
Infrequent Exploration in Linear Bandits

Harin Lee

University of Washington
Seattle, WA, USA
leeharin@cs.washington.edu

Min-hwan Oh

Seoul National University
Seoul, South Korea
minoh@snu.ac.kr

Abstract

We study the problem of infrequent exploration in linear bandits, addressing a significant yet overlooked gap between fully adaptive exploratory methods (e.g., UCB and Thompson Sampling), which explore potentially at every time step, and purely greedy approaches, which require stringent diversity assumptions to succeed. Continuous exploration can be impractical or unethical in safety-critical or costly domains, while purely greedy strategies typically fail without adequate contextual diversity. To bridge these extremes, we introduce a simple and practical framework, INFEX, explicitly designed for infrequent exploration. INFEX executes a base exploratory policy according to a given schedule while predominantly choosing greedy actions in between. Despite its simplicity, our theoretical analysis demonstrates that INFEX achieves instance-dependent regret matching standard provably efficient algorithms, provided the exploration frequency exceeds a logarithmic threshold. Additionally, INFEX is a general, modular framework that allows seamless integration of any fully adaptive exploration method, enabling wide applicability and ease of adoption. By restricting intensive exploratory computations to infrequent intervals, our approach can also enhance computational efficiency. Empirical evaluations confirm our theoretical findings, showing state-of-the-art regret performance and runtime improvements over existing methods.

1 Introduction

The multi-armed bandit (MAB) problem [Lattimore and Szepesvári, 2020] captures a fundamental dilemma in sequential decision-making under uncertainty: at each time step, an agent must select an action (or arm) and receives feedback only from the chosen action, without observing the outcomes of alternative choices. Linear bandits generalize this problem by assuming that rewards follow a linear structure with respect to known arm features [Abe and Long, 1999, Auer, 2002, Dani et al., 2008], modeling diverse real-world scenarios such as clinical trials, recommendation systems, and adaptive pricing, where simultaneous learning and optimization are critical.

A central challenge in bandit settings is balancing *exploration*—acquiring new information about uncertain arms—with *exploitation*—leveraging existing knowledge to maximize immediate rewards. Classical algorithms, including Upper Confidence Bound (UCB) [Auer et al., 2002, Abbasi-Yadkori et al., 2011] and Thompson Sampling (TS) [Thompson, 1933, Agrawal and Goyal, 2012], resolve this tension by exploring systematically at *every time step*. These methods provide robust theoretical guarantees and strong empirical performance, forming the backbone of the MAB literature.

However, persistent exploration can be costly, risky, or ethically problematic in certain domains. For example, in healthcare or safety-critical settings, consistently experimenting with potentially suboptimal actions might lead to adverse or unacceptable outcomes. Consequently, it is often desirable to minimize exploration, performing it only when absolutely necessary. A straightforward alternative is the purely *greedy policy*, which consistently selects the currently estimated optimal arm, offering simplicity and reduced risk by avoiding unnecessary experimentation.

Recent literature has studied conditions under which greedy algorithms achieve near-optimal performance in linear *contextual* bandits [Kannan et al., 2018, Sivakumar et al., 2020, Bastani et al., 2021, Raghavan et al., 2023, Kim and Oh, 2024]. Crucially, these favorable theoretical guarantees of the greedy policy rely on strong distributional assumptions, such as sufficient *diversity in observed contexts*, which naturally facilitates exploration. However, these guarantees fail to hold even in the standard linear bandit settings with fixed arm features, where the greedy approach typically incurs linear regret due to insufficient exploration and inadequate information acquisition (e.g., Example 1 in Jedor et al. [2021]).

Thus, we are left with two extremes: at one extreme, greedy policies can succeed under strong diversity conditions (and may otherwise fail); at the other extreme, exploratory methods such as UCB or TS explicitly balance exploration and exploitation at every time step. Surprisingly, there is a substantial gap between these extremes. Specifically, the literature lacks rigorous studies on the impact of infrequent exploration on regret in linear bandit problems.¹

This raises fundamental open questions:

1. In order to achieve near-optimal performance (e.g., logarithmic regret), are we forced to explore at every time step, or can infrequent exploration suffice?
2. Can we devise an analytical framework to rigorously analyze methods with infrequent exploration, given that existing techniques may not directly apply?
3. How does the frequency of exploration affect regret performance?
4. Can infrequent exploration methods also demonstrate practical advantages beyond theoretical considerations?

Answering these questions not only provides fundamental theoretical insights but also has significant practical implications, particularly in domains where frequent exploration carries substantial cost or risk. Moreover, even when exploration risks are low, thoroughly investigating these questions may still yield meaningful practical benefits, as exploration typically entails additional computational cost.

