

# Adaptive Data Augmentation for Thompson Sampling

**Wonyoung Kim**

WYK7@CAU.AC.KR

*Department of Artificial Intelligence*

*Chung-Ang University*

*Seoul, Republic of Korea*

**Editor:** My editor

## Abstract

Linear contextual bandits offer a robust framework for sequential decision-making under uncertainty. While Thompson Sampling (TS) exhibits strong empirical performance, it fails to achieve the minimax optimal regret bound in frequentist analysis because it requires more information to construct a posterior distribution that is robust against worst-case parameters. This work proposes a novel adaptive data augmentation method for TS to extract more information from the explored samples than is possible with the conventional methods. Our approach generates synthetic samples specifically designed to attain semiparametric efficiency in estimation, thereby achieving the minimax optimal regret bound. Empirical evaluations demonstrate that our method significantly outperforms existing methods.

**Keywords:** Thompson Sampling, minimax optimal regret bound, hypothetical bandit problem, adaptive data augmentation, coupling.

## 1 Introduction

In linear contextual bandits, the objective is to select actions that maximize cumulative rewards, modeled as a linear function with unknown parameters. At each decision epoch, an agent selects an action from a set of context vectors to maximize cumulative rewards, assumed to be linear functions of the chosen contexts. A special case is the multi-armed bandit (MAB) problem, in which the contexts are standard Euclidean basis vectors. Compared with more complex models—such as generalized linear models or deep neural networks, which demand substantial updates at each step—LinCB and MAB offer high computational efficiency. LinCB methods have been widely applied in fields such as e-commerce personalization (Hsu et al., 2020), revenue management (Ferreira et al., 2018), clinical trials (Murphy, 2005), political-science experiments (Offer-Westort et al., 2021), and A/B testing in marketing (Satyal et al., 2018), as comprehensively surveyed by Bouneffouf et al. (2020).

Two primary algorithmic paradigms dominate the LinCB literature: the upper-confidence-bound algorithm (LinUCB) and Thompson Sampling (LinTS). LinUCB selects the arm that maximizes an upper confidence bound on its reward, whereas LinTS samples a parameter from its posterior (or estimated) distribution and selects the arm with the highest sampled reward. Empirical studies (Chapelle and Li, 2011) consistently demonstrate the superiority of LinTS over LinUCB across various scenarios. Nevertheless, a significant gap persists between the best-known frequentist regret bound for LinTS (Abeille et al., 2017) and

Table 1: Comparison of regret bounds and assumptions on LinTS

Reference	Regret bound	Key assumptions
Agrawal and Goyal (2013)	$\tilde{O}(d^{3/2}\sqrt{T})$	Normalized contexts
Abeille et al. (2017)	$\tilde{O}(d^{3/2}\sqrt{T})$	Normalized contexts
Dimakopoulou et al. (2019)	$\tilde{O}(d^{3/2}\sqrt{T})$	Normalized contexts
Kim et al. (2021)	$\tilde{O}(d\sqrt{T})$	IID contexts from special distributions
Huix et al. (2023)	$\tilde{O}(d\sqrt{T})$	Gaussian prior on parameter
<b>This work</b>	$\tilde{O}(d\sqrt{T})$	Bounded contexts

the minimax lower bound (Lattimore and Szepesvári, 2020). Closing this gap is challenging due to the selective reward observation inherent in LinTS, complicating variance control for optimal-arm reward estimation.

Why TS is difficult for the Convergence of the distribution is slower than the convergence of the samples in frequentist analysis. As frequentist bound derives the regret bound over all possible distribution of parameters, LinTS requires more information to achieve the optimal regret bound.

This paper addresses this critical gap by introducing a novel estimation approach capable of learning rewards for all arms without relying on IID or diversity conditions. In the proposed framework, a hypothetical bandit problem tailored for efficient parameter estimation is constructed. This hypothetical setup employs a set of orthogonal basis vectors, preserving the covariance structure of the original contexts while significantly reducing the effective number of arms. By coupling the hypothetical and original problems, the resulting estimator achieves a novel self-normalized bound based on a Gram matrix encompassing contexts from all arms, including unselected ones. The proposed algorithm, equipped with this new estimator, attains the minimax optimal regret bound up to logarithmic factors without restrictive assumptions on context distributions.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature on LinCB, highlighting key contributions of the proposed approach. Section 3 presents the formal problem formulation. Section 5 details the proposed estimator and algorithm along with their theoretical justifications. Section 6 provides a rigorous regret analysis, establishing the minimax optimality of the proposed method. Finally, Section 7 empirically validates the effectiveness of the proposed algorithm across various benchmark scenarios.

## 2 Related Literature

The linear contextual bandit (LinCB) problem, introduced by Abe and Long (1999), has become foundational in sequential decision-making tasks. Two predominant algorithmic frameworks for LinCB are the upper-confidence-bound algorithm (LinUCB) and Thompson Sampling (LinTS). LinUCB, which selects the arm maximizing the upper confidence bound of its reward, has been extensively studied (Auer, 2002a; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011). In contrast, LinTS, incorporating randomization by sampling from an estimated or posterior distribution of rewards, has attracted significant attention (Agrawal and Goyal, 2013; Abeille et al., 2017).

Empirical studies, such as Chapelle and Li (2011), demonstrate that `LinTS` frequently outperforms `LinUCB` in practical scenarios, including online advertising and recommendation systems.

Theoretically, given contexts of dimension  $d$  and time horizon  $T$ , `LinUCB` achieves a regret bound of  $\tilde{O}(d\sqrt{T})$ , matching the minimax lower bound of  $\Omega(d\sqrt{T})$  up to logarithmic factors (Lattimore and Szepesvári, 2020). In contrast, `LinTS` currently achieves a higher regret bound of  $\tilde{O}(d^{3/2}\sqrt{T})$ , and improving this bound remains an open problem.

Table 1 summarizes existing regret bounds and associated assumptions for various `LinTS` methods. Recent studies have contributed to narrowing this gap. Kim et al. (2021) introduced a doubly robust (DR) estimator instead of the ridge estimator, achieving a regret bound of  $\tilde{O}(\alpha^{-1}\sqrt{T})$  under independent contexts with strictly positive minimum eigenvalue  $\alpha > 0$ . Special cases with  $\alpha^{-1} = O(d)$  are further analyzed by Bastani et al. (2021) and Kim et al. (2023a). However, if  $\alpha$  is extremely small (e.g., fixed or highly correlated contexts), this bound can be worse than previous guarantees. Huix et al. (2023) achieved the minimax rate of  $\tilde{O}(d\sqrt{T})$ , but under the assumption of a Gaussian prior distribution on the parameter, yielding Bayesian rather than worst-case frequentist guarantees. For the multi-armed bandit (MAB) setting, Agrawal and Goyal (2017) and Zhu and Tan (2020) obtained minimax-optimal bounds, but analogous results for `LinCB` with arbitrary contexts remain unresolved.

Statistical techniques, particularly those addressing missing data, have significantly advanced `LinCB` research. Methods such as inverse probability weighting (IPW) and doubly robust estimation (DR) (Bang and Robins, 2005) tackle the selective reward observation problem by treating unselected rewards as missing data. Dimakopoulou et al. (2019) employed IPW in `LinTS`, obtaining a regret bound of  $\tilde{O}(d^{3/2}\sqrt{T})$ . Kim and Paik (2019) adapted DR methods to sparse, high-dimensional linear bandits, leveraging information from unselected contexts. Subsequent works by Kim et al. (2021) and Kim et al. (2023a) enhanced DR estimators under assumptions of stochastic contexts and generalized linear rewards, respectively. Kim et al. (2023b) further generalized DR methods to scenarios involving zero-probability arm selections. Despite these advancements, their reliance on IID or diversity assumptions limits broader applicability, leaving open the challenge of improving regret bounds for arbitrary contexts. This paper addresses this challenge by developing a novel estimator and algorithm that achieve the minimax optimal regret bound of  $\tilde{O}(d\sqrt{T})$  without restrictive assumptions on the context distributions.

The theoretical analysis of Thompson Sampling (TS) generally bifurcates into frequentist and Bayesian settings. Frequentist analyses, such as those by Hamidi and Bayati (2019), provide robust per-instance guarantees but often fail to capture the informative value of the prior. Furthermore, these bounds typically exhibit a sub-optimal factor of  $\sqrt{d}$  compared to the minimax lower bound; our work specifically aims to close this gap. In contrast, Bayesian analyses of TS, exemplified by Kveton et al. (2021), effectively incorporate prior information. However, these results are generally on average over bandit instances sampled from the prior, making them inherently weaker than frequentist per-instance analyses. Additionally, Bayesian frameworks often necessitate restrictive assumptions, such as Gaussian noise, to ensure closed-form posteriors, which is less general than the sub-Gaussian noise models prevalent in frequentist literature. While earlier Bayesian bounds were criticized for failing to match the logarithmic lower bound established by Lai and Robbins (1985), recent work

by Riou et al. (2023) has resolved this discrepancy. Unlike these average-case approaches, our analysis remains strictly frequentist, providing minimax optimal guarantees that hold for any fixed bandit instance.

Recent LinCB research has leveraged advanced statistical methodologies to enhance algorithmic performance and theoretical guarantees. Techniques from high-dimensional parameter estimation (Bühlmann and Van De Geer, 2011), optimal experimental design (Smith, 1918; Guttorp and Lindgren, 2009), and Bayesian optimization (Mockus, 2005) have been adapted to bandit settings. More recently, missing-data techniques have been employed to bridge the regret-bound gap by estimating missing rewards as though rewards from all arms were observed at every round (Kim and Paik, 2019; Kim et al., 2021). Unlike conventional estimators, which reduce errors solely for selected arms, these methods seek convergence across all arms. However, their reliance on inverse-probability weighting introduces variance that scales with the number of arms. To mitigate this, existing methods typically impose restrictive assumptions—such as independent-and-identically-distributed (IID) contexts or particular diversity conditions—limiting broader applicability. Addressing this limitation necessitates resolving open issues in both missing-data and LinCB literatures.

Semiparametric for LinCB.

### 3 Linear Contextual Bandit Problem

This section presents the notation used throughout the paper and the formal definition of the linear contextual bandit (LinCB) problem.

#### 3.1 Notations

For a natural number  $n \in \mathbb{N}$ , define  $[n] := 1, 2, \dots, n$ . For a positive semidefinite matrix  $M \in \mathbb{R}^{d \times d}$  and a vector  $x \in \mathbb{R}^d$ , let  $\|x\|_M := \sqrt{x^\top M x}$ . For two matrices  $A$  and  $B$ , write  $A \succ B$  (respectively  $A \succeq B$ ) if  $A - B$  is positive definite (respectively positive semidefinite).

#### 3.2 Problem Formulation

In LinCB, the environment defines a sequence of distributions over  $d$ -dimensional context vectors for  $K$  arms, constrained to the set. Deterministic contexts can also be represented by setting each as a Dirac measure. The time horizon  $T$  is finite but not known to the learner. At each round  $t \in [T]$ , the environment draws context vectors  $(X_{1,t}, \dots, X_{K,t})$  from  $\mathcal{P}_t$ , where  $X_{k,t}$  denotes the context for arm  $k$ . Assume that  $x_{\max}$  is known; if unknown, it can be replaced with  $X_{\max,t} := \max_{s \in [t]} \max_{k \in [K]} \|X_{k,s}\|_2$ .

Let  $\mathcal{H}_t$  be the sigma-algebra generated by the observations until before selecting an action at round  $t$ , i.e.,

$$\mathcal{H}_t = \bigcup_{\tau=1}^{t-1} [\{X_{i,\tau}\}_{i=1}^K \cup \{a_\tau\} \cup \{Y_{a_\tau,\tau}\}] \cup \{X_{i,t}\}_{i=1}^K.$$

Based on  $\mathcal{H}_t$ , the learner selects an arm  $a_t \in [K]$  and receives a reward  $Y_{a_t,t}$ . In linear contextual bandits (LinCB), rewards are linear in the context, given by:

$$Y_{a_t,t} = X_{a_t,t}^\top \theta_\star + \eta_{a_t,t},$$

where  $\theta_\star \in \mathbb{R}^d$  is the unknown parameter such that  $\|\theta_\star\|_2 \leq \theta_{\max}$  for some unknown  $\theta_{\max} > 0$ , and  $\eta_{at,t}$  is conditionally zero-mean and  $\sigma$ -sub-Gaussian noise:

$$\mathbb{E}[\exp(\lambda\eta_{at,t}) | \mathcal{H}_t] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R},$$

for some  $\sigma \geq 0$ .

To normalize the scale of regret, following standard convention (e.g., Abbasi-Yadkori et al., 2011), assume that  $|X_{k,t}^\top \theta_\star| \leq 1$  for all  $k \in [K]$  and  $t \in [T]$ . At each round  $t$ , the optimal arm  $a_t^\star$  is defined as  $a_t^\star := \arg \max_{k \in [K]} (X_{k,t}^\top \theta_\star)$ , and the instantaneous regret is:

$$\text{regret}(t) := X_{a_t^\star,t}^\top \theta_\star - X_{at,t}^\top \theta_\star.$$

The goal is to minimize the cumulative regret over  $T$  rounds:

$$R(T) := \sum_{t=1}^T \text{regret}(t).$$

This general formulation aligns with the standard LinCB setting (see, e.g., Abbasi-Yadkori et al., 2011 and Lattimore and Szepesvári, 2020) and encompasses specific cases studied in Kim et al. (2021) and Kim et al. (2023b).

## 4 Augmenting Hypothetical Contexts for Linear Bandits

In linear contextual bandits (LinCB), the ridge estimator with  $\ell_2$ -regularization is a widely used approach for estimating the unknown parameter. This regularization can be interpreted as augmenting the dataset with artificial observations. Let  $\mathbf{e}_i \in \mathbb{R}^d$  denote the  $i$ -th Euclidean basis vector. Then, the ridge estimator at round  $t$  can be expressed as

$$\left( \sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + \sum_{i=1}^d \mathbf{e}_i \mathbf{e}_i^\top \right)^{-1} \left( \sum_{s=1}^t Y_{a_s,s} X_{a_s,s} + \sum_{i=1}^d 0 \cdot \mathbf{e}_i \right), \quad (1)$$

which is equivalent to augmenting the dataset with dummy context–reward pairs  $(\mathbf{e}_i, 0)$  for  $i \in [d]$ . Bishop (1995) showed that the inclusion of such artificial data can improve generalization, i.e., performance on test data that are not used in training. However, augmenting with zero-valued rewards induces shrinkage toward the origin, resulting in an estimator that is not adaptive to the observed data.

