} - N_{\tau-1}^{(r)} \right| &\leq \left\| N_\tau - N_{\tau-1} \right\|_2 \\
&= \left\| M_\tau - M_{\tau-1} \right\|_F \\
&= \left\| \mathbb{E}\left[M \middle| \mathcal{G}_\tau \right] - \mathbb{E}\left[M \middle| \mathcal{G}_{\tau-1} \right] \right\|_F \\
&= \begin{cases} \left\| \sum_{i=1}^N \left(1 - \frac{\mathbb{I}(\tilde{a}_\tau = i)}{\pi_{i,\tau}} \right) V_t^{-\frac{1}{2}} \boldsymbol{X}_{i,\tau} \right\|_F & \tau \in \Psi_t \\ 0 & \tau \notin \Psi_t \end{cases} \\
&\leq \begin{cases} \sqrt{\sum_{i=1}^N \left(1 - \frac{\mathbb{I}(\tilde{a}_\tau = i)}{\pi_{i,\tau}} \right)^2} \sqrt{\sum_{i=1}^N \left\| V_t^{-\frac{1}{2}} \boldsymbol{X}_{i,\tau} \right\|_F^2} & \tau \in \Psi_t \\ 0 & \tau \notin \Psi_t \end{cases} \\
&\leq \begin{cases} 2 \frac{N}{1-p} \sqrt{\sum_{i=1}^N \left\| X_{i,\tau} \right\|_{V_t^{-1}}^2} & \tau \in \Psi_t \\ 0 & \tau \notin \Psi_t \end{cases},
\end{aligned}$$

holds almost surely. The last inequality holds due to

$$\begin{aligned}
\left\| V_t^{-1/2} \boldsymbol{X}_{i,\tau} \right\|_F^2 &= \mathrm{Tr}\left(\boldsymbol{X}_{i,\tau}^T V_t^{-1} \boldsymbol{X}_{i,\tau} \right) \\
&= X_{i,\tau}^T V_t^{-1} X_{i,\tau} \mathrm{Tr}\left(X_{i,\tau} X_{i,\tau}^T \right) \\
&= \left\| X_{i,\tau} \right\|_{V_t^{-1}}^2 \left\| X_{i,\tau} \right\|_2 \\
&\leq \left\| X_{i,\tau} \right\|_{V_t^{-1}}^2 .
\end{aligned}$$

Using Lemma 2.8, for $x > 0$ and $r = 1, 2$,

$$\mathbb{P}\left(\left| N_\tau^{(r)} \right| > x \middle| \mathcal{G}_0 \right) \leq 2 \exp\left\{ - \frac{x^2}{2 \left(\frac{2N}{1-p} \right)^2 \sum_{\tau \in \Psi_t} \sum_{i=1}^N \left\| X_{i,\tau} \right\|_{V_t^{-1}}^2} \right\}$$

which implies that

$$\mathbb{P}\left(\left|N_\tau^{(r)}\right| > \frac{2N}{1-p}\sqrt{2\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\|X_{i,\tau}\|_{V_t^{-1}}^2\right)\log\frac{4t^2}{\delta}}\middle|\mathcal{G}_0\right) \le \frac{\delta}{2t^2}.$$

Since

$$\|M\|_F = \|M_t\|_F = \|N_t\|_2 \le \left|N_t^{(1)}\right| + \left|N_t^{(2)}\right|,$$

we have

$$\mathbb{P}\left(\|M\|_F > \frac{4N}{1-p}\sqrt{2\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\|X_{i,\tau}\|_{V_t^{-1}}^2\right)\log\frac{4t^2}{\delta}}\middle|\Psi_t\right) \le \frac{\delta}{t^2},$$

for any subset $\Psi_t \subseteq [t]$. Let $U_t := \sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T + \lambda_t I$. Since $V_t \succeq U_t$, we have $\left\|X_{i,\tau^{(u)}}\right\|_{V_t^{-1}}^2 \le \left\|X_{i,\tau^{(u)}}\right\|_{U_t^{-1}}^2$. By the definition of the Frobenous norm and $\boldsymbol{X}_{i,\tau}$, we have