In this work, we rigorously address this critical gap by introducing a novel and practical framework, **INFEX** (Infrequent Exploration), designed explicitly for infrequent exploration in linear bandits. Given a base exploratory policy Alg, our algorithm executes Alg according to a given schedule while predominantly making greedy action selections between these scheduled explorations. This hybrid approach naturally interpolates between fully exploratory and purely greedy strategies, offering fine-grained control over the exploration-exploitation trade-off. Notably, our approach is computationally efficient, which is particularly valuable in large-scale or real-time applications.

Our main contributions are summarized as follows:

- Our proposed framework **INFEX** is general and easily adoptable. It can seamlessly incorporate any (fully adaptive) linear bandit algorithm as the base policy, enabling broad applicability and straightforward integration into existing bandit implementations.
- We analyze the regret of **INFEX** within the linear bandit framework. We show that despite interleaving greedy actions—which individually could incur linear regret in naïve analysis—our algorithm achieves an instance-dependent regret matching that of **LinUCB** (or **OFUL**) [Abbasi-Yadkori et al., 2011], provided the total number of exploratory time steps exceeds the order of $\log T$. This result demonstrates that the asymptotic regret behavior remains unaffected by the infrequency of exploration. Furthermore, we complement the result by showing that the $\log T$ threshold is necessary (see Theorem 3).
- We construct a new analytical framework for infrequent exploration that establishes regret bounds for **INFEX** with arbitrary exploration schedules. Using this framework, we propose multiple exemplary exploration schedules and their resulting regret bounds. The main distinction of our analysis comes from the observation that the estimation error of the optimal arm directly affects the regret, and we show that this error decreases as the number of optimal selections increases.

¹While approaches such as the classic ϵ -greedy method—which introduces occasional stochastic exploration—and Explore-Then-Commit (ETC) algorithms—which perform an initial exploration phase followed by pure exploitation—are known to achieve suboptimal regret rates, in this work, we study whether infrequent exploration can be near-optimal.

- Furthermore, we derive a new instance-dependent regret bound for LinTS [Agrawal and Goyal, 2013, Abeille and Lazaric, 2017]. This new theoretical insight may independently interest the broader bandit research community.
- By limiting computationally intensive exploratory updates (e.g., posterior sampling or confidence set computations) to infrequent intervals, our algorithm significantly reduces runtime complexity compared to traditional approaches.
- Empirical results, provided in Section 5, substantiate our theoretical findings by demonstrating that, for suitable exploration schedules, INFEX outperforms both purely greedy and fully exploratory baselines in cumulative regret and computational efficiency.

1.1 Related Work

Full adaptive exploratory policies. Classical bandit algorithms, such as Upper Confidence Bound (UCB)[Auer et al., 2002, Abbasi-Yadkori et al., 2011] and Thompson Sampling (TS)[Thompson, 1933, Agrawal and Goyal, 2012], systematically balance exploration and exploitation at every time step. These approaches provide robust theoretical guarantees, including optimal logarithmic or sublinear regret bounds, and have been widely studied due to their effectiveness and simplicity. However, it remains an open question whether continuous exploration at every step is necessary or if infrequent exploration could suffice without compromising performance.

Greedy policies. Recently, significant research has investigated conditions under which purely greedy algorithms achieve near-optimal performance, particularly within contextual bandit frameworks. Studies by Bastani et al. [2021], Kannan et al. [2018], Sivakumar et al. [2020], Oh et al. [2021], Raghavan et al. [2023], Kim and Oh [2024] have shown that greedy policies can implicitly benefit from exploration when strong distributional assumptions, such as sufficient contextual diversity, are satisfied. For instance, Kannan et al. [2018], Sivakumar et al. [2020], Raghavan et al. [2023] assume that the context vectors are perturbed by a multivariate Gaussian distribution at each time step, forcing the context distribution to be diverse. Kim and Oh [2024] study a more general class of distributions under which greedy policies achieve polylogarithmic regret. While these findings identify specific scenarios favoring greedy methods, they leave unresolved how one should approach less ideal settings—such as linear bandit problems with fixed arm features lacking contextual diversity or stochastic variation, precisely the scenario addressed in our paper. In such standard linear bandit settings, purely greedy policies typically incur linear regret due to insufficient information gathering [Jedor et al., 2021], highlighting the necessity of explicit exploration.

Randomized/scheduled forced exploration. To incorporate explicit exploration in a simple manner, ϵ -greedy algorithms randomly explore arms with a small probability at each step [Lattimore and Szepesvári, 2020, Tirinzoni et al., 2022]. While intuitive and computationally efficient, ϵ -greedy policies are theoretically known to incur suboptimal regret. Another approach, forced-sampling [Goldenshluger and Zeevi, 2013, Bastani and Bayati, 2020, Lee et al., 2025], involves exploration at predetermined intervals. For instance, Goldenshluger and Zeevi [2013] demonstrate that scheduled forced-sampling combined with greedy exploitation can achieve polylogarithmic regret under favorable context distributions. Explore-Then-Commit (ETC) methods represent another scheduled exploration approach [Langford and Zhang, 2007, Abbasi-Yadkori et al., 2009, Garivier et al., 2016, Perchet et al., 2016, Hao et al., 2020], separating exploration and exploitation into distinct phases. ETC algorithms initially perform extensive exploration to identify promising actions, after which they commit exclusively to exploiting the best-identified arm. Despite their simplicity and intuitive appeal, ETC methods typically result in suboptimal regret compared to fully adaptive exploration strategies such as UCB and TS.