This observation motivates the use of alternative augmented samples that enhance parameter learning more effectively. Several estimators in the literature can be interpreted through the lens of such augmentation. For example, Kveton et al. (2020) proposed a method that adds multiple random perturbations to each reward observation; this is equivalent to augmenting the dataset with multiple context–reward pairs. Kim et al. (2021) introduced a doubly robust (DR) estimator:

$$\left( \sum_{s=1}^t \sum_{k=1}^K X_{k,s} X_{k,s}^\top + \sum_{i=1}^d \sqrt{\lambda_t} \mathbf{e}_i (\sqrt{\lambda_t} \mathbf{e}_i)^\top \right)^{-1} \left( \sum_{s=1}^t \sum_{k=1}^K X_{k,s} Y_{k,s}^{\text{DR}} \right), \quad (2)$$

where  $\lambda_t = \Omega(\sqrt{t})$  is a regularization parameter and  $Y_{k,s}^{\text{DR}}$  denotes a DR pseudo-reward. This estimator can be interpreted as augmenting the unselected contexts with their corresponding unbiased pseudo-rewards  $(X_{k,s}, Y_{k,s}^{\text{DR}})$  for all  $k \in [K]$  and  $s \in [t]$ .

These augmented observations construct the Gram matrix that controls the self-normalized error of the estimator and influences its generalization performance. In the ridge estimator (1), the Gram matrix includes only the contexts corresponding to selected arms, and thus the estimator converges within the span of those vectors. In contrast, the DR estimator (2) incorporates all contexts, yielding a more well-conditioned matrix and enabling convergence in all directions spanned by the  $K$  context vectors.

The well-conditioned Gram matrix plays a critical role in determining the convergence rate of the estimator in both linear regression and bandit settings. In the experimental design literature (e.g., Smith, 1918; Guttorp and Lindgren, 2009) and linear bandits (e.g., Soare et al., 2014; Tao et al., 2018), techniques such as E-optimal design aim to maximize the minimum eigenvalue of the Gram matrix to enhance estimation quality. Consequently, designing augmentations that yield well-conditioned Gram matrices is essential for accurate parameter estimation and low regret.

Although the Gram matrix of the DR estimator includes all  $K$  arms, the augmentation involves  $K$  augmented samples in each round, which introduces additional variance that scales linearly with  $K$ . To address this, Kim et al. (2021) assumed that the context covariance matrix has a strictly positive minimum eigenvalue, ensuring rapid increase in the minimum eigenvalue of the covariance.

Rather than augmenting with the full set  $\{(X_{k,s}, Y_{k,s}^{\text{DR}}) : k \in [K], s \in [t]\}$ , the proposed method constructs a hypothetical dataset with reduced number of augmented samples by identifying a set of orthogonal eigenvectors that are informative for learning all  $K$  rewards in each round. Moreover, instead of using dummy vectors such as  $(\mathbf{e}_i, 0)$ , the proposed estimator augments basis vectors orthogonal to the span of the observed contexts, thereby adaptively improving generalization to future inputs.

## 5 Proposed Method

This section introduces the proposed estimation scheme that enables LinTS to achieve a nearly minimax-optimal regret bound. Section 4 motivates the use of hypothetical sample augmentation for parameter estimation. Section 5.1 presents a construction of hypothetical contexts that efficiently support parameter learning while minimizing the number of augmented samples. Building on the constructed contexts, Section 5.3 defines an adaptive hypothetical bandit problem tailored for estimation. Section 5.4 describes a resampling strategy that couples the hypothetical and original bandit problems. Finally, Section 5.5 outlines the proposed algorithm, which incorporates the novel estimator derived from this framework.

### 5.1 Design of Hypothetical Contexts

For each round  $t$ , let  $a_t \sim \pi_t$  denote the arm drawn according to the policy  $\pi_t$ . Before selecting  $a_t$ , a set of hypothetical contexts is constructed to dominate the covariance structure of the original context vectors. Let  $\mathbf{X}_t = [X_{1,t}, \dots, X_{K,t}] \in \mathbb{R}^{d \times K}$  denote a concatenated context matrix and  $r_t$  denote its rank. Note that  $r_t \leq \min\{d, K\}$ . By applying

the singular value decomposition, there exists unitary matrices  $\mathbf{U}_t \in \mathbb{R}^{d \times d}$ ,  $\mathbf{V}_t \in \mathbb{R}^{K \times K}$ , and a rectangular diagonal matrix  $\Lambda_t \in \mathbb{R}^{d \times K}$  such that  $\mathbf{X}_t = \mathbf{U}_t \Lambda_t \mathbf{G}_t$ . For each  $k \in [K]$  let  $\mathbf{e}_k \in \mathbb{R}^K$  denote the Euclidean basis in  $\mathbb{R}^K$ . Then, writing  $\mathbf{U}_t = [u_{1,t}, \dots, u_{d,t}]$  for  $u_{i,t} \in \mathbb{R}^d$ ,

$$X_{k,t} = \sum_{i=1}^{r_t} \mathbf{V}_t(k, i) \lambda_{i,t} u_{i,t}, \quad (3)$$

where  $\mathbf{V}_t(k, i)$  is the entry on  $k$ -th row and  $i$ -th column of  $\mathbf{G}_t$ . Because  $\|X_k\|_2 \leq x_{\max}$  and  $u_{1,t}, \dots, u_{r_t,t}$  are orthonormal,

$$\|X_{k,t}\|_2^2 = \sum_{i=1}^{r_t} \mathbf{V}_t(k, i)^2 \lambda_{i,t}^2 \leq x_{\max}^2 \quad (4)$$

Let  $\mathbf{P}_{a_t,t} := X_{a_t,t} X_{a_t,t}^\top / \|X_{a_t,t}\|_2^2$  denote a projection matrix onto  $X_{a_t,t}$ . For brevity, let us write  $\bar{x}_{\max} := \max\{x_{\max}, 1\}$ . Define the hypothetical contexts for the round  $t$  by,

$$H_{i,t} := \begin{cases} \bar{x}_{\max} (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t}, & \text{for } i = 1, \dots, r_t, \\ \sqrt{\bar{x}_{\max}^2 - \|X_{a_t,t}\|_2^2} \frac{X_{a_t,t}}{\|X_{a_t,t}\|_2} & \text{for } i = r_t + 1 \\ X_{a_t,t}, & \text{for } i = r_t + 2. \end{cases}$$

The hypothetical contexts consist of three components: (i) bases that are projected onto the orthogonal space of the selected contexts (ii) elongated selected contexts, and (iii) the selected contexts. This construction ensures the following property.

**Lemma 1 (Hypothetical Gram Matrix Domination Lemma)** *For each  $t$ ,*

$$2 \sum_{i=1}^{r_t+2} H_{i,t} H_{i,t}^\top \succeq X_{k,t} X_{k,t}^\top$$

for any  $k \in [K]$ .

**Proof** For each  $k \in [K]$ , by (3),

$$(\mathbf{I}_d - \mathbf{P}_{a_t,t}) X_{k,t} = \sum_{i=1}^{r_t} \mathbf{V}_t(k, i) \lambda_{i,t} (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t}.$$

For any vector  $\xi \in \mathbb{R}^d$ , by Cauchy-Schwartz inequality,

$$\begin{aligned} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) X_{k,t})^2 &= \left( \sum_{i=1}^{r_t} \mathbf{V}_t(k, i) \lambda_{i,t} \xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t} \right)^2 \\ &\leq \left( \sum_{i=1}^{r_t} \mathbf{V}_t(k, i)^2 \lambda_{i,t}^2 \right) \left( \sum_{i=1}^{r_t} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t})^2 \right) \end{aligned}$$

By definition of  $H_{i,t}$ ,

$$(\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) X_{k,t})^2 \leq \frac{\left( \sum_{i=1}^{r_t} \mathbf{V}_t(k, i)^2 \lambda_{i,t}^2 \right)}{\bar{x}_{\max}^2} \left( \sum_{i=1}^{r_t} \xi^\top H_{i,t} \right)^2 \leq \left( \sum_{i=1}^{r_t} \xi^\top H_{i,t} \right)^2, \quad (5)$$

where the last inequality holds by (4). For  $a, b \in \mathbb{R}$ ,  $(a - b)^2 \geq a^2/2 - b^2$  and thus,

$$(\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) X_{k,t})^2 \geq \frac{(\xi^\top X_{k,t})^2}{2} - (\xi^\top \mathbf{P}_{a_t,t} X_{k,t})^2.$$

By Cauchy-Schwartz inequality,

$$(\xi^\top \mathbf{P}_{a_t,t} X_{k,t})^2 \leq (\xi^\top X_{a_t,t})^2 \frac{\|X_{k,t}\|_2^2}{\|X_{a_t,t}\|_2^2} \leq (\xi^\top X_{a_t,t})^2 \frac{\bar{x}_{\max}^2}{\|X_{a_t,t}\|_2^2}.$$

By definition of  $H_{r_t+1}$ , it follows that  $(\xi^\top \mathbf{P}_{a_t,t} X_{k,t})^2 \leq (\xi^\top H_{r_t+1,t})^2$  and

$$(\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) X_{k,t})^2 \geq \frac{(\xi^\top X_{k,t})^2}{2} - (\xi^\top H_{r_t+1,t})^2.$$

Plugging in (5) and rearranging the terms

$$\frac{(\xi^\top X_{k,t})^2}{2} \leq \left( \sum_{i=1}^{r_t+2} \xi^\top H_{i,t} \right)^2,$$

which completes the proof. ■

The novel Gram matrix is equivalent to the sum of basis vectors inflated by  $\bar{x}_{\max}^2$ . While Kim et al. (2021) augmented  $K$  arms which may be too excessive and cause additional error, the proposed method augments  $r_t + 1$  vectors;  $r_t$  vectors are orthogonal to the selected contexts and the other vectors are align

**Lemma 2 (Hypothetical Gram Matrix Equivalence Lemma)** *For each  $t$ ,*

$$\frac{\bar{x}_{\max}^2}{2} \sum_{i=1}^{r_t} u_{i,t} u_{i,t}^\top \preceq \sum_{i=1}^{r_t+2} H_{i,t} H_{i,t}^\top \preceq 5\bar{x}_{\max}^2 \sum_{i=1}^{r_t} u_{i,t} u_{i,t}^\top.$$

**Proof** For any vector  $\xi \in \mathbb{R}^d$ , by definition of  $H_{i,t}$ ,

$$\sum_{i=1}^{r_t+2} (\xi^\top H_{i,t})^2 = \bar{x}_{\max}^2 \sum_{i=1}^{r_t} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t})^2 + \frac{\bar{x}_{\max}^2}{\|X_{a_t,t}\|_2^2} (\xi^\top X_{a_t,t})^2. \quad (6)$$

To prove the lower Lowner order, since  $(a - b)^2 \geq a^2/2 - b^2$ ,

$$\sum_{i=1}^{r_t} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t})^2 \geq \frac{1}{2} \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2 - \sum_{i=1}^{r_t} (\xi^\top \mathbf{P}_{a_t,t} u_{i,t})^2.$$

By definition of  $\mathbf{P}_{a_t,t}$  and  $u_{1,t}, \dots, u_{r_t,t}$  are orthonormal,

$$\sum_{i=1}^{r_t} (\xi^\top \mathbf{P}_{a_t,t} u_{i,t})^2 \leq \frac{\xi^\top X_{a_t,t} X_{a_t,t}^\top (\sum_{i=1}^{r_t} u_{i,t} u_{i,t}^\top) X_{a_t,t} X_{a_t,t}^\top \xi}{\|X_{a_t,t}\|_2^4} \leq \frac{(\xi^\top X_{a_t,t})^2}{\|X_{a_t,t}\|_2^2}. \quad (7)$$

Thus,

$$\sum_{i=1}^{r_t} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t})^2 \geq \frac{1}{2} \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2 - \frac{(\xi^\top X_{a_t,t})^2}{\|X_{a_t,t}\|_2^2}.$$

Plugging in (6),

$$\sum_{i=1}^{r_t+2} (\xi^\top H_{i,t})^2 \geq \frac{\bar{x}_{\max}^2}{2} \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2,$$

which completes the lower Lowner order.

To prove the upper Lowner order, since  $(a - b)^2 \leq 2a^2 + 2b^2$

$$\bar{x}_{\max}^2 \sum_{i=1}^{r_t} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t})^2 \leq 2\bar{x}_{\max}^2 \sum_{i=1}^{r_t} ((\xi^\top u_{i,t})^2 + (\xi^\top \mathbf{P}_{a_t,t} u_{i,t})^2).$$

By (7),

$$\bar{x}_{\max}^2 \sum_{i=1}^{r_t} (\xi^\top (\mathbf{I}_d - \mathbf{P}_{a_t,t}) u_{i,t})^2 \leq 2\bar{x}_{\max}^2 \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2 + \frac{2\bar{x}_{\max}^2}{\|X_{a_t,t}\|_2^2} (X_{a_t,t}^\top \xi)^2.$$

Plugging in (6),

$$\sum_{i=1}^{r_t+2} (\xi^\top H_{i,t})^2 \leq 2\bar{x}_{\max}^2 \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2 + \frac{3\bar{x}_{\max}^2}{\|X_{a_t,t}\|_2^2} (\xi^\top X_{a_t,t})^2. \quad (8)$$

From the equation (3),

$$(\xi^\top X_{a_t,t})^2 = \left( \sum_{i=1}^{r_t} \mathbf{V}_t(a_t, i) \lambda_{i,t} (\xi^\top u_{i,t}) \right)^2 \leq \left( \sum_{i=1}^{r_t} \mathbf{V}_t(a_t, i)^2 \lambda_{i,t}^2 \right) \left( \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2 \right),$$

where the last inequality uses Cauchy-Schwartz inequality. Because  $\|X_{a_t,t}\|_2^2 = \sum_{i=1}^{r_t} \mathbf{V}_t(a_t, i)^2 \lambda_{i,t}^2$ ,

$$(\xi^\top X_{a_t,t})^2 \leq \|X_{a_t,t}\|_2^2 \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2.$$

Plugging in (8),

$$\sum_{i=1}^{r_t+2} (\xi^\top H_{i,t})^2 \leq 5\bar{x}_{\max}^2 \sum_{i=1}^{r_t} (\xi^\top u_{i,t})^2,$$

which completes the upper Lowner order. ■

Lemma ?? ensures that the compressed set  $\{Z_{i,t}\}_{i=1}^{r_t+1}$  dominates the outer product of any contexts, while reducing the number of arms to  $r_t + 1$ .