$$\begin{aligned}
\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left\|X_{i,\tau^{(u)}}\right\|_{U_t^{-1}}^2 &= \sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}^T U_t^{-1} X_{i,\tau} \\
&= \sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\mathrm{Tr}\left(X_{i,\tau}^T U_t^{-1} X_{i,\tau}\right) \\
&= \sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\mathrm{Tr}\left(X_{i,\tau} X_{i,\tau}^T U_t^{-1}\right) \\
&= \mathrm{Tr}\left(\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T\right)U_t^{-1}\right) \\
&\le \mathrm{Tr}\left(\left(\sum_{\tau\in\Psi_t}\sum_{i=1}^{N} X_{i,\tau}X_{i,\tau}^T + \lambda_t I\right)U_t^{-1}\right) \\
&= \mathrm{Tr}\left(I_d\right) = d.
\end{aligned}$$

Thus, we have

$$\mathbb{P}\left(\|M\|_F > \frac{4N}{1-p}\sqrt{2d\log\frac{4t^2}{\delta}}\middle|\Psi_t\right) \leq \frac{\delta}{t^2},$$

and

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\sum_{i=1}^{N}\left(1-\frac{\mathbb{I}\left(\tilde{a}_\tau=i\right)}{\pi_{i,\tau}}\right)V_t^{-\frac{1}{2}}\boldsymbol{X}_{i,\tau}\left(\breve{\beta}_t-\beta\right)\right\|_2 > \frac{4}{1-p}\sqrt{2d\log\frac{4t^2}{\delta}}\right)\\
\leq&\mathbb{P}\left(\frac{1}{N}\|M\|_F > \frac{4}{1-p}\sqrt{2d\log\frac{4t^2}{\delta}}\right)\\
\leq&\mathbb{E}\mathbb{P}\left(\|M\|_F > \frac{4N}{1-p}\sqrt{2d\log\frac{4t^2}{\delta}}\middle|\Psi_t\right)\\
\leq&\frac{\delta}{t^2}.
\end{aligned} \tag{2.29}$$

Now for the second term in (2.28), we have for any $x > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{\tilde{a}_\tau,\tau}}{\pi_{\tilde{a}_\tau,\tau}}X_{\tilde{a}_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_{V_t^{-1}} > x\right)\\
\leq&\mathbb{P}\left(\left\{\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{\tilde{a}_\tau,\tau}}{\pi_{\tilde{a}_\tau,\tau}}X_{\tilde{a}_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_{V_t^{-1}} > x\right\}\bigcap\left\{\bigcap_{\tau\in\Psi_t}\{\tilde{a}_\tau=a_\tau\}\right\}\right)\\
&+\mathbb{P}\left(\bigcup_{\tau\in\Psi_t}\{\tilde{a}_\tau\neq a_\tau\}\right)\\
\leq&\mathbb{P}\left(\left\|\sum_{\tau\in\Psi_t}\frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}}X_{a_\tau,\tau}+\sum_{\tau\notin\Psi_t}\eta_{a_\tau,\tau}X_{a_\tau,\tau}\right\|_{V_t^{-1}} > x\right).
\end{aligned}$$

Since $\pi_{a_\tau,\tau}=p$, we observe that $\frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}}$ and $\eta_{a_\tau,\tau}$ are $\frac{\sigma}{p}$-sub-Gaussian. Define

$W_t := \sum_{\tau=1}^{t} X_{a_\tau,\tau} X_{a_\tau,\tau}^T + \lambda_t I$. Since $V_t \succeq W_t$, we have

$$\left\| \sum_{\tau \in \Psi_t} \frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}} X_{a_\tau,\tau} + \sum_{\tau \notin \Psi_t} \eta_{a_\tau,\tau} X_{a_\tau,\tau} \right\|_{V_t^{-1}} \leq \left\| \sum_{\tau \in \Psi_t} \frac{\eta_{a_\tau,\tau}}{\pi_{a_\tau,\tau}} X_{a_\tau,\tau} + \sum_{\tau \notin \Psi_t} \eta_{a_\tau,\tau} X_{a_\tau,\tau} \right\|_{W_t^{-1}}.$$