Infrequent exploration. To the best of our knowledge, approaches combining greedy exploitation with infrequent exploration have received limited attention, particularly in linear bandit contexts. One related work by Jin et al. [2023] studies multi-armed bandits without features and proposes a hybrid method that randomly chooses between Thompson Sampling and greedy selections. Their results highlight the potential theoretical benefits of strategically interleaving exploration and exploitation. Nevertheless, extending this hybridization concept rigorously to linear bandits and establishing near-optimal regret guarantees remains an important open question.

Despite extensive research on adaptive exploration methods, greedy algorithms, and scheduled exploration, significant gaps remain in understanding how exploration frequency affects regret in linear bandits. Key questions include: Is continuous exploration necessary for near-optimal performance, and can infrequent exploration achieve similar guarantees? Current analytical frameworks primarily address frequent exploration, highlighting the need for rigorous approaches tailored specifically to infrequent exploration scenarios.

2 Problem Setting

We consider the stochastic linear bandit problem. The agent is presented with a finite arm set $\mathcal{X} \subset \mathbb{B}^d$ with $|\mathcal{X}| = K$, where $\mathbb{B}^d$ is the d -dimensional unit ball. At each time step $t = 1, 2, \dots$, the agent selects an arm $X_t \in \mathcal{X}$ and receives a real-valued reward $Y_t = X_t^\top \theta^* + \eta_t$, where $\theta^* \in \mathbb{R}^d$ is an unknown parameter vector and η_t is zero-mean σ -subGaussian noise.² We assume that $\|\theta^*\|_2 \leq S$, and that this bound is known to the agent, where $\|\cdot\|_2$ denotes the ℓ_2 norm. The *optimal arm* is the arm with the highest expected reward and is denoted by $x^* := \operatorname{argmax}_{x \in \mathcal{X}} x^\top \theta^*$. We assume that it is unique for simplicity.

A linear bandit algorithm Alg is one that (possibly randomly) selects X_t based on the history $X_1, Y_1, \dots, X_{t-1}, Y_{t-1}$. The cumulative regret $\mathcal{R}_{\text{Alg}}(T)$ of an algorithm Alg over T time steps is defined as follows:

$$\mathcal{R}_{\text{Alg}}(T) := \sum_{t=1}^T (x^{*\top} \theta^* - X_t^\top \theta^*).$$

The goal of the agent is to minimize the cumulative regret. We primarily focus on instance-dependent regret, meaning that we study the growth of $\mathcal{R}_{\text{Alg}}(T)$ for a fixed problem instance $(\mathcal{X}, \theta^*)$.

3 Algorithmic Framework: INFEX

INFEX is a versatile and broadly applicable algorithmic framework designed for linear bandits and explicitly controls the frequency of exploration. The framework takes as input a base exploratory algorithm Alg and a predetermined exploration schedule $\mathcal{T}_e$ (i.e., a set of time-step indices). At each time step in $\mathcal{T}_e$, INFEX executes the exploratory algorithm Alg, while at all other steps it acts greedily based on the ridge estimator. We denote the resulting hybrid algorithm as $\text{INFEX}(\text{Alg}, \mathcal{T}_e)$.

One notable advantage of INFEX is its generic design, enabling seamless integration of virtually any linear bandit algorithm as the exploratory component. This flexibility facilitates straightforward adaptation to various application domains and existing algorithmic frameworks. Furthermore, by clearly separating exploration and exploitation phases, INFEX achieves computational efficiency by limiting the frequency of computationally intensive exploratory procedures.

The pseudocode describing the procedure is provided in Algorithm 1.

Remark 1 (Substituting the ridge estimator.). The only properties of the ridge estimator used in our analysis are the boundedness of the online squared-loss regret, $\sum_{t=1}^T (X_t^\top \hat{\theta}_{t-1} - X_t^\top \theta^*)^2 = \mathcal{O}(d^2 \log^2 T)$, and the fact that the estimation error $|x^\top \hat{\theta}_t - x^\top \theta^*|$ decreases proportionally to $1/\sqrt{n}$ when there are n samples of x in the data. Therefore, any estimator that satisfies similar properties may be used in place of the ridge estimator.

4 Theoretical Analysis

4.1 Notations and Definitions

Define $\text{reg}_t := x^{*\top} \theta^* - X_t^\top \theta^*$ to be the instantaneous regret at time step t . The main quantity that measures an instance's difficulty is the *minimum gap*, defined as $\Delta := x^{*\top} \theta^* - \max_{x \in \mathcal{X} \setminus \{x^*\}} x^\top \theta^*$. It represents the smallest possible non-zero instantaneous regret.