## 5.2 Orthogonal Basis Augmentation

To replace the artificial augmentation  $\{(\mathbf{e}_i, 0) : i \in [d]\}$  in the ridge estimator (1), an orthogonal basis is constructed at selected rounds. Given hyperparameters  $\delta \in (0, 1)$  and  $\gamma \in (0, 1)$ , define:

$$h_t := \left\lceil \frac{2}{\frac{1}{2} - e^{-1}} \frac{d}{1 - \gamma} \log \frac{d(t+1)^2}{\delta} \right\rceil, \quad (9)$$

which specifies the number of rounds allocated for orthogonal basis augmentation. The subset of rounds for the augmentation  $\mathcal{A}_t \subseteq [t]$  is defined recursively as:

$$\mathcal{A}_0 = \emptyset, \quad \mathcal{A}_t = \begin{cases} \mathcal{A}_{t-1} \cup \{t\}, & \text{if } |\mathcal{A}_{t-1}| < h_t, \\ \mathcal{A}_{t-1}, & \text{otherwise.} \end{cases} \quad (10)$$

Also define:

$$T_1 := \inf\{t \geq 1 : t \geq h_t\}. \quad (11)$$

By Proposition 13, the number of rounds  $T_1$  is bounded by

$$T_1 \leq \frac{8}{\frac{1}{2} - e^{-1}} \frac{d}{1 - \gamma} \left( 1 + \log \frac{4}{e/2 - 1} \frac{d}{1 - \gamma} \sqrt{\frac{d}{\delta}} \right).$$

Thus, for all  $t \geq T_1$ , it holds that  $h_t \leq |\mathcal{A}_t| \leq h_t + 1$

Now we introduce the orthogonal augmentation at round  $s \in \mathcal{A}_t$ . If  $r_s < d$ , the Gram–Schmidt process is applied to  $\{u_{i,s} : i \in [r_s]\}$  to construct an orthonormal set  $\{u_{i,s}\}_{i=r_s+1}^d$ . Recall that  $\bar{x}_{\max} = \max\{x_{\max}, 1\}$  and we define the hypothetical contexts for the orthogonal augmentation round by,

$$\tilde{H}_{i,s} := \begin{cases} \bar{x}_{\max} u_{i,s}, & \text{for } i = 1, \dots, d, \\ X_{a_s, s} & \text{for } i = d+1 \end{cases}$$

The equivalent of the sum of hypothetical contexts are as follows:

**Lemma 3 (Gram matrix with hypothetical contexts)** *For any  $t \geq T_1$  and any combinations of all actions  $(k_1, \dots, k_t) \in [K]^t$ , the Gram matrix  $\mathbf{G}_t$  satisfies*

$$\bar{x}_{\max} \mathbf{I}_d \preceq \sum_{i=1}^{d+1} \tilde{H}_{i,s} \tilde{H}_{i,s}^\top \preceq 2\bar{x}_{\max} \mathbf{I}_d.$$

For the augmentation round  $\mathcal{A}_t$ , the proposed method focuses on the construction of the orthogonal basis that may be required for the future rounds.

Now, define the number of arms in the hypothetical bandit problem at round  $s$  as:

$$N_s := \begin{cases} r_s + 2 & \text{if } s \in [t] \setminus \mathcal{A}_t, \\ d + 1 & \text{if } s \in \mathcal{A}_t. \end{cases} \quad (12)$$

The number of the hypothetical problem solely depends on whether the round  $s$  is included in  $\mathcal{A}_t$  or not. For all  $s \in [t]$  and  $i \in [N_s]$ , the **hypothetical contexts** are constructed as

$$Z_{i,s} := \begin{cases} H_{i,s}, & s \in [t] \setminus \mathcal{A}_t, \\ \tilde{H}_{i,s}, & s \in \mathcal{A}_t \end{cases}. \quad (13)$$

At round  $t$ , the Gram matrix of all hypothetical contexts is given by:

$$\mathbf{G}_t := \sum_{s=1}^t \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top = \sum_{s \in [t] \setminus \mathcal{A}_t} \sum_{i=1}^{r_s+2} H_{i,s} H_{i,s}^\top + \sum_{s \in \mathcal{A}_t} \sum_{i=1}^{d+1} \tilde{H}_{i,s} \tilde{H}_{i,s}^\top,$$

and satisfies the following bounds.

**Lemma 4 (Gram matrix with hypothetical contexts)** *For any  $t \geq T_1$  and any combinations of all actions  $(k_1, \dots, k_t) \in [K]^t$ , the Gram matrix  $\mathbf{G}_t$  satisfies*

$$\mathbf{G}_t \succeq \sum_{s \in [t] \setminus \mathcal{A}_t} X_{k_s,s} X_{k_s,s}^\top + \max\{x_{\max}^2, 1\} h_t I_d.$$

**Proof** For  $s \in \mathcal{A}_t$ , by definition of  $\tilde{H}_{i,s}$ ,

$$\sum_{i=1}^{d+1} \tilde{H}_{i,s} \tilde{H}_{i,s}^\top \succeq \max\{x_{\max}^2, 1\} \sum_{i=1}^d u_{i,s} u_{i,s}^\top.$$

Because  $\{u_{i,s} : i \in [d]\}$  are  $d$  orthonormal vectors in  $\mathbb{R}^d$ , we obtain  $\sum_{i=1}^d u_{i,s} u_{i,s}^\top = I_d$ . Since  $t \geq T_1$  and  $|\mathcal{A}_t| \geq h_t$ ,

$$\sum_{s \in \mathcal{A}_t} \sum_{i=1}^{d+1} \tilde{H}_{i,s} \tilde{H}_{i,s}^\top \succeq h_t \max\{x_{\max}^2, 1\} I_d.$$

For  $s \in [t] \setminus \mathcal{A}_t$  and  $k_s \in [K]$ , from (??),

$$\sum_{i=1}^{r_s+1} H_{i,s} H_{i,s}^\top \succeq X_{k_s,s} X_{k_s,s}^\top.$$

Thus,

$$\mathbf{G}_t = \sum_{s \in [t] \setminus \mathcal{A}_t} X_{k_s,s} X_{k_s,s}^\top + h_t \max\{x_{\max}^2, 1\} I_d,$$

which completes the proof. ■

Lemma 4 ensures that the hypothetical contexts preserve the Gram matrix of all  $K$  context vectors while reducing the number of augmented samples to  $N_s$  in each round. The proposed method retains statistical efficiency comparable to full augmentation (Kim et al., 2021), with less number of augmented context samples.

### 5.3 A Hypothetical Linear Contextual Bandit

Based on the sample  $a_t \sim \pi_t$ , construct the hypothetical contexts  $Z_{i,s}$  as previously described. For each  $s \in [t]$  and  $i \in [N_s]$ , define the corresponding hypothetical rewards:

$$W_{i,s} := Z_{i,s}^\top \theta_\star + \eta_{a_s,s},$$

where  $\eta_{a_s,s}$  is shared with the original bandit problem.

Let  $\tilde{a}_s \in [N_s]$  be a hypothetical action sampled from the distribution:

$$\mathbb{P}(\tilde{a}_s = i) := \phi_{i,s} = \begin{cases} \frac{1-\gamma}{N_s-1}, & \text{if } i \in [N_s-1], \\ \gamma, & \text{if } i = N_s, \end{cases} \quad (14)$$

where  $\gamma \in (0, 1)$  determines the probability mass assigned to the original context  $X_{a_s,s}$ . As  $\gamma$  increases, the sampling concentrates on arm  $N_s$ , increasing the variance of inverse-probability weights for the other arms. To mitigate this, the number of rounds with orthogonal basis augmentation,  $|\mathcal{A}_t| \geq h_t$ , must be sufficiently large. This construction defines a hypothetical linear bandit problem  $\{(Z_{i,s}, W_{i,s}) : i \in [N_s], s \in [t]\}$  that shares the parameter  $\theta_\star$  with the original problem.

To perform estimation, construct the following ridge estimator as a reference:

$$\check{\theta}_t := \left( \sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + \gamma I_d \right)^{-1} \left( \sum_{s=1}^t X_{a_s,s} Y_{a_s,s} \right), \quad (15)$$

with regularization parameter  $\gamma \in (0, 1)$  as used in (14). Using  $\check{\theta}_t$ , define the pseudo-rewards:

$$\tilde{W}_{i,s}^{H(\check{\theta}_t)} := \left( 1 - \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \right) Z_{i,s}^\top \check{\theta}_t + \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} W_{i,s}. \quad (16)$$

This pseudo-reward is unbiased:  $\mathbb{E}[\tilde{W}_{i,s}^{H(\check{\theta}_t)}] = Z_{i,s}^\top \theta_\star$  for all  $i \in [N_s]$ . The estimator for  $\theta_\star$  is then:

$$\tilde{\theta}_t^{H(\check{\theta}_t)} := \left( \sum_{s=1}^t \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top \right)^{-1} \left( \sum_{s=1}^t \sum_{i=1}^{N_s} \tilde{W}_{i,s}^{H(\check{\theta}_t)} Z_{i,s} \right). \quad (17)$$

However, this estimator cannot be computed directly, as the rewards  $W_{i,s}$  for  $i \in [N_s-1]$  are unobserved. Only  $W_{N_s,s} = Y_{a_s,s}$  is available. Thus, (16) is computable for all  $i \in [N_s]$  only when  $W_{\tilde{a}_s,s} = Y_{a_s,s}$ , i.e., when the sampled context in the hypothetical bandit matches the observed context from the original bandit. Since both models share the same parameter  $\theta_\star$  and noise  $\eta_{a_s,s}$ , this condition is equivalent to  $Z_{\tilde{a}_s,s} = X_{a_s,s}$ . This matching event provides the motivation for the coupling technique introduced in the next section.

### 5.4 Coupling the Hypothetical and Original Linear Contextual Bandits

This section introduces a probabilistic method to couple the hypothetical bandit problem with the original contextual bandit. Figure 1 illustrates the overall coupling and resampling process. The key observation is that the hypothetical pseudo-reward in (16) is computable under the event  $\{Z_{\tilde{a}_s,s} = X_{a_s,s}\}$ , which is implied by  $\{\tilde{a}_s = N_s\}$ . To ensure this condition, we resample both  $a_s \sim \pi_s$  and  $\tilde{a}_s \sim \tilde{\pi}_s$  using the distribution in (14).

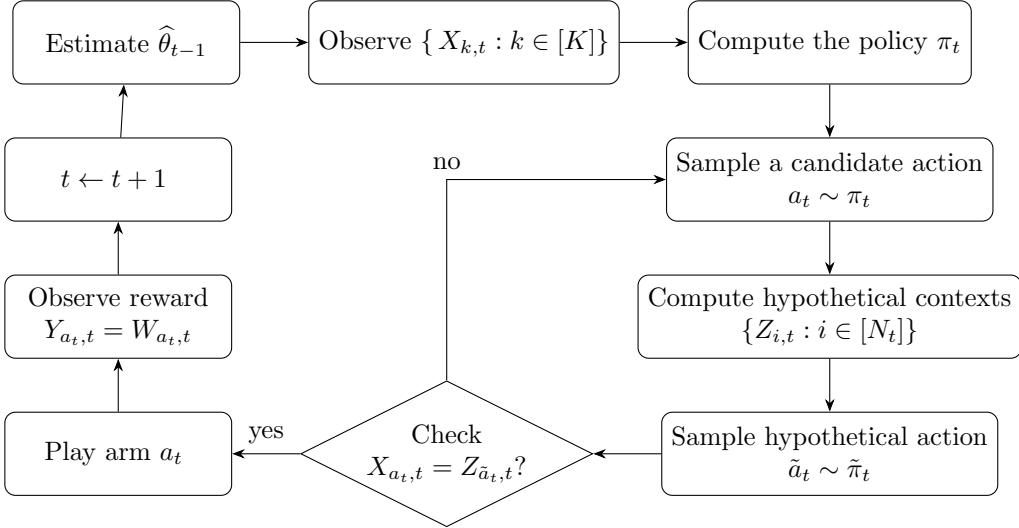


Figure 1: Flow diagram of the proposed coupling and resampling scheme

Each resampling iteration generates updated hypothetical contexts  $\{Z_{i,s} : i \in [N_s]\}$ , effectively randomizing the hypothetical contexts and rewards until the pseudo-reward becomes computable. Although rewards are collected from the original bandit, the estimation of the parameter  $\theta_\star$  is performed using compressed and augmented samples from the hypothetical bandit problem.

Let  $\tilde{a}_s(m)$  and  $a_s(m)$  denote the actions sampled during the  $m$ -th resampling trial in the hypothetical and original bandits, respectively. These actions are IID across trials  $m$  given  $\mathcal{H}_t$ . Define the stopping time for successful coupling as  $\xi_s := \inf\{m \geq 1 : X_{a_s(m)} = Z_{\tilde{a}_s(m)}\}$ . Then, define the matching event:

$$\mathcal{M}_s := \{\xi_s \leq M_s\}, \quad M_s := \left\lceil \frac{\log((s+1)^2/\delta)}{\log(1/(1-\gamma))} \right\rceil \quad (18)$$

which ensures a successful coupling within  $M_s$  trials. Since  $\mathbb{P}(\tilde{a}_s(m) = N_s) = \gamma$ , the number of trials  $M_s$  is selected to guarantee  $\mathbb{P}(\mathcal{M}_s) \geq 1 - \delta/(s+1)^2$ .

The hyperparameter  $\gamma$  controls the trade-off: as  $\gamma$  increases, the probability of coupling success increases (thus requiring fewer resampling trials), while the size of the regularization set  $h_t$  must increase. Table 2 illustrates the data structures for successful and failed couplings during resampling.

The proposed resampling-coupling scheme is described in Algorithm 1 as candidate arm sampler (CAS). The resampling in CAS is distinctive from that in Kim et al. (2021) and Xu and Zeevi (2020). The resampling in Kim et al. (2021) resamples the action to find the arm whose selection probability is greater than a prespecified threshold value. Xu and Zeevi (2020) resamples the previous counterfactual actions and contexts to impose randomization and generalization on the estimator. In contrast, our resampling is to couple the hypothetical samples with the original samples and this coupling is the first method a novel innovative part in this work.

Hypothetical Bandit Problem		Original Bandit Problem			
	Arm 1	Arm 2	Arm 1	Arm 2	Arm 3
Failure	$(Z_{\tilde{a}_t,t}, W_{\tilde{a}_t,t})$	$(Z_{N_t,t}, ?)$	$(X_{1,t}, ?)$	$(X_{2,t}, ?)$	$(X_{a_t,t}, Y_{a_t,t})$
Success	$(Z_{1,t}, ?)$	$(Z_{N_t,t}, W_{N_t,t})$	$(X_{1,t}, ?)$	$(X_{a_t,t}, Y_{a_t,t})$	$(X_{3,t}, ?)$

Table 2: Illustration of coupling success and failure during resampling for  $N_t = 2$  and  $K = 3$ . By construction,  $Z_{2,t} := X_{a_t,t}$  and  $W_{2,t} := Y_{a_t,t}$ . Gray cells indicate the selected actions.