By assumption 2, $\eta_{a_\tau,\tau}$ is a σ-sub-Gaussian random variable given $\mathcal{H}_\tau$, and $\mathcal{H}_{\tau+1}$-measurable. Since $X_{a_\tau,\tau}$ is $\mathcal{H}_\tau$-measurable, we can use Lemma 9 in Abbasi-Yadkori et al. [2011] to have

$$\left\| \sum_{\tau \in \Psi_t} \frac{\eta_{a_\tau,\tau}}{p} X_{a_t,\tau} + \sum_{\tau \notin \Psi_t} \eta_{a_\tau,\tau} X_{a_\tau,\tau} \right\|_{W_t^{-1}}^2 \leq \frac{\sigma^2}{p^2} d \log\left(\frac{t}{\delta}\right), \tag{2.30}$$

for all $t \geq 0$ with probability at least $1 - \delta$. Now with (2.26)-(2.30), we can conclude that

$$\begin{aligned} \left\| \widehat{\beta}_t - \beta \right\|_{V_t} \leq & \frac{4}{1-p} \sqrt{2d \log \frac{4t^2}{\delta}} + \frac{\sigma}{p} \sqrt{d \log\left(\frac{t}{\delta}\right)} + \sqrt{\lambda_t} \\ \leq & \left(\frac{4\sqrt{2}}{1-p} + \frac{\sigma}{p} \right) \sqrt{d \log \frac{4t^2}{\delta}} + \sqrt{\lambda_t}, \end{aligned}$$

with probability at least $1 - 6\delta$. □

2.7.5 Proof of Lemma 2.5

Proof. The proof follows from Chernoff's lower bound. In Algorithm 2.1, Ψ_t is constructed as $\Psi_t = \{\tau \in [t] : \tilde{a}_\tau = a_\tau\}$. Thus we have

$$|\Psi_t| = \sum_{\tau=1}^{t} \mathbb{I}\left(\tilde{a}_\tau = a_\tau\right).$$

Then for any $\epsilon \in (0,1)$ and $s < 0$,

$$\begin{aligned}\mathbb{P}\left(|\Psi_t| \leq \epsilon pt\right) =& \mathbb{P}\left(s\sum_{\tau=1}^{t}\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right) \geq s\epsilon pt\right) \\ \leq& \exp\left(-s\epsilon pt\right)\mathbb{E}\left[\exp\left(s\sum_{\tau=1}^{t}\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right)\right)\right].\end{aligned}$$

Let $\mathcal{G}_\tau = \mathcal{F}_\tau \cup \{\tilde{a}_1, \ldots, \tilde{a}_{\tau-1}\}$. Then $\mathbb{E}\left[\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right)\middle|\,\mathcal{G}_\tau\right] = p$, for all $\tau \in [t]$ and

$$\begin{aligned}&\mathbb{E}\left[\exp\left(s\sum_{\tau=1}^{t}\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right)\right)\right] \\ =&\mathbb{E}\mathbb{E}\left[\exp\left(s\sum_{\tau=1}^{t}\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right)\right)\middle|\,\mathcal{G}_t\right] \\ =&\mathbb{E}\left[\exp\left(s\sum_{\tau=1}^{t-1}\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right)\right)\mathbb{E}\left[\exp\left\{s\mathbb{I}\left(\tilde{a}_t = a_t\right)\right\}\middle|\,\mathcal{G}_t\right]\right] \\ =&\left\{(1-p) + pe^s\right\}\mathbb{E}\left[\exp\left(s\sum_{\tau=1}^{t-1}\mathbb{I}\left(\tilde{a}_\tau = a_\tau\right)\right)\right] \\ =&\vdots \\ =&\left\{(1-p) + pe^s\right\}^t \\ \leq&\left\{\exp\left(-p + pe^s\right)\right\}^t.\end{aligned}$$