² η_t satisfies $\mathbb{E}[\exp(s\eta_t) \mid X_1, Y_1, \dots, X_t] \leq \exp(s^2 \sigma^2 / 2)$ for all $s \in \mathbb{R}$.

Algorithm 1 INFEX(Alg, $\mathcal{T}_e$): INFrequent EXploration

```
1: Input : Base algorithm Alg, exploration schedule  $\mathcal{T}_e \subset \mathbb{N}$ 
2: Initialize  $V_0 = I_d$ 
3: for  $t = 1, 2, \dots$ , do
4:   if  $t \in \mathcal{T}_e$  then
5:     Choose  $X_t$  according to Alg and observe  $Y_t$ 
6:   else
7:     Compute ridge estimator  $\hat{\theta}_{t-1} = V_{t-1}^{-1} \sum_{i=1}^{t-1} X_i Y_i$ 
8:     Choose  $X_t = \operatorname{argmax}_{x \in \mathcal{X}} x^\top \hat{\theta}_{t-1}$  and observe  $Y_t$ 
9:   end if
10:  Update  $V_t = V_{t-1} + X_t X_t^\top$ 
11: end for
```

For two positive functions $f(x)$ and $g(x)$, we write $f(x) = \mathcal{O}(g(x))$ if there exists a constant $C > 0$ such that $f(x) \leq Cg(x) + C$ for all x . When x is a positive real number and $\lim_{x \rightarrow \infty} \frac{g(x)}{f(x)} = 0$, we write $f(x) = \omega(g(x))$. In our analysis, we treat d, T, K , and Δ as variables, and regard all other quantities such as σ and S as constants.

We say an algorithm Alg attains (high-probability instance-dependent) *polylogarithmic regret* if $\mathcal{R}_{\text{Alg}}(T) = \mathcal{O}\left(\frac{d^a}{\Delta^b} \log^c T\right)$ for some constants $a, b, c \geq 0$ with probability at least $1 - 1/T$. Note that our analysis holds for an arbitrary failure probability $\delta \in (0, 1]$. For simplicity, we will mainly focus on the common choice $\delta \approx 1/T$. Such high-probability bounds that hold with probability at least $1 - 1/T$ immediately imply comparable expected-regret bounds.

Let $f(t) := |\mathcal{T}_e \cap \{1, 2, \dots, t\}|$ be the number of time steps at which Alg is executed by INFEX(Alg, $\mathcal{T}_e$) up to time step t . Hence, $f(t)$ is the frequency of exploratory steps up to time t . Let $f^{-1}(n) := \min\{t \in \mathbb{N} : f(t) \geq n\}$ be the time step at which Alg is executed for the n -th time. One particular exploration schedule of interest is the periodic schedule that executes Alg at a fixed interval. For a positive integer m , let $m\mathbb{N} := \{m, 2m, 3m, \dots\}$ denote the set of positive multiples of m . Then, the exploration schedule that executes Alg every m time steps corresponds to $\mathcal{T}_e = m\mathbb{N}$, and the resulting algorithm is denoted by INFEX(Alg, $m\mathbb{N}$).

Let $N_{\text{opt}}(T) := \sum_{t=1}^T \mathbb{1}\{x^* = X_t\}$ denote the number of times the optimal arm is selected up to time step T . We define $\alpha_t := \log \frac{\det V_t}{\det V_0}$ and $\beta_t(\delta) := \sigma \sqrt{\alpha_t + 2 \log(1/\delta)} + S$, which are key quantities in the analysis of many linear bandit algorithms [Abbasi-Yadkori et al., 2011]. For simplicity, we let $\beta_t := \beta_t(1/T)$ for all t .

4.2 Main Results

In this section, we analyze the regret bound of INFEX(Alg, $\mathcal{T}_e$).

Theorem 1 (Regret of INFEX). *Let Alg be a linear bandit algorithm that attains polylogarithmic regret, specifically $\mathcal{R}_{\text{Alg}}(T) = \mathcal{O}\left(\frac{d^a}{\Delta^b} \log^c T\right)$ with probability at least $1 - 1/T$ for some constants $a, b, c \geq 0$. Let $\mathcal{T}_e \subset \mathbb{N}$ be the set of exploratory time steps and $f(t) := |\mathcal{T}_e \cap \{1, 2, \dots, t\}|$ be the number of exploratory time steps up to time step t . Assume that $f(t) = \omega(\log t)$ as $t \rightarrow \infty$. Then, with probability at least $1 - 2/T$, the regret of INFEX(Alg, $\mathcal{T}_e$) is bounded as*

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}(T) \leq \mathcal{R}_{\text{Alg}}(f(T)) + G_{\text{const}}(\tau_{\text{Alg}}, f) + G(T),$$

where $G_{\text{const}}(\tau_{\text{Alg}}, f)$ is independent of T , $\tau_{\text{Alg}} \in \mathbb{N}$ is a constant determined by Alg satisfying $\tau_{\text{Alg}} = \mathcal{O}\left(\frac{d^a}{\Delta^{b+1}} \log^c \frac{d}{\Delta}\right)$, and

$$G(T) = \mathcal{O}\left(\frac{(\log T + d \log \log T + d \log \frac{d}{\Delta})^2}{\Delta}\right).$$

Bounds on $G_{\text{const}}(\tau_{\text{Alg}}, f)$ for some functions f are provided in Table 1.