---

**Algorithm 1** Candidate-Arm Sampler (CAS) for Round  $t$ 


---

- 1: **Input:** contexts  $\{X_{k,t}\}_{k \in [K]}$ , posterior mean  $\hat{\theta}_{t-1}$ , exploration variance  $v_{t-1}$ , Gram matrix  $V_{t-1}$ , pseudo-index  $N_t$ , coupling parameter  $\gamma$ , confidence  $\delta$ .
  - 2: **Set**  $M_t$  as in (18) { // maximum retries}
  - 3: **Initialise**  $m \leftarrow 1$
  - 4: **repeat**
  - 5:   Sample  $\tilde{\theta}_{k,t}^{(m)} \sim \mathcal{N}(\hat{\theta}_{t-1}, v_{t-1}^2 V_{t-1}^{-1})$  independently for all  $k \in [K]$
  - 6:    $a_t^{(m)} \leftarrow \arg \max_{k \in [K]} X_{k,t}^\top \tilde{\theta}_{k,t}^{(m)}$
  - 7:   Sample  $\tilde{a}_t^{(m)}$  from the distribution in (14)
  - 8:    $m \leftarrow m + 1$
  - 9: **until**  $\tilde{a}_t^{(m-1)} = N_t$  **or**  $m > M_t$
  - 10: **Output:**  $a_t^* \leftarrow a_t^{(m-1)}$ ,  $\tilde{a}_t^* \leftarrow \tilde{a}_t^{(m-1)}$
- 

Upon obtaining the coupled contexts  $Z_{\tilde{a}_s(M_s),s} = X_{a_s(M_s),s}$ , we construct the coupled pseudo-reward as:

$$W_{i,s}^{Co(\tilde{\theta}_t)} := \left( 1 - \frac{\mathbb{I}(\tilde{a}_s(M_s) = i)}{\phi_{i,s}} \right) Z_{i,s}^\top \check{\theta}_t + \frac{\mathbb{I}(\tilde{a}_s(M_s) = i)}{\phi_{i,s}} W_{i,s}, \quad (19)$$

which is computable for all  $i \in [N_s]$  because  $a_s(M_s)$  is selected and  $W_{\tilde{a}_s(M_s),s} = Y_{a_s(M_s),s}$  is observable.

Given an reference estimator  $\check{\theta}_t$  defined in (15), the proposed *hypothetical coupled sample augmented* (HCSA) estimator is defined as:

$$\hat{\theta}_t := \left\{ \sum_{s=1}^t \mathbb{I}(\mathcal{M}_s) \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top \right\}^{-1} \left( \sum_{s=1}^t \mathbb{I}(\mathcal{M}_s) \sum_{i=1}^{N_s} W_{i,s}^{Co(\tilde{\theta}_t)} Z_{i,s} \right). \quad (20)$$

The indicator  $\mathbb{I}(\mathcal{M}_s)$  makes the estimator use the coupled pseudo-rewards (19) only when  $\mathcal{M}_s$  occurs; otherwise, it skips round  $s$  and relies on the previous estimator. Since  $\mathcal{M}_s$  occurs with high probability, we can couple the HCSA estimator with the hypothetical sample augmented estimator from (17).

While the DR estimator in (2) where the  $K$  pseudo-rewards are augmented, the proposed (HCSA) estimator (17) adds  $N_s \leq d+1$  for each round  $s \in [t]$ . This reduction in the number

of augmented pseudo-reward samples paves a way to reduce the error and eliminate the IID and minimum eigenvalue assumption on contexts, by which Kim et al. (2021) used to obtain a regret bound that depends on the minimum eigenvalue of the context covariance.

Next, we provide a coupling inequality that relates the HCSA estimator to hypothetical sample augmented estimator  $\tilde{\theta}_t^{H(\check{\theta}_t)}$ .

**Lemma 5 (A coupling inequality)** *For  $t \geq 1$ , let  $\mathcal{S}_t := \cap_{s=1}^t \mathcal{M}_s$ , where  $\mathcal{M}_s$  is the matching event defined in (18). For the reference estimator  $\theta_t$  defined in (15) and for  $x > 0$ ,*

$$\mathbb{P}\left(\left\{\left\|\hat{\theta}_t^{A(\check{\theta}_t)} - \theta_\star\right\|_{\mathbf{G}_t} > x\right\}\right) \leq \mathbb{P}\left(\left\{\left\|\tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star\right\|_{\mathbf{G}_t} > x\right\} \cap \mathcal{S}_t\right) + \mathbb{P}(\mathcal{S}_t^c),$$

and the failure probability satisfies  $\mathbb{P}(\mathcal{S}_t^c) \leq \delta$ .

**Proof** Fix  $t \in [T]$  throughout the proof. For any  $\check{\theta} \in \mathbb{R}^d$  and  $x > 0$ , decompose the probability as follows:

$$\mathbb{P}\left(\left\|\hat{\theta}_t - \theta_\star\right\|_{\mathbf{G}_t} > x\right) \leq \mathbb{P}\left(\left\{\left\|\hat{\theta}_t - \theta_\star\right\|_{\mathbf{G}_t} > x\right\} \cap \mathcal{S}_t\right) + \mathbb{P}(\mathcal{S}_t^c).$$

On the event  $\mathcal{S}_t := \cap_{s=1}^t \mathcal{M}_s$ , the HCSA estimator in (20) is simplified to

$$\hat{\theta}_t^{A(\check{\theta}_t)} = \left\{\sum_{s=1}^t \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top\right\}^{-1} \left(\sum_{s=1}^t \sum_{i=1}^{N_s} W_{i,s}^{Co(\check{\theta}_t)} Z_{i,s}\right) = \mathbf{G}_t \left(\sum_{s=1}^t \sum_{i=1}^{N_s} W_{i,s}^{Co(\check{\theta}_t)} Z_{i,s}\right).$$

Define the function

$$F(\tilde{a}_1(M_1), \dots, \tilde{a}_t(M_t)) := \left\|\hat{\theta}_t - \theta_\star\right\|_{\mathbf{G}_t} = \left\|\sum_{s=1}^t \sum_{i=1}^{N_s} (W_{i,s}^{Co(\check{\theta}_t)} - Z_{i,s}^\top \theta_\star) Z_{i,s}\right\|_{\mathbf{G}_t}.$$

Using the definition of  $M_s$ , where  $\tilde{a}_s(M_s) = N_s$ , we have

$$\begin{aligned} &\mathbb{P}\left(\left\{F(\tilde{a}_1(M_1), \dots, \tilde{a}_t(M_t)) > x\right\} \cap \mathcal{S}_t\right) \\ &= \mathbb{P}\left(\left\{F(\tilde{a}_1(M_1), \dots, \tilde{a}_t(M_t)) > x\right\} \cap \mathcal{S}_t \cap \bigcap_{s=1}^t \{\tilde{a}_s(M_s) = N_s\}\right) \\ &= \mathbb{P}\left(\left\{F(\tilde{a}_1(1), \dots, \tilde{a}_t(1)) > x\right\} \cap \mathcal{S}_t \cap \bigcap_{s=1}^t \{\tilde{a}_s(1) = N_s\}\right), \end{aligned}$$

where the last equality holds because  $\{\tilde{a}_s(m) : m \in \mathbb{N}\}$  are IID for each  $s \in [t]$ . Then,

$$\begin{aligned} &\mathbb{P}\left(\left\{F(\tilde{a}_1(1), \dots, \tilde{a}_t(1)) > x\right\} \cap \mathcal{S}_t \cap \bigcap_{s=1}^t \{\tilde{a}_s(1) = N_s\}\right) \\ &\leq \mathbb{P}\left(\left\{F(\tilde{a}_1(1), \dots, \tilde{a}_t(1)) > x\right\} \cap \mathcal{S}_t\right) \end{aligned}$$

Replacing  $\{\tilde{a}_s(M_s) : s \in [t]\}$  in coupled pseudo-rewards (19) with  $\{\tilde{a}_s(1) : s \in [t]\}$  gives the hypothetical pseudo-rewards in (16). Thus, the distribution of the normalized error  $\|\tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star\|_{\mathbf{G}_t}$  is equivalent to that of  $F(\tilde{a}_1(1), \dots, \tilde{a}_t(1))$  and we obtain

$$\mathbb{P}\left(\{F(\tilde{a}_1(1), \dots, \tilde{a}_t(1)) > x\} \cap \mathcal{S}_t\right) = \mathbb{P}\left(\{\|\tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star\|_{\mathbf{G}_t} > x\} \cap \mathcal{S}_t\right),$$

which proves the coupling inequality. The bound for the failure probability  $\mathbb{P}(\mathcal{S}_t^c) \leq \delta$  is proved by the fact that  $\mathbb{P}(\mathcal{M}_s^c) \leq \delta/(s+1)^2$  by construction of the maximum number of resampling trials.  $\blacksquare$

With the coupling inequality, we can leverage augmented samples from the hypothetical bandit problem to closely approximate the hypothetical sample augmented estimator with high probability. While the coupling technique can be applied to any choice of hypothetical problem — and the hypothetical contexts  $\{Z_{i,t}\}$  may be arbitrarily defined — it is crucial that they are compatible with the original contextual bandit problem and the reference estimator  $\check{\theta}_t$ .

The design of suitable hypothetical contexts is key, as it enables control of the maximum deviation in the original problem through the bound:

$$\max_{k \in [K]} \left| X_{k,t}^\top (\tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star) \right| \leq \left\| \tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star \right\|_{\tilde{G}_t} \cdot \max_{k \in [K]} \|X_{k,t}\|_{\tilde{G}_t^{-1}},$$

where  $\tilde{G}_t := \sum_{s=1}^t \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top$  denotes the Gram matrix constructed from the hypothetical contexts.

This upper bound consists of two components: (i) the self-normalized error of the hypothetical sample augmented estimator, and (ii) the maximum norm of the original contexts normalized by  $\tilde{G}_t$ . Each component is sensitive to how the Gram matrix  $\tilde{G}_t$  is constructed.

If  $\tilde{G}_t$  is defined using only the played contexts — i.e.,  $\tilde{G}_t = \sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + I_d$  — then the norm term  $\max_{k \in [K]} \|X_{k,t}\|_{\tilde{G}_t^{-1}}$  may become unbounded due to insufficient exploration. Conversely, if  $\tilde{G}_t$  is constructed from an overly large set of hypothetical contexts, the self-normalized error  $\|\tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star\|_{\tilde{G}_t}$  may increase significantly, making it harder to guarantee tight estimation bounds.

Therefore, the design of the hypothetical contexts must carefully balance the number of augmenting samples to ensure both the estimator's accuracy and the tightness of high-probability bounds. Section 6.1 presents a formal analysis of this trade-off, and shows that the proposed construction achieves this balance effectively, leading to a well-conditioned estimator with high-probability regret guarantees.

## 5.5 Hypothetical Coupled Sample Augmented Thompson Sampling

The proposed algorithm, *Hypothetical Coupled Sample Augmented Thompson Sampling* (HCSA+TS), is detailed in Algorithm 2. This algorithm builds upon the structure of LinTS but introduces two key innovations: (i) resampling to couple the hypothetical bandit with the original bandit, and (ii) the HCSA estimator, which leverages a compressed orthogonal basis for improved efficiency.

**Algorithm 2** Hypothetical Coupled Sample Augmented Thompson Sampling (HCSA+TS)

---

```

1: Input: confidence level  $\delta \in (0, 1)$ , coupling parameter  $\gamma \in (0, 1)$ , exploration parameter
    $\mathbf{G}_t := \{2 \log \frac{K(t+1)^2}{\delta}\}^{-1/2}$ , orthogonal basis regularization parameter  $h_t$  as in (9).
2: Initialize the estimator  $\hat{\theta}_0 = \mathbf{0}$ , Gram matrix  $V_0 = O$ , and a subset of rounds for the
   orthogonal basis regularization  $\mathcal{A}_0 = \emptyset$ .
3: for  $t = 1$  to  $T$  do
4:   Observe contexts  $\{X_{k,t} : k \in [K]\}$ .
5:   Update  $\mathcal{A}_t$  as in (10) and compute  $N_t$  as in (12).
6:   Set  $m = 1$  sample  $\tilde{a}_t(m)$  from the multinomial distribution (14).
7:    $(a_t(m), \tilde{a}_t(m)) \leftarrow \text{CAS}(\{X_{k,t}\}, \hat{\theta}_{t-1}, v_{t-1}, V_{t-1}, N_t, \gamma, \delta)$ 
8:   if  $\tilde{a}_t(m) = N_t$  // Resampling succeeded then
9:     Pull arm  $a_t(m)$  and observe  $Y_{a_t(m),t}$ .
10:    Compute the reference estimator  $\hat{\theta}_t$  defined in (15).
11:    if  $t \in \mathcal{A}_t$  then
12:      Compute hypothetical contexts  $Z_{i,t} := H_{i,t}$  as in (13)
13:    else
14:      Compute hypothetical contexts  $Z_{i,t} := \tilde{H}_{i,t}$  with orthogonal basis as in (13).
15:    end if
16:    Update  $\mathbf{G}_t = V_{t-1} + \sum_{i=1}^{N_t} Z_{i,s} Z_{i,s}^\top$ 
17:    Compute the estimator  $\hat{\theta}_t$  as in (20).
18:  else
19:     $\hat{\theta}_t \leftarrow \hat{\theta}_{t-1}$ 
20:  end if
21: end for

```

---

For the resampling step (i), the algorithm repeatedly resamples  $a_t$  from LinTS policy equipped with HCSA estimator and the novel Gram matrix until the condition  $\{Z_{\tilde{a}_t,t} = X_{a_t,t}\}$  is satisfied. This ensures that the hypothetical bandit aligns with the original bandit by augmenting the randomized contexts. The number of resampling attempts, set to  $\lceil \log \frac{(t+1)^2}{\delta} / \log \frac{1}{1-\gamma} \rceil$ , guarantees the resampling process succeeds with probability at least  $1 - \delta/(t+1)^2$ . In practice, this resampling typically succeeds after only a few iterations.

For the HCSA estimator (ii), although computing it might appear computationally demanding, efficient implementation strategies significantly reduce its complexity. Theoretically, the algorithm must compute hypothetical context vectors  $\{Z_{i,t} : i \in [N_t]\}$  for each resampling iteration. However, in practice, the algorithm first checks if  $\{\tilde{a}_t = N_t\}$  occurs and then computes the hypothetical contexts based on the resampled  $a_t$ .

The worst-case computational complexity of the algorithm is  $O(d^2(K+d)T + T \log(T+1) / \log(\frac{1}{1-\gamma}))$ . The primary computational bottleneck arises from calculating the Gram matrix  $\mathbf{G}_t$  and performing eigenvalue decomposition to construct hypothetical contexts, repeated for the specified number of resampling attempts in each round  $t \in [T]$ . In practice, the computational efficiency of the A estimator can be further enhanced by applying the Sherman-Morrison formula, enabling rank-1 updates and reducing memory usage.

## 6 Regret Analysis

The following theorem establishes a nearly minimax-optimal cumulative regret bound for the HCSA+TS algorithm.