The last inequality holds due to $1 + x \leq e^x$ for all $x \in \mathbb{R}$. Thus, we have

$$\mathbb{P}\left(|\Psi_t| \leq \epsilon pt\right) \leq \exp\left\{\left(e^s - s\epsilon - 1\right)pt\right\}.$$

The right hand side is minimized when $s = \log\epsilon$. Setting $s = \log\epsilon$ gives

$$\mathbb{P}\left(|\Psi_t| \leq \epsilon pt\right) \leq \exp\left\{\left(\epsilon - \epsilon\log\epsilon - 1\right)pt\right\} \leq \exp\left\{-2\left(1-\epsilon\right)pt\right\},$$

where the last inequality holds due to $\log x \geq 1 - x^{-1}$ for all $x > 0$. Setting the right hand side smaller than δ/T gives

$$t \geq \frac{1}{2p\,(1-\epsilon)} \log \frac{T}{\delta}. \tag{2.31}$$

For t that satisfies (2.31), $\mathbb{P}\left(|\Psi_t| \leq \epsilon p t\right) \leq \frac{\delta}{T}$ holds. □

2.7.6 Proof of Theorem 2.6

Proof. The proof is inspired by that of Theorem 5.1 in Auer et al. [2002], and that of Theorem 24.2 in Lattimore and Szepesvári [2020]. Define the context distribution $\mathcal{P}_X$ sampled from

$$\left(\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) \in \left(\mathbb{R}^d\right)^N.$$

Here, the covariance matrix $\mathbb{E}\left[N^{-1}\sum_{i=1}^{N} X_{i,t}X_{i,t}\right]$ is positive definite. Let $\eta_{i,t}$ be a random variable sampled from the normal distribution $\mathcal{N}(0, 1^2)$, independently. Then the reward distribution is Gaussian with mean $X_{i,t}^T\beta$, and variance 1^2. For each $i \in [d]$ let $\beta_i = (0, \ldots 0, \Delta, 0 \ldots, 0)$ where $\Delta > 0$ is in i-th component only. Then we have

$$\mathbb{E}_{\beta_i}\left[\sum_{t=1}^{T} X_{a_t^*,t}^T \beta\right] = \Delta T. \tag{2.32}$$

For each $i \in [d]$, we have

$$\mathbb{E}_{\beta_i}\left[\sum_{t=1}^{T} X_{a_t,t}^T \beta_i\right] = \Delta \mathbb{E}_{\beta_i}\left[\sum_{t=1}^{T} \mathbb{I}\left(a_t = i\right)\right].$$

Now set $\beta_0 = \mathbf{0}$. Let $\mathbb{P}_{\beta_i}$ and $\mathbb{P}_{\beta_0}$ be the laws of $\sum_{t=1}^{T} \mathbb{I}\left(a_t = i\right)$ with respect to the bandit/learner interaction measure induced by β_i and β_0 respectively. Then by the result in Exercise 14.4 in Lattimore and Szepesvári [2020],

$$\mathbb{E}_{\beta_i}\left[\sum_{t=1}^{T} \mathbb{I}\left(a_t = i\right)\right] \leq \mathbb{E}_{\beta_0}\left[\sum_{t=1}^{T} \mathbb{I}\left(a_t = i\right)\right] + T\sqrt{\frac{1}{2} D\left(\mathbb{P}_{\beta_0}, \mathbb{P}_{\beta_i}\right)},$$

where $D(\cdot,\cdot)$ is the relative entropy between two probability measures. Set $\mathcal{X}_t := (X_{1,t}, \ldots, X_{N,t})$. By the chain rule for the relative entropy,