Table 1: Example bounds on $G_{\text{const}}(\tau, f)$ for various functions f . *Epoch length* refers to the length between two consecutive executions of the base algorithm.

Example of $f(t)$	Description	$G_{\text{const}}(\tau_{\text{Alg}}, f)$
t/m	Epoch length is constant m	$\mathcal{O}\left(m\tau_{\text{Alg}} + \frac{md}{\Delta} \log^2 \frac{md}{\Delta}\right)$
$t/(\log t)^r$	Epoch length increases by $(\log t)^r$	$\mathcal{O}\left(\tau_{\text{Alg}} \log^r \tau_{\text{Alg}} + \frac{d}{\Delta} \log^{2+r} \frac{d}{\Delta}\right)$
t^r ($r \in (0, 1]$)	Epoch length increases by t^{1-r}	$\mathcal{O}\left(\tau_{\text{Alg}}^{1/r} + \frac{d^{1/r}}{\Delta^{2/r-1}} \log^{2/r} \frac{d}{\Delta}\right)$
$(\log t)^r$ ($r > 1$)	Epoch length increases exponentially	$e^{\mathcal{O}(\tau_{\text{Alg}}^{1/r})} + \Delta e^{\mathcal{O}((d/\Delta^2)^{\frac{1}{r-1}})}$

Discussion of Theorem 1. In the regret bound of Theorem 1, only the terms $\mathcal{R}_{\text{Alg}}(f(T))$ and $\mathcal{O}\left(\frac{1}{\Delta}(\log T + d \log \log T)^2\right)$ depend on T . The first term corresponds to the regret of the base algorithm Alg. The second term bounds the additional regret incurred by the interleaved greedy selections, and it matches the instance-dependent bound of LinUCB [Abbasi-Yadkori et al., 2011]. We emphasize that these terms do not increase as the number of explorations decreases; in fact, the first term decreases. Therefore, choosing a sparse exploration schedule does not worsen the asymptotic regret of INFEX(Alg, $\mathcal{T}_e$), as long as it satisfies the condition $f(t) = \omega(\log t)$. The trade-off from reduced exploration only appears in the constant term. $G_{\text{const}}(\tau_{\text{Alg}}, f)$ is the cumulative regret incurred by the greedy selections for some initial time steps, where greedy selections do not have strong guarantees. As shown in Table 1, an excessively small number of explorations may result in exponential growth of the constant term with respect to d/Δ , which may significantly degrade the algorithm's finite-time performance. Meanwhile, exploration with constant periods or logarithmically growing epochs increases $G_{\text{const}}(\tau_{\text{Alg}}, f)$ only by a constant or a logarithmic factor. For finite T , the least amount of exploration required to ensure that $G_{\text{const}}(\tau_{\text{Alg}}, f)$ does not exceed the order of $G(T)$ is determined by the relative magnitudes of d , T , and Δ . While it may be possible to allocate a minimal amount of exploration if all of these quantities are known, Δ is typically unknown to the agent, making it challenging to determine the optimal schedule. In practice, we suggest that periodic or logarithmically growing epochs would be efficient. However, it is very important to note that, even without knowing these quantities, INFEX achieves the same order of the regret compared to the vanilla fully adaptive exploration methods. In Section 5, we demonstrate through numerical simulations that exploration with a fixed period of 5 to 100, so that 80% to 99% of the actions are greedy, yields favorable performance in terms of both regret and computational efficiency.

Obtaining minimax bound. We mainly focus on the instance-dependent bounds in this paper to show how the exploration schedule affects the regret for a fixed instance. Meanwhile, providing the worst-case minimax regret bounds for infrequent exploration would also be an interesting problem. While the asymptotic behavior of the instance-dependent bounds achieves the same order of polylogarithmic regret as long as the exploration number satisfies $\omega(\log t)$, we conjecture that this threshold would be too small to achieve the optimal $\mathcal{O}(\sqrt{T})$ minimax guarantees. Finding the optimal infrequent exploration strategy and trade-offs for the minimax regret bound would be an interesting open problem.

As an instantiation of INFEX, we can choose Alg = LinUCB [Abbasi-Yadkori et al., 2011] or Alg = LinTS [Abeille and Lazaric, 2017], which are representative linear bandit algorithms. To show that Theorem 1 applies to both algorithms, we present their instance-dependent polylogarithmic regret bounds. To the best of our knowledge, the instance-dependent bound for LinTS is explicitly shown for the first time. The proof of Theorem 2 is deferred to Appendix B.