**Theorem 6 (Regret Bound for HCSA+TS)** *In Algorithm 2, set the exploration parameter as  $\mathbf{G}_t = \{2 \log(K(t+1)^2/\delta)\}^{-1/2}$ . Then, with probability at least  $1 - 3\delta$ , the cumulative regret of the HCSA+TS algorithm by round  $T$  satisfies:*

$$R(T) \leq \frac{4d}{(1/2 - e^{-1})(1 - \gamma)} \log \frac{d(T+1)^2}{\delta} + \left\{ 20\sqrt{2}\theta_{\max} + \frac{21\sigma}{\gamma} \sqrt{2d \log \frac{2T}{\delta}} + 1 \right\} \sqrt{8dT \log \frac{T}{d}}. \quad (21)$$

The leading order term of the regret bound is  $O(d\sqrt{T} \log T)$ , which matches the minimax lower bound  $\Omega(d\sqrt{T})$  established in Lattimore and Szepesvári (2020), up to logarithmic factors. This result provides the first nearly minimax-optimal regret guarantee for linear contextual bandits under arbitrary context distributions.

For comparison, Kim et al. (2021) achieved  $\tilde{O}(d\sqrt{T})$  regret bound under IID contexts with some special distributions of which the minimum eigenvalue of the average covariance matrix is  $\Omega(1/d)$ . Kim et al. (2023b) achieved a regret bound of  $O(\sqrt{dT \log T})$  under the assumption of IID contexts with a strictly positive minimum eigenvalue for their average covariance matrix. Similarly, Huix et al. (2023) obtained a regret bound of  $\tilde{O}(d\sqrt{T})$ , but their analysis assumes a Gaussian prior on the unknown parameter  $\theta_*$ .

Earlier works such as Kim et al. (2021) and Agrawal and Goyal (2013) impose unit-norm assumptions on the contexts. In contrast, the regret bound (21) does not require normalization of the context vectors. The only assumption made is that the absolute inner product between any context and the true parameter is bounded, i.e.,  $|X_{k,t}^\top \theta_*| \leq 1$ . Notably, increasing the context norm bound  $x_{\max}$  does not cause the regret to grow linearly, avoiding the scaling issues encountered in prior analyses.

The key technical contributions enabling this result include: (i) the development of a self-normalized bound for the HCSA estimator using a carefully constructed Gram matrix (Section 6.1), (ii) the identification of a set of low-regret arms selected with high probability (Section 6.2), and (iii) a novel maximal elliptical potential bound based on the augmented Gram matrix  $\mathbf{G}_t$  (Section 6.3).

### 6.1 A Self-Normalized Bound for the Proposed Estimator

With the coupling inequality (Lemma 5), we can bound the error of the proposed estimator by obtaining an error bound for the hypothetical estimator (17), which is proven in the following lemma.

**Lemma 7 (A self-normalized bound of the HSA estimator)** *For each  $t \geq 1$ , define the matrix  $A_t := \sum_{s=1}^t \phi_{\tilde{a}_{s,s}}^{-1} Z_{\tilde{a}_{s,s}} Z_{\tilde{a}_{s,s}}^\top$ . and. Then the self-normalized bound of the hypothetical sample augmented estimator is decomposed as:*

$$\left\| \tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_* \right\|_{\mathbf{G}_t} \leq \left\| \mathbf{G}_t^{-1/2} (\mathbf{G}_t - A_t) (\check{\theta}_t - \theta_*) \right\|_2 + \left\| \sum_{s=1}^t \phi_{\tilde{a}_{s,s}}^{-1} (W_{\tilde{a}_{s,s}} - Z_{\tilde{a}_{s,s}}^\top \theta_*) Z_{\tilde{a}_{s,s}} \right\|_{\mathbf{G}_t}$$

The proof is in Appendix A.1. The decomposition shows the two sources of error for the HDR estimator for rewards of all arms: (i) from the reference estimator  $\tilde{\theta}$  used as an reference estimator in HDR pseudo-rewards (16), and (ii) the noise error of the rewards. In error term (i),  $\mathbf{G}_t(\mathbf{G}_t - A_t)$  is the matrix martingale with bounded eigenvalues, which is bounded by the newly developed matrix concentration inequality (Lemma 14). The error term (ii) is bounded by a modified martingale inequality developed by Abbasi-Yadkori et al. (2011).

With the suitable choice of  $\tilde{\theta}_t$  as in (15), we obtain an  $O(\sqrt{d \log t})$  error bound for  $\tilde{\theta}_t^{H(\tilde{\theta}_t)}$  estimator, which is normalized by the novel augmented Gram matrix  $\mathbf{G}_t$  instead of the conventional Gram matrix that includes only selected contexts.

**Theorem 8 (A Self-Normalized Bound for the HCSA Estimator)** *The HCSA estimator (20) satisfies,*

$$\|\hat{\theta}_t - \theta_\star\|_{\mathbf{G}_t} \leq 20\theta_{\max} + \frac{21\sigma}{\gamma} \sqrt{d \log \frac{1+t}{\delta}},$$

with probability at least  $1 - 3\delta$  for all  $t \geq T_1$ .

The proof is provided in Appendix A.2. Unlike the conventional self-normalized bound of Abbasi-Yadkori et al. (2011), which is normalized by the Gram matrix  $\sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + I_d$  built solely from selected contexts, Theorem 8 establishes a bound normalized by the full Gram matrix  $\mathbf{G}_t$ , which includes contexts from all  $K$  arms. While Kim et al. (2023b) also considered self-normalization with respect to a full Gram matrix, their estimator incorporates contexts from all arms only in a fraction of the rounds, and their analysis is restricted to IID contexts with a strictly positive-definite covariance matrix. In contrast, Theorem 8 applies to arbitrary (including non-IID, non-stationary) context sequences, establishing a uniform self-normalized bound with a full Gram matrix. This result enables a novel regret analysis of Thompson Sampling that yields a  $\tilde{O}(d\sqrt{T})$  bound for HCSA+TS, applicable under arbitral context distributions.

## 6.2 Low-Regret Arms with a High-Probability Guarantee

For each  $k \in [K]$  and  $t \in [T]$ , define the instantaneous regret gap between the optimal arm and arm  $k$  as

$$\Delta_{k,t} := X_{a_t^\star,t}^\top \theta_\star - X_{k,t}^\top \theta_\star.$$

Using this, we define the set of low-regret arms at round  $t$  as

$$\mathcal{P}_t := \left\{ k \in [K] : \Delta_{k,t} \leq 2x_t \|\hat{\theta}_{t-1} - \theta_\star\|_{\mathbf{G}_{t-1}} + \sqrt{\|X_{a_t^\star,t}\|_{\mathbf{G}_{t-1}}^2 + \|X_{k,t}\|_{\mathbf{G}_{t-1}}^2} \right\}, \quad (22)$$

where  $x_t := \max_{k \in [K]} \|X_{k,t}\|_{\mathbf{G}_{t-1}}$ . The self-normalized confidence bound with respect to the hypothetical Gram matrix  $\mathbf{G}_{t-1}$ , which includes contexts beyond the selected arms, allows the construction of an effective set  $\mathcal{P}_t$  that has lower regret than arms in sets built from Gram matrices based solely on selected contexts such as  $\sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + I_d$ .

The following lemma provides a high-probability guarantee that the arm selected by Algorithm 2 belongs to the low-regret set.

**Lemma 9 (High-Probability Selection of Low-Regret Arms)** *Let  $a_t$  be the arm selected by Algorithm 2, and let  $\mathcal{P}_t$  be the set defined in (22). If the exploration parameter is set as  $\mathbf{G}_t = \{2\log(K(t+1)^2/\delta)\}^{-1/2}$ , then*

$$\mathbb{P}(a_t \in \mathcal{P}_t \mid \mathcal{H}_t) \geq 1 - \frac{\delta}{(t+1)^2}.$$

The proof is deferred to Appendix A.4. In contrast to Agrawal and Goyal (2013), where bounding the probability of selecting a saturated arm required setting  $v = \sqrt{9d\log(t/\delta)}$ , thereby introducing a  $\sqrt{d}$ -scaling, Lemma 9 establishes that such dimensional dependence is unnecessary. By leveraging the structure of the HCSA estimator and its associated Gram matrix, the proposed approach guarantees high-probability selection of low-regret arms without incurring additional dependence of the variance parameter  $v$  on the dimension  $d$ .

### 6.3 Maximal Elliptical Potential Bound

While continuing the proof of the regret bound, we face a novel terms that needs to be analyzed. By Lemma 9, the algorithm selects arms from the low-regret set  $\mathcal{P}_t$ , i.e.,  $a_t \in \mathcal{P}_t$ , with high probability. This leads to the following novel regret decomposition:

$$\text{regret}(t) = \Delta_{a_t, t} \leq 2x_t \left\| \widehat{\theta}_{t-1} - \theta_\star \right\|_{\mathbf{G}_{t-1}} + \sqrt{\|X_{a_t^\star, t}\|_{\mathbf{G}_{t-1}^{-1}}^2 + \|X_{a_t, t}\|_{\mathbf{G}_{t-1}^{-1}}^2},$$

for  $t \geq T_1$ . Thus, the cumulative regret is bounded as follows:

$$R(T) \leq 2h_T + \sum_{t \in [T] \setminus \mathcal{A}_T} \left\{ 2x_t \left\| \widehat{\theta}_{t-1} - \theta_\star \right\|_{\mathbf{G}_{t-1}} + \sqrt{\|X_{a_t^\star, t}\|_{\mathbf{G}_{t-1}^{-1}}^2 + \|X_{a_t, t}\|_{\mathbf{G}_{t-1}^{-1}}^2} \right\}.$$

Because  $x_t := \max_{k \in [K]} \|X_{k, t}\|_{V_{t-1}^{-1}}$ , we get:

$$R(T) \leq 2h_T + \sum_{t \in [T] \setminus \mathcal{A}_T} (2x_t \|\widehat{\theta}_{t-1} - \theta_\star\|_{\mathbf{G}_{t-1}} + \sqrt{2}x_t).$$

Factoring out  $x_t$ :

$$R(T) \leq 2T_1 + \sum_{t \in [T] \setminus \mathcal{A}_T} (2\|\widehat{\theta}_{t-1} - \theta_\star\|_{\mathbf{G}_{t-1}} + \sqrt{2})x_t.$$

Since we obtain  $\|\widehat{\theta}_{t-1} - \theta_\star\|_{V_{t-1}} = O(\sqrt{d\log t})$  from Theorem 8, we need a bound for  $\sum_{t \in [T] \setminus \mathcal{A}_T} x_t$ , which is in the following lemma.

**Lemma 10 (Maximal elliptical potential lemma)** *For  $x_t := \max_{k \in [K]} \|X_{k, t}\|_{\mathbf{G}_{t-1}^{-1}}$ , we have:*

$$\sum_{t=T_1}^T x_t^2 \leq 4d \log \frac{T}{d}.$$

The proof is in Appendix A.5. Previous elliptical lemmas only obtain the bounds the normalized norm of the selected contexts  $X_{a_t,t}$ , while we need a bound for  $\max_{k \in K} \|X_{k,t}\|_{\mathbf{G}_{t-1}^{-1}}$ . The maximal elliptical potential lemma bounds the normalized norm of the contexts for all  $K$  arms, which previous analyses could not bound effectively. This is possible because we augmented suitable sample to obtain the Gram matrix  $\mathbf{G}_t$  consisting of contexts from all  $K$  arms.

## 7 Experimental Results

This section is for evaluating the empirical performance of the proposed algorithm, HCSA+TS, against several benchmark algorithms for LinCB with simulated data. The benchmarks include **LinTS**, **LinUCB**, **DRTS** (Kim et al., 2021), **HyRan** (Kim et al., 2023b), and **SupLinUCB** (Chu et al., 2011).

For the experiment setting, the parameter  $\theta_*$  is defined as:

$$\theta_* := \frac{1}{\sqrt{d}} \left( \underbrace{1, \dots, 1}_{\lceil d/2 \rceil}, \underbrace{-1, \dots, -1}_{d-\lceil d/2 \rceil} \right)^\top,$$

where  $d \in \{10, 30\}$  is the dimension of the parameter. The  $i$ -th entry of the context vectors for  $K \in \{20, 30\}$  arms are independently sampled from a Gaussian distribution with a mean of  $-1 + \frac{3(i-1)}{d-1}$  and variance of 1 for each  $i \in [d]$ . These vectors are normalized and then scaled by a scalar drawn uniformly from  $[0, 1]$ . To simulate missing context information, with probability 1/2, the last  $d - \lceil d/2 \rceil$  entries of the context vectors are set to zero at each round. This setting reflects practical scenarios where certain context features may be unavailable with some probability, making it challenging to estimate the corresponding entries in  $\theta_*$ .

The hyperparameter optimization was conducted as follows: For **LinTS**, the variance parameter was selected from  $\{0.01, 0.1, 1\}$ . For **LinUCB** and **SupLinUCB**, the confidence bound inflation parameter was chosen from  $\{0.01, 0.1, 1\}$ . For **HyRan** and **HCSA+TS**, the regularization parameters  $p$  and  $\gamma$  (only  $\gamma$  for **HCSA+TS**) were tuned from  $\{0.1, 0.5, 0.9\}$ . The hyperparameters for **DRTS** were fixed as specified in Kim et al. (2021). Values outside the specified ranges showed negligible differences in performance, suggesting robustness to hyperparameter selection for all methods.

Figure 2 compares the cumulative regret of **HCSA+TS** with other benchmark algorithms across various configurations of  $d$  and the number of arms  $K$ . Each line represents the average cumulative regret, and the shaded regions indicate the standard deviation across 20 independent trials.

The results show that **HCSA+TS** achieves the lowest cumulative regret in all tested settings. Compared to **LinTS**, **LinUCB**, and **SupLinUCB**, which do not leverage information from all arms, **HCSA+TS** demonstrates robustness to missing context data. When compared to **DRTS** and **HyRan**, which uses the original context vectors, **HCSA+TS** consistently identifies low-regret arms more effectively, even under significant masking of context features.