$$\begin{aligned}
&D\left(\mathbb{P}_{\beta_0}, \mathbb{P}_{\beta_i}\right) \\
=&\sum_{t=1}^{T} D\left(\mathbb{P}_{\beta_0}\left(Y_{a_t,}|Y_{a_1,}, \ldots, Y_{a_{t-1},}, \mathcal{X}_1, \ldots, \mathcal{X}_t\right), \mathbb{P}_{\beta_i}\left(Y_{a_t,}|Y_{a_1,}, \ldots, Y_{a_{t-1},}, \mathcal{X}_1, \ldots, \mathcal{X}_t\right)\right) \\
&+\sum_{t=1}^{T} D\left(\mathbb{P}_{\beta_0}\left(\mathcal{X}_t|Y_{a_1,}, \ldots, Y_{a_{t-1},}, \mathcal{X}_1, \ldots, \mathcal{X}_{t-1}\right), \mathbb{P}_{\beta_i}\left(\mathcal{X}_t|Y_{a_1,}, \ldots, Y_{a_{t-1},}, \mathcal{X}_1, \ldots, \mathcal{X}_{t-1}\right)\right) \\
=&\sum_{t=1}^{T} \mathbb{E}_{\beta_0} \frac{\left\{X_{a_t,t}^T\left(\beta_i - \beta_0\right)\right\}^2}{2} \\
=&\frac{\Delta^2}{2} \mathbb{E}_{\beta_0}\left[\sum_{t=1}^{T} \mathbb{I}\left(a_t = i\right)\right],
\end{aligned}$$

where the second equality holds since the distribution of $\mathcal{X}_t$ is fixed, and

$$
\begin{aligned}
&D\left(\mathbb{P}_{\beta_0}\left(Y_{a_t,|}Y_{a_1,},\ldots,Y_{a_{t-1},,}\mathcal{X}_1,\ldots,\mathcal{X}_t\right),\mathbb{P}_{\beta_i}\left(Y_{a_t,|}Y_{a_1,},\ldots,Y_{a_{t-1},,}\mathcal{X}_1,\ldots,\mathcal{X}_t\right)\right)\\
&\quad=\int\int\log\frac{d\mathbb{P}_{\beta_i}(y|a_t)}{d\mathbb{P}_{\beta_0}(y|a_t)}d\mathbb{P}_{\beta_0}(y|a_t)d\mathbb{P}_{\beta_0}(a_t)\\
&\quad=\int\frac{\left\{X_{a_t,t}^T\left(\beta_i-\beta_0\right)\right\}^2}{2}d\mathbb{P}_{\beta_0}(a_t)\\
&\quad=\mathbb{E}_{\beta_0}\frac{\left\{X_{a_t,t}^T\left(\beta_i-\beta_0\right)\right\}^2}{2}.
\end{aligned}
$$

Thus we have

$$
\mathbb{E}_{\beta_i}\left[\sum_{t=1}^T X_{a_t,t}^T\beta_i\right]\leq\Delta\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]+\frac{\Delta^2T}{2}\sqrt{\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]}.
$$

With (2.32),

$$
\mathbb{E}_{\beta_i}\left[R(T)\right]\geq\Delta T-\Delta\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]-\frac{\Delta^2T}{2}\sqrt{\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]}.
$$

Taking average over $i\in[d]$ gives

$$
\begin{aligned}
\frac{1}{d}\sum_{i=1}^d\mathbb{E}_{\beta_i}\left[R(T)\right]&\geq\Delta T-\frac{\Delta}{d}\sum_{i=1}^d\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]-\frac{\Delta^2T}{2d}\sum_{i=1}^d\sqrt{\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]}\\
&\geq\Delta T-\frac{\Delta T}{d}-\frac{\Delta^2T\sqrt{d}}{2d}\sqrt{\sum_{i=1}^d\mathbb{E}_{\beta_0}\left[\sum_{t=1}^T\mathbb{I}\left(a_t=i\right)\right]}\\
&\geq\frac{\Delta T}{2}-\frac{\Delta^2T\sqrt{T}}{2\sqrt{d}}.
\end{aligned}
$$

Setting $\Delta=\frac{1}{2}\sqrt{\frac{d}{T}}$ gives

$$
\frac{1}{d}\sum_{i=1}^d\mathbb{E}_{\beta_i}\left[R(T)\right]\geq\frac{1}{8}\sqrt{dT}.
$$

Thus, there exists β_i such that $\mathbb{E}_{\beta_i}[R(T)] \geq \frac{1}{8}\sqrt{dT}$. $\square$

Bibliography

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.