Theorem 2. *LinTS [Abeille and Lazaric, 2017] achieves the following instance-dependent bound with probability at least $1 - \delta$:*

$$\mathcal{R}_{\text{LinTS}}(T) = \mathcal{O}\left(\frac{\min\{d \log \frac{dT}{\delta}, \log \frac{KT}{\delta}\} (\alpha_T + \log \frac{1}{\delta})^2}{\Delta}\right),$$

$$\text{where } \alpha_T = \mathcal{O}\left(\min\left\{d \log T, \log T + d \log \log T + d \log \frac{d}{\Delta}\right\}\right).$$

Furthermore, Theorem 5 in Abbasi-Yadkori et al. [2011] states that the regret of LinUCB is $\mathcal{R}_{\text{LinUCB}}(T) = \mathcal{O}(\alpha_T^2/\Delta)$ with the same bound on α_T as in Theorem 2. Then, combined with the result of Theorem 1, we obtain the regret bounds for specific base algorithms. We show some example regret bounds for INFEX when Alg is LinUCB or LinTS with varying exploration schedule in Table 2. It demonstrates that the regret of INFEX, instantiated with LinUCB or LinTS, matches the regret bounds of the corresponding algorithms without infrequent exploration, up to factors independent of T .

Table 2: Example regret bounds of $\text{INFEX}(\text{Alg}, \mathcal{T}_e)$ with specific instantiations of Alg and $f(t)$. Each column shows the regret corresponding to each base algorithm. The final regret bound is the sum of the regret shown in the base regret row and the constant regret shown in the row with the corresponding exploration schedule.

Frequency of exploration	Regret bound of $\text{INFEX}(\text{Alg}, \mathcal{T}_e)$	
	Alg = LinUCB	Alg = LinTS
$f(t) = t$ (base)	$\mathcal{O}\left(\frac{1}{\Delta} (\log T + d \log \log T)^2\right)$	$\mathcal{O}\left(\frac{1}{\Delta} (d \log T) (\log T + d \log \log T)^2\right)$
$f(t) = t/m$	$\mathcal{O}\left((m + \frac{d}{\Delta}) \frac{d}{\Delta} \log^2 \frac{d}{\Delta}\right)$	$\mathcal{O}\left(\frac{d^3}{\Delta^2} \log^3 \frac{d}{\Delta} + \frac{md}{\Delta} \log^2 \frac{d}{\Delta}\right)$
$f(t) = t/(\log t)^r$	$\mathcal{O}\left(\frac{d^2}{\Delta^2} \log^{2+r} \frac{d}{\Delta}\right)$	$\mathcal{O}\left(\frac{d^3}{\Delta^2} \log^{3+r} \frac{d}{\Delta}\right)$
$f(t) = t^r$	$\mathcal{O}\left(\left(\frac{d}{\Delta} \log \frac{d}{\Delta}\right)^{\frac{2}{r}}\right)$	$\mathcal{O}\left(\left(\frac{d^3}{\Delta^2} \log^3 \frac{d}{\Delta}\right)^{\frac{1}{r}}\right)$
$f(t) = (\log t)^r$	$e^{\mathcal{O}\left(\left(\frac{d}{\Delta} \log \frac{d}{\Delta}\right)^{\frac{2}{r}} + \left(\frac{d}{\Delta^2}\right)^{\frac{1}{r-1}}\right)}$	$e^{\mathcal{O}\left(\left(\frac{d^3}{\Delta^2} \log^3 \frac{d}{\Delta}\right)^{\frac{1}{r}} + \left(\frac{d}{\Delta^2}\right)^{\frac{1}{r-1}}\right)}$

Computational complexity. The computational time complexity of a single greedy selection is $\mathcal{O}(d^2 + dK)$: using the Sherman-Morrison formula [Sherman and Morrison, 1950], one can maintain V_t^{-1} in $\mathcal{O}(d^2)$ time per step, so updating $\hat{\theta}_t$ also takes $\mathcal{O}(d^2)$ time, and the remaining $\mathcal{O}(dK)$ is required to find the arm with the highest estimated reward. The computational complexity of LinUCB is $\mathcal{O}(d^2 + d^2K)$ per time step, where the additional $\mathcal{O}(d^2K)$ term is required to compute the upper confidence bound of rewards $x^\top \hat{\theta}_t + \beta_t \|x\|_{V_t^{-1}}$ for all $x \in \mathcal{X}$. The computational complexity of LinTS is $\mathcal{O}(d^3 + dK)$, where the additional $\mathcal{O}(d^3)$ term corresponds to sampling parameter $\tilde{\theta}_t$ from a multivariate Gaussian distribution. Both algorithms have strictly greater computational complexity than performing a greedy selection, meaning that replacing them with greedy selections reduces the total computational cost.