Initially, due to the orthogonal basis regularization, **HCSA+TS** incurs higher regret during the exploration phase, particularly when the effective rank of the context matrix is low. However, it rapidly adapts and identifies the optimal arm, ultimately outperforming the

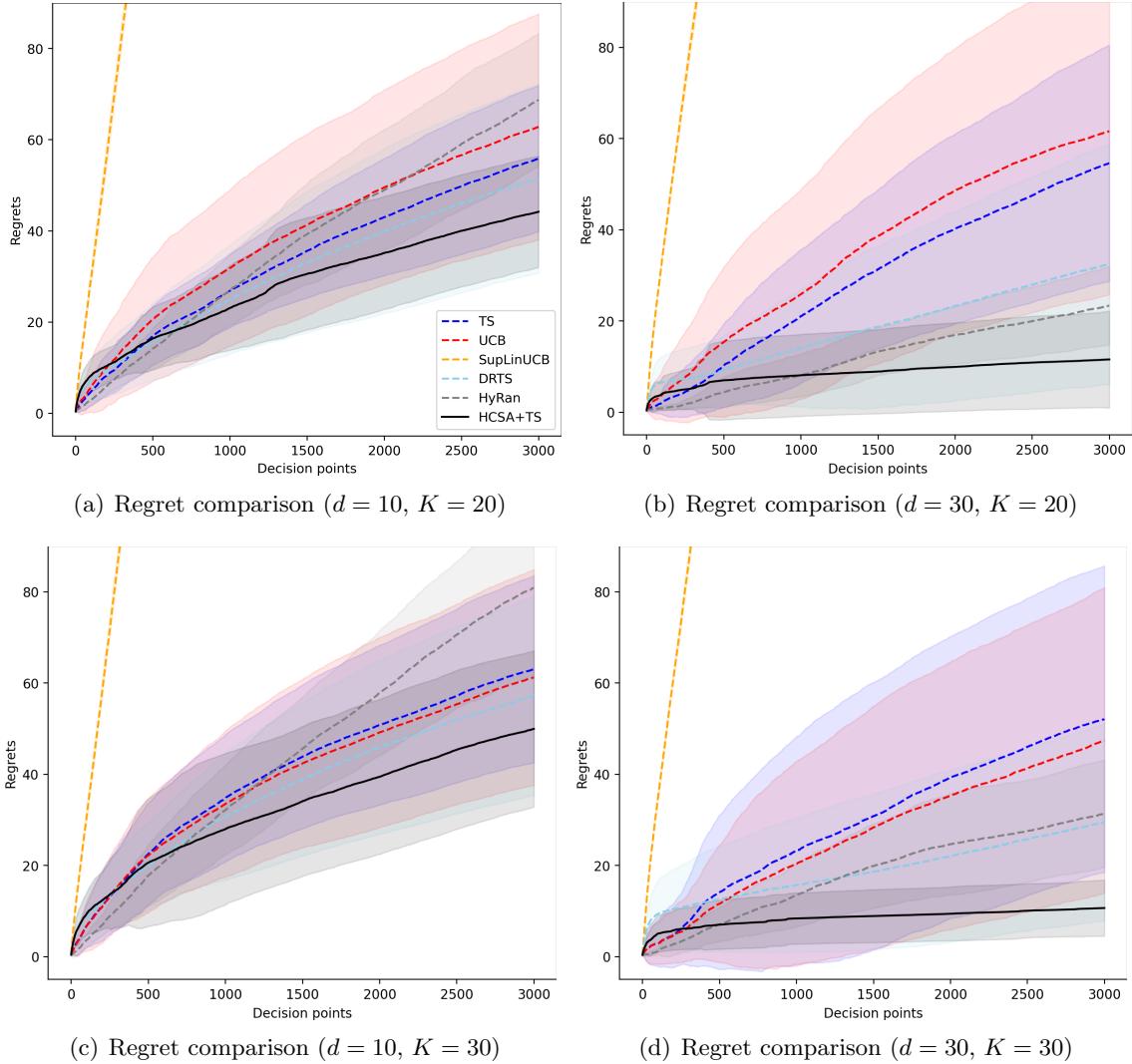


Figure 2: Comparison of the regrets of the proposed HCSA+TS algorithm with other benchmark methods. The lines represent the average, and the shaded areas indicate the standard deviation based on twenty experiments. The results demonstrate that the proposed HCSA+TS effectively identifies the optimal arm using orthogonal regularization.

other algorithms, which continue to suffer regret due to their inability to handle missing context data effectively.

Figure 3 illustrates the prediction error across all arms, measured as  $\sum_{i=1}^K \{X_{i,t}^\top (\hat{\theta}_t - \theta_\star)\}^2$ . Similar to the regret results, the averages and standard deviations are computed over 20 trials. The initial convergence of the estimators in DRTS and HyRan is faster due to their reliance on imputed contexts. However, their prediction errors increase over time because the imputed contexts, often containing many zero entries, provide incomplete information

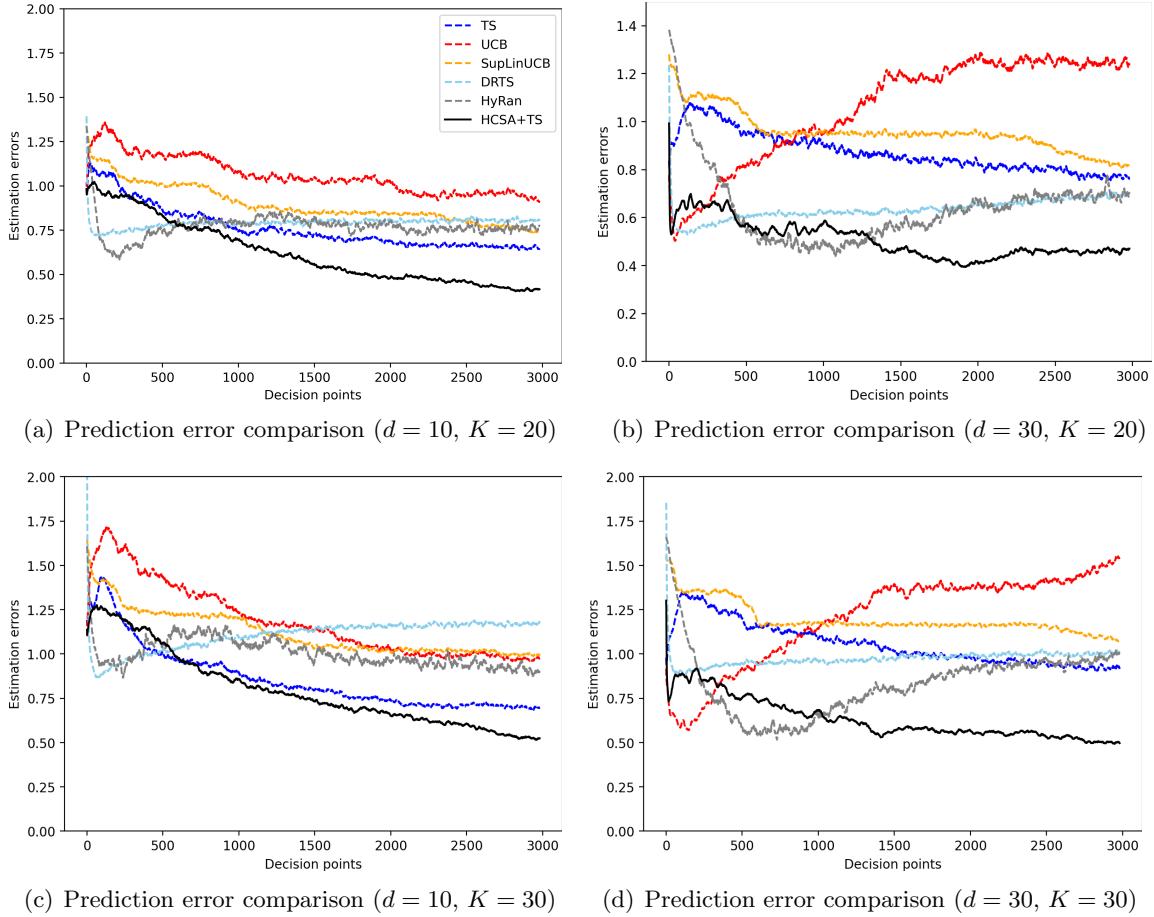


Figure 3: Comparison of the prediction error across all arms, calculated as  $\sum_{i=1}^K \{X_{i,t}^\top (\hat{\theta}_t - \theta_\star)\}^2$ , for the proposed HCSA+TS and other benchmark methods. The lines represent the averages, and the shaded areas indicate the standard deviations based on twenty experiments. The results demonstrate that the proposed estimator, enhanced with orthogonal augmentation, learns the reward more accurately than other estimators.

and hinder accurate estimation. In contrast, HCSA+TS demonstrates steady and consistent convergence throughout the experimental horizon. Its orthogonal augmentation strategy allows it to extract useful information even when parts of the context vectors are masked, leading to superior prediction accuracy compared to both traditional ridge-based estimators (LinTS, LinUCB, SupLinUCB) and other augmented methods (DRTS, HyRan).

The experiments validate that HCSA+TS achieves significant improvements in both cumulative regret and prediction accuracy over existing benchmarks. This shows the efficacy of the generalization and convergence on rewards for all arms of the proposed HCSA estimator due to the adaptive augmentation with carefully constructed hypothetical contexts and rewards.

## 8 Conclusion

This work introduces novel resampling and coupling techniques for integrating adaptive data augmentation into linear contextual bandits (LinCB), achieving a nearly minimax-optimal regret bound without imposing structural assumptions on the context vectors. By leveraging these techniques, the proposed approach enables the algorithm to effectively utilize contextual information from all arms as if full reward feedback were available in each round, while minimizing the number of required augmentations. Beyond improving regret performance, this methodology marks a conceptual advancement in reward estimation. Through the use of hypothetical contexts and coupling-based resampling, we show that accurate estimation for rewards of all arms is feasible even under arbitrary context distributions, thereby broadening the applicability of LinCB methods in non-IID and adversarial settings.

The proposed framework offers a foundation for extending adaptive augmentation techniques to more general bandit models and reinforcement learning environments. In particular, these ideas have the potential to reduce variance in reward estimates for optimal arms or general policies, thus enabling more efficient exploration and decision-making in complex and high-dimensional settings. Future work may explore such extensions, including applications to nonlinear bandits, structured decision-making problems, and policy optimization in reinforcement learning, further amplifying the impact of the techniques developed in this study.

## Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341], Artificial Intelligence Graduate School Program (Chung-Ang University) and the Chung-Ang University Research Grants in 2025. Wonyoung Kim also appreciates the proofreading and advising provided by Myunghee Cho Paik.

## Appendix A. Missing Proofs

### A.1 Proof of Lemma 7

**Proof** Recall that

$$\begin{aligned} \mathbf{G}_t &:= \sum_{s=1}^t \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top, \\ A_t &:= \sum_{s=1}^t \frac{1}{\phi_{\tilde{a}_s, s}} Z_{\tilde{a}_s, s} Z_{\tilde{a}_s, s}^\top = \sum_{s=1}^t \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} Z_{i,s} Z_{i,s}^\top. \end{aligned}$$

By definition of the estimator,

$$\begin{aligned}
 & \left\| \tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star \right\|_{\mathbf{G}_t} \\
 &= \left\| \sum_{s=1}^t \sum_{i=1}^{r_s+1} \left( \tilde{W}_{i,s}^{H(\check{\theta}_t)} - Z_{i,s}^\top \theta_\star \right) Z_{i,s} \right\|_{\mathbf{G}_t^{-1}} \\
 &= \left\| \sum_{s=1}^t \sum_{i=1}^{r_s+1} \left\{ 1 - \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \right\} Z_{i,s} Z_{i,s}^\top (\check{\theta}_t - \theta_\star) + \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \left( W_{i,s} - Z_{i,s}^\top \theta_\star \right) Z_{i,s} \right\|_{\mathbf{G}_t^{-1}} \\
 &= \left\| (\mathbf{G}_t - A_t)(\check{\theta}_t - \theta_\star) + \sum_{s=1}^t \sum_{i=1}^{r_s+1} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \left( W_{i,s} - Z_{i,s}^\top \theta_\star \right) Z_{i,s} \right\|_{\mathbf{G}_t^{-1}},
 \end{aligned}$$

where the second equality holds by the definition (16). Recall that

Then, by the triangle inequality, we have

$$\left\| \tilde{\theta}_t^{H(\check{\theta}_t)} - \theta_\star \right\|_{\mathbf{G}_t} \leq \|(\mathbf{G}_t - A_t)(\check{\theta}_t - \theta_\star)\|_{\mathbf{G}_t^{-1}} + \left\| \sum_{s=1}^t \sum_{i=1}^{r_s+1} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \left( W_{i,s} - Z_{i,s}^\top \theta_\star \right) Z_{i,s} \right\|_{\mathbf{G}_t^{-1}}.$$

Because

$$\|(\mathbf{G}_t - A_t)(\check{\theta}_t - \theta_\star)\|_{\mathbf{G}_t^{-1}} = \left\| \mathbf{G}_t^{-1/2} (\mathbf{G}_t - A_t)(\check{\theta}_t - \theta_\star) \right\|_2,$$

which completes the proof.  $\blacksquare$

## A.2 Proof of Theorem 8

**Proof** By the definition of the proposed estimator and the coupling inequality (Lemma 5), for any  $x > 0$ ,

$$\mathbb{P} \left( \left\| \widehat{\theta}_t - \theta_\star \right\|_{\mathbf{G}_t} > x \right) \leq \mathbb{P} \left( \left\{ \left\| \widehat{\theta}_t^{HDR(\check{\theta}_t)} - \theta_\star \right\|_{\mathbf{G}_t} > x \right\} \cap \mathcal{S}_t \right) + \delta.$$

From the error decomposition in Lemma 7, for  $x > 0$ ,

$$\begin{aligned}
 & \mathbb{P} \left( \left\{ \left\| \widehat{\theta}_t^{HDR(\check{\theta}_t)} - \theta_\star \right\|_{\mathbf{G}_t} > x \right\} \cap \mathcal{S}_t \right) \\
 & \leq \mathbb{P} \left( \left\{ \left\| \mathbf{G}_t^{-1/2} (A_t - \mathbf{G}_t)(\check{\theta}_t - \theta_\star) \right\|_2 + \|S_t\|_{\mathbf{G}_t^{-1}} > x \right\} \cap \mathcal{S}_t \right),
 \end{aligned}$$

where

$$S_t := \sum_{s=1}^t \sum_{i=1}^{r_s+1} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \left( W_{i,s} - Z_{i,s}^\top \theta_\star \right) Z_{i,s}.$$

By the definition of  $\check{\theta}_t$ ,

$$\check{\theta}_t - \theta_\star := \left( \sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + \gamma I_d \right)^{-1} \left\{ \sum_{s=1}^t (Y_{a_s,s} - X_{a_s,s}^\top \theta_\star) X_{a_s,s} - \gamma \theta_\star \right\},$$

and similarly,

$$\check{\theta}_t - \theta_\star = (A_t + I_d)^{-1} (S_t - \theta_\star).$$

Substituting into the error bound, we have

$$\begin{aligned} \|\hat{\theta}_t - \theta_\star\|_{\mathbf{G}_t} &\leq \|(\mathbf{G}_t - A_t)(A_t + I_d)^{-1}(S_t - \theta_\star)\|_{\mathbf{G}_t^{-1}} + \|S_t\|_{\mathbf{G}_t^{-1}} \\ &= \left\| \mathbf{G}_t^{-1/2} (\mathbf{G}_t - A_t)(A_t + I_d)^{-1}(S_t - \theta_\star) \right\|_2 + \|S_t\|_{\mathbf{G}_t^{-1}}, \end{aligned}$$

where the last equality holds by the definition of  $\|\cdot\|_{\mathbf{G}_t^{-1}}$ . By the definition of the spectral norm  $\|\cdot\|_2$ ,

$$\begin{aligned} &\left\| \mathbf{G}_t^{-1/2} (\mathbf{G}_t - A_t)(A_t + I_d)^{-1}(S_t - \theta_\star) \right\|_2 \\ &\leq \underbrace{\left\| \mathbf{G}_t^{-1/2} (\mathbf{G}_t - A_t)(A_t + I_d)^{-1} \mathbf{G}_t^{1/2} \right\|_2}_{P_t} \|S_t - \theta_\star\|_{\mathbf{G}_t^{-1}}. \end{aligned}$$