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

Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs, 2014.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262, 2019.

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

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

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

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

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

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Mathematics*, 3(1):79–127, 2006.

Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. pages 355–366, 01 2008.

Maria Dimakopoulou, Zhengyuan Zhou, Susan Athey, and Guido Imbens. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3445–3453, 2019.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. An efficient algorithm for generalized linear bandit: Online stochastic gradient descent and thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 1585–1593. PMLR, 2021.

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

Nima Hamidi and Mohsen Bayati. On worst-case regret of linear thompson sampling, 2020.

Gisoo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pages 5869–5879, 2019.

Wonyoung Kim, Gisoo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling for linear payoffs, 2021. URL `https://arxiv.org/abs/2102.01229`.

Leonid Aryeh Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

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

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

James R. Lee, Yuval Peres, and Charles K. Smart. A gaussian upper bound for martingale small-ball probabilities. *Ann. Probab.*, 44(6):4184–4197, 11 2016. doi: 10.1214/15-AOP1073.

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, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2071–2080. JMLR. org, 2017.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459.

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

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. ISSN 00063444.

Joel A Tropp. User-friendly tail bounds for matrix martingales. Technical report, CALIFORNIA INST OF TECH PASADENA, 2011.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Aad W. van der Vaart and Jon A. Wellner. *Symmetrization and Measurability*, pages 107–121. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2545-2.

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

국 문 초 록

본 학위논문은 순차적 결정 문제(Sequential decision making problem)를 위한 효율적인 선형 다중 슬롯 머신 알고리즘(Linear Contextual Bandit Algorithm)을 제안한다. 선형 다중 슬롯 머신 알고리즘은 유한 개의 선택지(Arm)가 주어진 특정 환경 안에서 학습자가 그 선택지의 내용(Context)을 관찰하고 이들 중 보상 (Reward)을 최대화하는 행동을 파악하고 선택하는 방법론이다. 보상은 선택지의 내용과 선형 관계를 가지고 있다. 현재까지 제안된 선형 다중 슬롯 머신 알고리즘은 내용과 보상의 관계를 추정할 때, 선택된 내용과 보상으로만 추정하고 있다. 이는 선택되지 않은 내용들을 관찰만 하고 추정에는 사용할 수 없는 비효율성을 유발한다. 이로 인해 다중 슬롯 머신이 활용되는 뉴스 기사 배치 알고리즘이나 광고 추천 알고리즘이나 모바일 건강관리 시스템 등에서 선택받지 못한 기사, 광고, 건강관리법과 같은 내용이 추정에 사용될 수 없는 비효율성이 발생한다.

본 학위논문에서는 선택받지 못한 내용들도 추정에 활용할 수 있는 새로운 선형 다중 슬롯 머신 알고리즘 두 가지를 제안하였다. 첫째는 결측자료 분석법 중 이중 강건법(Doubly Robust)을 적용하여 관측하지 못한 보상을 유사 보상(Pseudo-reward)으로 대체하면서 선택되지 못한 내용도 추정에 활용할 수 있도록 하였고, 이를 통해 내용의 차원의 제곱근만큼 후회 상한 (Regret bound)를 개선하였다. 둘째는 간단한 랜덤화(Randomization)를 적용하여 선택받지 못한 내용을 활용하는 방법과 선택한 내용만 사용하는 방법을 혼합하여 만든 혼합 추정량(Compound Estimator)을 정의하고, 이 알고리즘이 최적(Optimal rate)의 후회 상한을 가졌음을 증명하였다. 본 학위논문이 제안하는 새 알고리즘은 선택받지 못한 내용을 활용하면서 이론적으로 성능이 개선되었음이 증명되었고, 시뮬레이션 데이터에 적용한 결과를 통해서도 기존 알고리즘보다 성능이 개선되었음을 확인하였다.