4.3 Necessity of $\omega(\log t)$ Exploration.

We provide a lower-bound result that implies the condition $f(t) = \omega(\log t)$ is necessary to obtain a polylogarithmic regret bound that holds for any T . Specifically, we show that if $f(t) = \omega(\log t)$ does not hold, that is, either the limit $\lim_{t \rightarrow \infty} \frac{\log t}{f(t)}$ does not exist or is above zero, then there exists a problem instance such that the regret of INFEX scales almost linearly in T using the standard information-theoretical method.

Theorem 3. *Let Alg be an arbitrary policy and $\mathcal{T}_e \subset \mathbb{N}$ be a set of natural numbers. If $f(t) \neq \omega(\log t)$, then for an arbitrary constant $\varepsilon \in (0, 1)$, there exists a problem instance $(\mathcal{X}, \theta^*)$ and a constant $c(f, \varepsilon) > 0$ that depends on f and ε such that*

$$\mathbb{E} [\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}(T)] \geq c(f, \varepsilon) T^{1-\varepsilon}$$

for infinitely many $T \in \mathbb{N}$.

We note that this result applies to predetermined exploration schedules, and the $\omega(\log t)$ threshold might not be necessary when the exploration schedule is adaptive to the observations.

The proof of Theorem 3 is presented in Appendix A.4.

4.4 Sketch of Proof

In this subsection, we provide a sketch of the proof of Theorem 1. Throughout this subsection, we work under the high-probability event that $\mathcal{R}_{\text{Alg}}(T)$ is polylogarithmic in T and the event of Lemma 9 that ensures the concentration of $\hat{\theta}_t$ toward θ^* .

We first explain how τ_{Alg} is chosen. Assuming that Alg is independently run, τ_{Alg} is defined as the time step such that for all $T \geq \tau_{\text{Alg}}$, at least a quarter of the selections made by Alg are optimal, that is, the optimal arm is chosen in at least $T/4$ of the T time steps. The existence and order of τ_{Alg} are guaranteed by the following lemma:

Lemma 1. *Suppose a linear bandit algorithm Alg' attains a polylogarithmic regret bound of $\mathcal{R}_{\text{Alg}'}(T) = \mathcal{O}\left(\frac{d^a}{\Delta^b} \log^c T\right)$ for some constants $a, b, c \geq 0$. Then, there exists $\tau_{\text{Alg}'} \in \mathbb{N}$ such that for all $T \geq \tau_{\text{Alg}'}$, at least a quarter of the T selections made by Alg' are optimal. Furthermore, $\tau_{\text{Alg}'} = \mathcal{O}\left(\frac{d^a}{\Delta^{b+1}} \log^c \frac{d}{\Delta}\right)$.*

We mainly focus on the sum of regret incurred after the time step $f^{-1}(\tau_{\text{Alg}})$, that is, after Alg is executed for τ_{Alg} times. For $\tau, T \in \mathbb{N}$, let $\mathcal{G}(\tau, T) := \{t \in \mathbb{N} : \tau + 1 \leq t \leq T, t \notin \mathcal{T}_e\}$, which denotes the set of time steps with greedy selections between $\tau + 1$ and T , inclusively. Let $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau, T) := \sum_{t \in \mathcal{G}(\tau, T)} \text{reg}_t$ be the cumulative regret incurred at the time steps in $\mathcal{G}(\tau, T)$. In the remainder of this section, we show that $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(f^{-1}(\tau_{\text{Alg}}) + \tau_1, T)$ has the polylogarithmic bound stated in Theorem 1 for some constant τ_1 .

The following lemma shows that the regret of greedy selections is related to the number of optimal selections.

Lemma 2. *For any $\tau, T \in \mathbb{N}$ with $\tau < T$, it holds that*

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau, T) \leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau, T)} \frac{\beta_{t-1}^2}{1 + N_{\text{opt}}(t-1)}.$$

The intuition behind this lemma is that the estimator $\hat{\theta}_t$ becomes more accurate in estimating $x^{*\top} \theta^*$ as the optimal arm x^* is selected more often. The conclusion of the lemma implies that if $N_{\text{opt}}(t)$ increases linearly in t , then the additional regret caused by the greedy selections remains polylogarithmic in T . By the choice of τ_{Alg} , at least a quarter of the selections made by Alg are optimal for all $t \geq f^{-1}(\tau_{\text{Alg}})$, implying that $N_{\text{opt}}(t) \geq \frac{1}{4}f(t)$. This fact leads to the following regret bound:

Lemma 3. *Let τ_{Alg} be defined as in Theorem 1. Then, for any $T > f^{-1}(\tau_{\text{Alg}})$, it holds that*

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(f^{-1}(\tau_{\text{Alg}}), T) \leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(f^{-1}(\tau_{\text{Alg}}), T)} \frac{\beta_t^2}{f(t)}.$$

Furthermore, this bound is sublinear in T when $f(t) = \omega(\log t)$.