Thus, the bound becomes

$$\|\hat{\theta}_t - \theta_\star\|_{\mathbf{G}_t} \leq (\|P_t\|_2 + 1) \|S_t\|_{\mathbf{G}_t^{-1}} + \|P_t\|_2 \theta_{\max} \lambda_{\min}(\mathbf{G}_t)^{-1/2}.$$

By the property of  $\mathbf{G}_t$  (Lemma 4)  $\lambda_{\min}(\mathbf{G}_t) \geq \max\{x_{\max}^2, 1\} h_t \geq 1$ . It follows that

$$\|\hat{\theta}_t - \theta_\star\|_{\mathbf{G}_t} \leq (\|P_t\|_2 + 1) \|S_t\|_{\mathbf{G}_t^{-1}} + \|P_t\|_2 \theta_{\max}.$$

Thus,

$$\begin{aligned} &\mathbb{P} \left( \bigcap_{t:|\mathcal{A}_t| \geq h_t} \left[ \left\{ \|\hat{\theta}_t - \theta_\star\|_{\mathbf{G}_t} > x \right\} \cap \mathcal{S}_t \right] \right) \\ &\leq \mathbb{P} \left( \bigcap_{t:|\mathcal{A}_t| \geq h_t} \left[ \left\{ (\|P_t\|_2 + 1) \|S_t\|_{\mathbf{G}_t^{-1}} + \|P_t\|_2 \theta_{\max} > x \right\} \cap \mathcal{S}_t \right] \right) + \delta. \end{aligned}$$

By Lemma 11, we have  $\mathbb{P}(\|P_t\|_2 > 5) \leq \delta/t^2$  for all  $t$  such that  $|\mathcal{A}_t| \geq h_t$ . Thus,

$$\begin{aligned} &\mathbb{P} \left( \bigcap_{t:|\mathcal{A}_t| \geq h_t} \left[ \left\{ (\|P_t\|_2 + 1) \|S_t\|_{\mathbf{G}_t^{-1}} + \|P_t\|_2 \theta_{\max} > x \right\} \cap \mathcal{S}_t \right] \right) + \delta \\ &\leq \mathbb{P} \left( \bigcap_{t:|\mathcal{A}_t| \geq h_t} \left[ \left\{ 21 \|S_t\|_{\mathbf{G}_t^{-1}} + 20\theta_{\max} > x \right\} \cap \mathcal{S}_t \right] \right) + \mathbb{P} \left( \cup_{t:|\mathcal{A}_t| \geq h_t} \{\|P_t\| > 20\} \right) + \delta \\ &\leq \mathbb{P} \left( \bigcap_{t:|\mathcal{A}_t| \geq h_t} \left[ \left\{ \|S_t\|_{\mathbf{G}_t^{-1}} > \frac{x}{21} - \frac{20}{21}\theta_{\max} \right\} \cap \mathcal{S}_t \right] \right) + 2\delta. \end{aligned}$$

By Lemma 12, setting  $x = 20\theta_{\max} + \frac{21\sigma}{\gamma}\sqrt{d \log \frac{1+t}{\delta}}$  completes the proof of the theorem. ■

### A.3 Probability Bounds for the Norms

This section provides the probability inequalities for the norms of the core terms:  $P_t$  and  $S_t$ . These bounds are key to proving the convergence of the estimator. We will use tail inequalities for random matrices and sums to obtain the desired results.

**Lemma 11** Suppose  $t \geq T_1$ . Then, with probability at least  $1 - \frac{\delta}{t^2}$ , the spectral norm  $\|P_t\|_2 \leq 5$ .

**Proof** By the definition of  $P_t$ , we have

$$\|P_t\|_2 = \|\mathbf{G}_t^{-1/2}(\mathbf{G}_t - A_t)(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2.$$

Expanding this,

$$\|P_t\|_2 = \|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2} - \mathbf{G}_t^{-1/2}A_t(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2.$$

Next, we simplify further:

$$\|P_t\|_2 = \|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2} - I_d + \mathbf{G}_t^{-1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2.$$

This gives us the bound:

$$\|P_t\|_2 \leq \|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2} - I_d\|_2 + \|\mathbf{G}_t^{-1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2.$$

Since  $\mathbf{G}_t$  and  $A_t$  are real symmetric and positive semidefinite, we have the following:

$$\|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2 \leq \frac{1}{\lambda_{\min}(\mathbf{G}_t)}\|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2.$$

Thus, we obtain:

$$\|P_t\|_2 \leq \|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2} - I_d\|_2 + \frac{1}{\lambda_{\min}(\mathbf{G}_t)}\|\mathbf{G}_t^{1/2}(A_t + I_d)^{-1}\mathbf{G}_t^{1/2}\|_2.$$

Let

$$\mathbf{F}_t := \bar{x}_{\max}^2 \left( \sum_{s \in [t] \setminus \mathcal{A}_t} \sum_{i=1}^{r_s} u_{i,s} u_{i,s}^\top + \sum_{s \in [t] \setminus \mathcal{A}_t} \sum_{i=1}^d u_{i,s} u_{i,s}^\top \right).$$

By the equivalence lemmas (Lemma ?? and Lemma ??)  $\mathbf{G}_t \preceq 5\mathbf{F}_t$ . Note that  $\mathbf{F}_t$  does not depend on both  $a_t$  and  $\tilde{a}_t$ . Thus the matrix,

$$\mathbf{G}_t^{-1/2}(A_t + I_d)\mathbf{G}_t^{-1/2} \succeq \frac{1}{5} \sum_{s=1}^t \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2}$$

Let  $\mathcal{F}_t$  denote the sigma-algebra generated by  $\{(X_{k,s}, a_s, \tilde{a}_s) : k \in [K], s \in [t]\}$  and  $\{(a_s, \tilde{a}_s) : s \in [t]\} \cup \{a_{t+1}\}$ . Because the distribution of the hypothetical action  $\tilde{a}_t$  is  $\tilde{\pi}_t$  given  $a_t$  and  $\mathbf{F}_t$  depends only on the original contexts  $\{X_{k,s} : k \in [K], s \in [t]\}$ ,

$$\begin{aligned} & \mathbb{E} \left[ \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2} \middle| \mathcal{F}_{t-1} \right] \\ &= \mathbf{F}_t^{-1/2} \mathbb{E} \left[ \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \middle| \mathcal{F}_{t-1} \right] Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2} \\ &= \sum_{i=1}^{N_s} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2}, \end{aligned}$$

where the first equality holds because hypothetical contexts  $\{Z_{i,s} : i \in [N_s]\}$  is deterministic given the original contexts  $\{X_{k,t} : k \in [K]\}$  and the action  $a_t$ . Observe that for each  $s \in [t]$ ,

$$\begin{aligned} & \lambda_{\max} \left( \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2} \right) \\ & \leq \frac{d}{1-\gamma} \max_{i \in [N_s]} \lambda_{\max} (\mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2}) \\ & = \frac{d}{1-\gamma} \max_{i \in [N_s]} \|Z_{i,s}\|_{\mathbf{F}_t^{-1}}^2. \end{aligned}$$

By definition of the hypothetical contexts (13), for any  $s \in [t]$  and  $i \in [N_s]$ ,

$$\begin{aligned} \mathbf{F}_t &\succeq 2\bar{x}_{\max}^2 h_t I_d \\ &\succeq Z_{i,s} Z_{i,s}^\top + \bar{x}_{\max}^2 h_t I_d, \end{aligned}$$

where the last inequality uses  $\|Z_{i,s}\|_2 \leq \max\{x_{\max}, 1\}$ . Using Sherman-Morrison formula, for any  $s \in [t]$  and  $i \in [N_s]$ ,

$$\begin{aligned} \|Z_{i,s}\|_{\mathbf{F}_t^{-1}}^2 &\leq Z_{i,s}^\top (Z_{i,s} Z_{i,s}^\top + \bar{x}_{\max}^2 h_t I_d)^{-1} Z_{i,s} \\ &\leq \frac{\|Z_{i,s}\|_2^2}{\bar{x}_{\max}^2 h_t + \|Z_{i,s}\|_2^2} \\ &\leq \frac{1}{h_t + 1}, \end{aligned}$$

where the last inequality uses  $\|Z_{i,s}\|_2 \leq \bar{x}_{\max}$ . By definition of  $h_t$ , we obtain,

$$\lambda_{\max} \left( \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2} \right) \leq \frac{d}{(1-\gamma)h_t} \leq \frac{1/2 - e^{-1}}{2} \left( \log \frac{dt^2}{\delta} \right)^{-1}$$

Then by Lemma 14, with probability at least  $1 - \delta/t^2$

$$\begin{aligned} & 2(1/2 - e^{-1})^{-1} \log \frac{dt^2}{\delta} \sum_{s=1}^t \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2} \\ & \succeq 2(1/2 - e^{-1})^{-1} \log \frac{dt^2}{\delta} (1 - e^{-1}) \mathbf{F}_t^{-1} \mathbf{G}_t \mathbf{F}_t^{-1} - \log \frac{dt^2}{\delta} I_d \\ & \succeq (1/2 - e^{-1})^{-1} \log \frac{dt^2}{\delta} (1 - e^{-1}) I_d - \log \frac{dt^2}{\delta} I_d, \end{aligned}$$

where the last inequality uses the fact that  $\mathbf{G}_t \succeq \mathbf{F}_t/2$  (Lemma 4). Thus,

$$\sum_{s=1}^t \sum_{i=1}^{N_s} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} \mathbf{F}_t^{-1/2} Z_{i,s} Z_{i,s}^\top \mathbf{F}_t^{-1/2} \succeq \frac{1 - e^{-1}}{2} - \frac{1/2 - e^{-1}}{2} = \frac{1}{4},$$

which implies,

$$\mathbf{G}_t^{-1} (A_t + I_d) \mathbf{G}_t^{-1} \succeq \frac{1}{20} I_d.$$

Thus,

$$\|\mathbf{G}_t (A_t + I_d) \mathbf{G}_t\|_2 \leq 20,$$

which implies

$$\|P_t\|_2 \leq 20 - 1 + \frac{20}{\lambda_{\min}(\mathbf{G}_t)}.$$

Since  $\min(\mathbf{G}_t) \geq h_t \max\{x_{\max}^2, 1\} \geq 20$ , we have

$$\|P_t\|_2 \leq 20$$

■

**Lemma 12** For any  $\delta \in (0, 1)$ ,

$$\mathbb{P} \left( \bigcap_{t: |\mathcal{A}_t| \geq h_t} \left[ \left\{ \|S_t\|_{\mathbf{G}_t} > \frac{\sigma}{\gamma} \sqrt{d \log \frac{1+t}{\delta}} \right\} \cap \mathcal{S}_t \right] \right) \leq \delta.$$

**Proof** Under the event  $\mathcal{S}_t$ , we have:

$$\|S_t\|_{\mathbf{G}_t^{-1}} := \left\| \sum_{s=1}^t \sum_{i=1}^{r_s+1} \frac{\mathbb{I}(\tilde{a}_s = i)}{\phi_{i,s}} (W_{i,s} - Z_{i,s}^\top \theta_\star) Z_{i,s} \right\|_{\mathbf{G}_t^{-1}}.$$

Simplifying this, we get:

$$\|S_t\|_{\mathbf{G}_t^{-1}} = \frac{1}{\gamma} \left\| \sum_{s=1}^t \eta_{a_s, s} X_{a_s, s} \right\|_{\mathbf{G}_t^{-1}}.$$

Next, we observe that the Gram matrix  $\mathbf{G}_t$  satisfies:

$$\mathbf{G}_t \succeq \sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + \sum_{s \in \mathcal{A}_t} \sum_{i=1}^d Z_{i,s} Z_{i,s}^\top.$$

By the definition of the new contexts, we have:

$$\sum_{s \in \mathcal{A}_t} \sum_{i=1}^d Z_{i,s} Z_{i,s}^\top \succeq h_t \max\{x_{\max}^2, 1\} I_d.$$

Thus,

$$\mathbf{G}_t \succeq \sum_{s=1}^t X_{a_s,s} X_{a_s,s}^\top + \max\{x_{\max}^2, 1\} I_d.$$

Using Lemma 9 and Lemma 10 from Abbasi-Yadkori et al. (2011), with probability at least  $1 - \delta$ , we get:

$$\frac{1}{\gamma} \left\| \sum_{s=1}^t \eta_{a_s,s} X_{a_s,s} \right\|_{\mathbf{G}_t} \leq \frac{\sigma}{\gamma} \sqrt{d \log \left( \frac{1+t}{\delta} \right)}.$$

This completes the proof. ■

#### A.4 Proof of Lemma 9

**Proof** For  $k \in [K]$  and  $t \in [T]$ , let  $\tilde{Y}_{k,t} := X_{k,t}^\top \tilde{\theta}_{k,t}$  denote the estimated reward for arm  $k$ . Define the maximizer  $\tilde{M}_t := \arg \max_{k \in [K]} \tilde{Y}_{k,t}$ . Since HCSA+TS selects the arm that maximizes  $\tilde{Y}_{k,t}$ , the distribution of  $a_t$  matches that of  $\tilde{M}_t$ , i.e.,

$$\mathbb{P}(a_t \in \mathcal{P}_t | \mathcal{H}_t) = \mathbb{P}\left(\tilde{M}_t \in \mathcal{P}_t \middle| \mathcal{H}_t\right).$$

Suppose the estimated reward for the optimal arm,  $\tilde{Y}_{a_t^*,t}$ , exceeds  $\tilde{Y}_{j,t}$  for all  $j \in [K] \setminus \mathcal{P}_t$ . Since the optimal arm  $a_t^*$  is always in  $\mathcal{P}_t$  by definition, it follows that  $a_t^* \in \mathcal{P}_t$ . Thus:

$$\mathbb{P}\left(\tilde{M}_t \in \mathcal{P}_t \middle| \mathcal{H}_t\right) \geq \mathbb{P}\left(\bigcap_{j \in [K] \setminus \mathcal{P}_t} \{\tilde{Y}_{j,t} < \tilde{Y}_{a_t^*,t}\} \middle| \mathcal{H}_t\right).$$