주요어 : 다중 슬롯 머신, 효율적인, 성능개선, 결측자료, 랜덤화
학 번 : 2016-20263

감 사 의 글

2015 년 10 월, 대학 졸업을 앞둔 저의 마음은 불확실함과 두려움으로 가득했습니다. 군대를 가는 것도 두려웠고, 대학원에는 지원조차 할 자신도 없었습니다. 어디로 가야할지 알 수 없었습니다. 마침 아버지로부터 전화 한 통이 걸려 왔습니다. 저는 아버지가 '졸업 후 계획은 있냐?' '왜 아직도 계획이 없냐?' 하며 꾸짖으실 것 같아 전화 받는 것이 두려웠습니다. 그런데 아버지는 제가 졸업하고 어디로 가든지 필요한 것이 있으면 다 지원하겠다고만 하셨습니다. 여전히 저는 어디로 가야 할지 알지 못했지만, 저를 꾸짖지 않으시고 지원해주고자 하시는 아버지의 마음이 느껴졌을 때 힘이 되었습니다. 통화를 마치고 교회에 와서 성경을 읽는데, 야고보서 1 장 5 절 말씀이 보였습니다.

"너희 중에 누구든지 지혜가 부족하거든
모든 사람에게 후히 주시고 꾸짖지 아니하시는 하나님께 구하라
그리하면 주시리라"

이 말씀을 보았을 때, 저를 꾸짖지 않으시고 지원하겠다는 아버지의 그 마음이 하나님의 마음과 같다는 것이 느껴졌습니다. 하나님은 아버지를 통해 제게 후히 주시고 꾸짖지 아니하시는 하나님의 마음을 알려주셨습니다. 후히 주시고 꾸짖지 아니하시는 하나님의 마음을 알게 되었을 때, 제가 지혜가 부족하다고 솔직하게 기도할 수 있게 되었고, 하나님께 지혜를 구하게 되었습니다. 하나님은 야고보서 1 장 5 절 말씀대로 제게 지혜를 후하게 주셔서 통계학과 대학원에 지원할 자신감을 주시고 좋은 성적으로 입학하게 해주셨습니다. 그리고 그 주신 지혜 덕분에 이렇게 좋은 박사학위 논문도 쓸 수 있게 되었습니다.

그런데 하나님은 제게 지혜만 주신 것이 아니었습니다. 사실 저는 자기만을 사랑하고 하나님과 다른 사람들에게 무관심하고 무정하여 많은 상처를 준 죄인이었습니다. 이 죄로 인해 저는 하나님께 꾸짖음을 받고 영원한 지옥 형벌을 받을 수밖에 없었습니다. 그런데 하나님은 이러한 죄인을 꾸짖지 않으시고 다만 후한 지혜를 주셨습니다. 그것은 예수님이 십자가에 못박히신 채로 저의 죄로 인한 꾸짖음을 다 대신 받으시고 죽으셨기에 가능한 일이었습니다. 하나님은 제게 지혜만이 아니라, 하나뿐인 아들 예수님을 십자가에 못박도록 내주시고 제게 영원한 용서와 사랑까지 베풀어주셨던 것이었습니다.

이 하나님의 한없는 용서와 사랑, 그리고 예수님의 십자가와 부활의 복음이 없었다면 저도, 이 논문도 이 세상에 없었을 것입니다. 죄인을 꾸짖지 않으시고 한없는 용서와 사랑으로 후한 지혜를 베풀어주신 하나님과 예수님께 감사와 경배를 드립니다!