Let  $Z_{j,t} := \tilde{Y}_{a_t^*,t} - \tilde{Y}_{j,t} - (X_{a_t^*,t} - X_{j,t})^\top \hat{\theta}_{t-1}$ . Then:

$$\mathbb{P}\left(\tilde{M}_t \in \mathcal{P}_t \middle| \mathcal{H}_t\right) \geq \mathbb{P}\left(\bigcap_{j \in [K] \setminus \mathcal{P}_t} \{Z_{j,t} > (X_{j,t} - X_{a_t^*,t})^\top \hat{\theta}_{t-1}\} \middle| \mathcal{H}_t\right).$$

Given  $\mathcal{H}_t$ ,  $\{Z_{j,t} : j \in [K] \setminus \mathcal{P}_t\}$  are Gaussian random variables with mean 0 and variance  $\mathbf{G}_t(\|X_{a_t^*,t}\|_{V_{t-1}}^2 + \|X_{j,t}\|_{V_{t-1}}^2)$ . For each  $j \notin \mathcal{P}_t$ , we have:

$$(X_{j,t} - X_{a_t^*,t})^\top \hat{\theta}_{t-1} \leq 2x_t \left\| \hat{\theta}_{t-1} - \theta_\star \right\|_{V_{t-1}} - \Delta_{j,t} \leq -\sqrt{\|X_{a_t^*,t}\|_{V_{t-1}}^2 + \|X_{j,t}\|_{V_{t-1}}^2}.$$

Thus:

$$\mathbb{P}\left(\tilde{M}_t \in \mathcal{P}_t \mid \mathcal{H}_t\right) \geq \mathbb{P}\left(\frac{Z_{j,t}}{\mathbf{G}_t \sqrt{\|X_{a_t^*, t}\|_{V_{t-1}^{-1}}^2 + \|X_{j,t}\|_{V_{t-1}^{-1}}^2}} > -\frac{1}{\mathbf{G}_t}, \forall j \notin \mathcal{P}_t \mid \mathcal{H}_t\right).$$

Since  $Z_{j,t}$  is Gaussian with variance normalized to 1,

$$\begin{aligned} \mathbb{P}\left(\tilde{M}_t \in \mathcal{P}_t \mid \mathcal{H}_t\right) &\geq \mathbb{P}\left(\frac{Z_{j,t}}{\mathbf{G}_t \sqrt{\|X_{a_t^*, t}\|_{V_{t-1}^{-1}}^2 + \|X_{j,t}\|_{V_{t-1}^{-1}}^2}} > -\frac{1}{\mathbf{G}_t}, \forall j \notin \mathcal{P}_t \mid \mathcal{H}_t\right) \\ &:= \mathbb{P}(Y_j > -\mathbf{G}_t, \forall j \notin \mathcal{P}_t \mid \mathcal{H}_t), \end{aligned}$$

where

$$Y_j := \frac{Z_{j,t}}{\mathbf{G}_t \sqrt{\|X_{a_t^*, t}\|_{V_{t-1}^{-1}}^2 + \|X_{j,t}\|_{V_{t-1}^{-1}}^2}}$$

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

$$\mathbb{P}(Y_j \leq -\mathbf{G}_t \mid \mathcal{H}_t) \leq \exp\left(-\frac{1}{2\mathbf{G}_t}\right).$$

Now, setting  $\mathbf{G}_t = \left(2 \log \frac{K(t+1)^2}{\delta}\right)^{-1/2}$ , we get:

$$\mathbb{P}(Y_j \leq -\mathbf{G}_t \mid \mathcal{H}_t) \leq \exp\left(-\log \frac{(t+1)^2}{\delta}\right) = \frac{\delta}{K(t+1)^2}.$$

Thus, we obtain:

$$\begin{aligned} \mathbb{P}\left(\tilde{M}_t \in \mathcal{P}_t \mid \mathcal{H}_t\right) &\geq 1 - \mathbb{P}(Y_j \leq -\mathbf{G}_t, \exists j \notin \mathcal{P}_t \mid \mathcal{H}_t) \\ &\geq 1 - \sum_{j \notin \mathcal{P}_t} \mathbb{P}(Y_j \leq -\mathbf{G}_t \mid \mathcal{H}_t) \\ &\geq 1 - \frac{\delta}{(t+1)^2}. \end{aligned}$$

This completes the proof. ■

## A.5 Proof of Lemma 10

**Proof** By definition of  $\mathcal{A}_t$ , for  $t$  such that  $t \leq h_t$ , we have  $t \in \mathcal{A}_t$  and  $1 \in \mathcal{A}_t$  for all  $t \in [T]$ . By definition of the new contexts (13), for any  $t \in [T]$ ,

$$\begin{aligned} \mathbf{G}_t &= \sum_{s=1}^t \sum_{i=1}^{N_s} Z_{i,t} Z_{i,t}^\top \succeq \sum_{s \in [t] \setminus \mathcal{A}_t} \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top + \sum_{i=1}^d u_{i,1} u_{i,1}^\top \\ &\succeq \sum_{s \in [t] \setminus \mathcal{A}_t} \sum_{i=1}^{N_s} Z_{i,s} Z_{i,s}^\top + \bar{x}_{\max}^2 \mathbf{I}_d. \end{aligned}$$

For  $t \in [T] \setminus \mathcal{A}_T$ , define

$$\nu_t := \operatorname{argmax}_{a \in [K]} \|X_{a,t}\|_{(\sum_{s \in [t-1] \setminus \mathcal{A}_t} X_{\nu_s,s} X_{\nu_s,s}^\top + \bar{x}_{\max}^2 \mathbf{I}_d)^{-1}}$$

By Lemma ??,

$$\mathbf{G}_t \succeq \frac{1}{2} \underbrace{\sum_{s \in [t] \setminus \mathcal{A}_t} X_{\nu_s,s} X_{\nu_s,s}^\top}_{W_t} + \bar{x}_{\max}^2 \mathbf{I}_d.$$

and  $x_t^2 \leq 2\|X_{\nu_t,t}\|_{W_{t-1}^{-1}}^2$ . Applying Lemma 11 from Abbasi-Yadkori et al. (2011):

$$\sum_{t \in [T] \setminus \mathcal{A}_T} x_t^2 \leq 2 \sum_{t \in [T] \setminus \mathcal{A}_T} \|X_{\nu_t,t}\|_{W_{t-1}^{-1}}^2 \leq 4 \log \frac{\det(W_T)}{\det(\max\{x_{\max}^2, 1\} I_d)}.$$

By AM-GM inequality,  $\det(W_T) \leq \left(\frac{\operatorname{Tr}(W_T)}{d}\right)^d \leq \left(\frac{T \max\{x_{\max}^2, 1\}}{d}\right)^d$  holds. Thus,

$$\sum_{t \in [T] \setminus \mathcal{A}_T} x_t^2 \leq 4d \log \frac{\operatorname{Tr}(W_T)}{\max\{x_{\max}^2, 1\} d} \leq 4d \log \frac{T \max\{x_{\max}^2, 1\}}{\max\{x_{\max}^2, 1\} d} \leq 4d \log \frac{T}{d},$$

which completes the proof. ■

## Appendix B. Technical Results

### B.1 Threshold for logarithmic inequality.

**Proposition 13 (Threshold for logarithmic inequality Lemma C.6 in Kim et al. (2025))**  
*For  $a > 1/2$  and  $b > e^2$ ,  $t \geq 4a \left(1 + \log \frac{2a\sqrt{b}}{e}\right)$  implies  $t \geq a \log bt^2$ .*

### B.2 A Matrix Concentration Inequality

**Lemma 14 (Matrix concentration inequality)** *Let  $\{M_s : s \in [t]\}$  be a  $\mathbb{R}^{d \times d}$ -valued stochastic process adapted to the filtration  $\{\mathcal{F}_s : s \in [t]\}$ . Suppose  $M_s$  are nonnegative definite symmetric matrices such that  $\lambda_{\max}(M_s) \leq 1$ . Then, with probability at least  $1 - \delta$ ,*

$$\sum_{s=1}^t M_s \succeq (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] - \log \frac{d}{\delta} I_d.$$

In addition, with probability at least  $1 - \delta$ ,

$$\sum_{s=1}^t M_s \preceq (e - 1) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] + \log \frac{d}{\delta} I_d.$$

**Proof** This proof is an adapted version of the argument from Tropp (2012).

For the lower bound, it is sufficient to prove that

$$\lambda_{\max} \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \leq \log \frac{d}{\delta},$$

with probability at least  $1 - \delta$ . By the spectral mapping theorem,

$$\begin{aligned} & \exp \left( \lambda_{\max} \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \\ & \leq \lambda_{\max} \left( \exp \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \\ & \leq \text{Tr} \left( \exp \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right). \end{aligned}$$

Taking the expectation of both sides gives:

$$\begin{aligned} & \mathbb{E} \exp \left( \lambda_{\max} \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \\ & \leq \mathbb{E} \text{Tr} \left( \exp \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \\ & = \mathbb{E} \text{Tr} \left( \mathbb{E} \left[ \exp \left( - \sum_{s=1}^{t-1} M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] + \log \exp(-M_t) \right) \middle| \mathcal{F}_{t-1} \right] \right) \\ & \leq \mathbb{E} \text{Tr} \left( \exp \left( - \sum_{s=1}^{t-1} M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] + \log \mathbb{E}[\exp(-M_t) | \mathcal{F}_{t-1}] \right) \right). \end{aligned}$$

The last inequality follows from Lieb's theorem (Tropp, 2015). Define a function  $f_\lambda : [0, 1] \rightarrow \mathbb{R}$  as  $f_\lambda(x) = e^{\lambda x} - x(e^\lambda - 1) - 1$ . Then  $f_\lambda(x)$  is convex with  $f_\lambda(0) = f_\lambda(1) = 0$  for all  $\lambda \in \mathbb{R}$ . Thus,  $e^{-x} \leq 1 + x(e^{-1} - 1)$  for  $x \in [0, 1]$ . Because the eigenvalues of  $M_s$  lie in  $[0, 1]$ , by the spectral mapping theorem, we have

$$\mathbb{E}[\exp(-M_t) | \mathcal{F}_{t-1}] \preceq I + (e^{-1} - 1)\mathbb{E}[M_t | \mathcal{F}_{t-1}] \preceq \exp(-(1 - e^{-1})\mathbb{E}[M_t | \mathcal{F}_{t-1}]).$$

Thus, we obtain:

$$\begin{aligned}
& \mathbb{E} \exp \left( \lambda_{\max} \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \\
& \leq \mathbb{E} \text{Tr} \left( \exp \left( - \sum_{s=1}^{t-1} M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] + \log \exp(-M_t) \right) \right) \\
& = \mathbb{E} \text{Tr} \left( \exp \left( - \sum_{s=1}^{t-1} M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] + \log \exp(-(1 - e^{-1}) \mathbb{E}[M_t | \mathcal{F}_{s-1}]) \right) \right) \\
& = \mathbb{E} \text{Tr} \left( \exp \left( - \sum_{\tau=1}^{t-1} M_\tau + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \\
& \leq \vdots \\
& \leq \mathbb{E} \text{Tr}(\exp(O)) = d.
\end{aligned}$$

Now, by Markov's inequality:

$$\begin{aligned}
& \mathbb{P} \left( \lambda_{\max} \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) > \log \frac{d}{\delta} \right) \\
& \leq \mathbb{E} \exp \left( \lambda_{\max} \left( - \sum_{s=1}^t M_s + (1 - e^{-1}) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \right) \frac{\delta}{d} \\
& \leq \delta.
\end{aligned}$$

For the upper bound, we can prove:

$$\lambda_{\max} \left( \sum_{s=1}^t M_s - (e - 1) \sum_{s=1}^t \mathbb{E}[M_s | \mathcal{F}_{s-1}] \right) \leq \log \frac{d}{\delta},$$

in a similar manner, using the fact that  $e^x \leq 1 + (e - 1)x$  for  $x \in [0, 1]$ . ■

## References

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- 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. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002a.
- Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.
- Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.
- Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.
- Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- 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.
- Varsha Dani, Thomas Hayes, and Sham Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, 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.
- Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.
- Peter Guttorp and Georg Lindgren. Karl pearson and the scandinavian school of statistics. *International Statistical Review*, 77(1):64–71, 2009.
- Nima Hamidi and Mohsen Bayati. Worst-case analysis of Thompson sampling for linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Jyh-Yih Hsu, Wei-Kuo Tseng, Jia-You Hsieh, Chao-Jen Chang, and Huan Chen. The recommending agricultural product sales promotion mode in e-commerce using reinforcement learning with contextual multiarmed bandit algorithms. *Mathematical Problems in Engineering*, 2020(1):8836000, 2020.

- Tom Huix, Matthew Zhang, and Alain Durmus. Tight regret and complexity bounds for thompson sampling via langevin monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pages 8749–8770. PMLR, 2023.
- Gisoo Kim and Myunghee Cho Paik. Doubly-robust lasso bandit. In *Advances in Neural Information Processing Systems*, pages 5869–5879, 2019.
- Wonyoung Kim, Gi-Soo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling with linear payoffs. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.
- Wonyoung Kim, Kyungbok Lee, and Myunghee Cho Paik. Double doubly robust thompson sampling for generalized linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8300–8307, 2023a.
- Wonyoung Kim, Myunghee Cho Paik, and Min-Hwan Oh. Squeeze all: Novel estimator and self-normalized bound for linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3098–3124. PMLR, 2023b.
- Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. Learning the pareto front using bootstrapped observation samples. In *International Conference on Artificial Intelligence and Statistics*, pages 667–675. PMLR, 2025.
- Branislav Kveton, Csaba Szepesvári, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 530–540. PMLR, 22–25 Jul 2020.
- Branislav Kveton, Mikhail Konobeev, Manzil Zhao, Seohyun Hong, Chao-Kai Cheng, and Tanveer Ahmed. Meta-TS: Meta-learned Thompson sampling. In *International Conference on Machine Learning*, pages 5858–5867. PMLR, 2021.
- 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.
- Jonas Mockus. The bayesian approach to global optimization. In *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981*, pages 473–481. Springer, 2005.
- Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005.
- Molly Offer-Westort, Alexander Coppock, and Donald P Green. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4):826–844, 2021.

Charles Riou, Junya Honda, and Shinji Matsushima. Matching Lai and Robbins lower bound for Thompson sampling. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Suhrid Satyal, Ingo Weber, Hye-young Paik, Claudio Di Cicco, and Jan Mendling. Ab testing for process versions with contextual multi-armed bandit algorithms. In *Advanced Information Systems Engineering: 30th International Conference, CAiSE 2018, Tallinn, Estonia, June 11-15, 2018, Proceedings 30*, pages 19–34. Springer, 2018.

Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR, 2018.

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

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

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

Qiuyu Zhu and Vincent Tan. Thompson sampling algorithms for mean-variance bandits. In *International Conference on Machine Learning*, pages 11599–11608. PMLR, 2020.