We further improve this bound by observing that the quantity $N_{\text{opt}}(t)$ must grow linearly with t for sufficiently large t as we now have a sublinear bound on $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}$. Using this fact, we obtain the following stronger regret bound.

Proposition 1. *There exists a constant $\tau_1 \in \mathbb{N}$ that depends on $d, \Delta, \tau_{\text{Alg}}$, and the function f , is independent of T , and satisfies*

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(f^{-1}(\tau_{\text{Alg}}), f^{-1}(\tau_{\text{Alg}}) + \tau_1) \leq \frac{7}{16} \Delta \tau_1$$

and

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(f^{-1}(\tau_{\text{Alg}}) + \tau_1, T) \leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{16\beta_T^2 \log T}{\Delta}$$

for all $T > f^{-1}(\tau_{\text{Alg}}) + \tau_1$.

Note that $\beta_T^2 = \mathcal{O}(\alpha_T)$, so we have derived a bound of $\mathcal{O}(\alpha_T(\alpha_T + \log T)/\Delta)$ with some additional constant amount. The proof is completed by providing an appropriate bound on α_T . We apply the following lemma, which is derived from the proof of Theorem 5 in Abbasi-Yadkori et al. [2011].

Lemma 4. *If the data $X_1, X_2, \dots, X_T$ is collected through a linear bandit algorithm Alg' , then*

$$\alpha_T \leq \log(1+T) + (d-1) \log \left(1 + \frac{\mathcal{R}_{\text{Alg}'}(T)}{(d-1)\Delta} \right).$$

Consequently, if Alg' attains polylogarithmic regret, then

$$\alpha_T = \mathcal{O} \left(\log T + d \log \log T + d \log \frac{d}{\Delta} \right).$$

The detailed proof of Theorem 1 is presented in Appendix A.

Remark 2. The analysis of Theorem 1 requires positivity of the minimum gap Δ and a fixed optimal arm. Therefore, the analysis holds as long as the two conditions are satisfied, even for infinite and time-varying arm sets, although it does not fully generalize to the linear contextual bandit setting with arbitrary arm sets. For a detailed discussion on the possibility of extending the analysis to time-varying arm sets, refer to Appendix E.

5 Numerical Experiments

To complement our theoretical analysis, we conduct numerical simulations to empirically investigate the behavior and practical benefits of INFEX. Our main objectives are to (i) assess whether infrequent exploration strategies maintain strong regret performance compared to fully adaptive methods, (ii) evaluate computational efficiency improvements due to reduced exploration frequency, and (iii) demonstrate the general applicability and robustness of our proposed framework across different base exploratory algorithms and exploration schedules.

We select $\text{Alg} = \text{LinUCB}$ and $\text{Alg} = \text{LinTS}$ as the base algorithms for exploration and use an exploration schedule $\mathcal{T}_e = m\mathbb{N} := \{mn : n \in \mathbb{N}\}$, meaning Alg executes every m steps. Specifically, we examine three choices of m : $m = 5$, $m = 20$, and $m = 100$, corresponding to 80%, 95%, and 99% greedy selections, respectively. For benchmarking, we also compare our framework against other policies: the purely greedy policy, a single-parameter version of `OLSBandit` [Goldenshluger and Zeevi, 2013], and an ε -greedy approach with $\varepsilon_t = t^{-1/3}$.

We randomly generate problem instances for given d and K as follows. We construct the arm set $\mathcal{X}$ by sampling K arms i.i.d. from a multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}_d, \frac{1}{2d}I_d)$ and rescaling each vector to have a norm at most 1 when it exceeds 1. We sample θ^* uniformly from the unit sphere in $\mathbb{R}^d$. The random reward is given as either $+1$ or -1 , with its expectation being $X_t^\top \theta^*$. We repeat the process for 20 randomly generated instances and report the mean and standard deviation of the cumulative regret over $T = 10000$ time steps for each algorithm.

Figure 1 shows the total regret and computation time of each algorithm. Interestingly, we observe that certain exploration schedules *improve* the total regret. Especially for $\text{Alg} = \text{LinTS}$, all values of $m = 5, 20, 100$ reduce the regret significantly. The performance of $\text{Alg} = \text{LinUCB}$ is also improved when $m = 5$. These configurations outperform both the base algorithm and the purely greedy policy, exhibiting strong practicality. We also observe a reduction of computational time for any value of m .

`OLSBandit` is inefficient because it spends most of the time steps, specifically at least $\Omega(d^2 \log T)$ steps, on forced sampling. While ε -greedy appears to show decent performance, we note that the choice $\varepsilon_t = t^{-1/3}$ implies a regret lower bound of $\Omega(T^{2/3})$ and it is its best bound, precluding the possibility of achieving polylogarithmic regret.

Refer to Appendix F for additional experiments with different dimensions d and experiment details.

6 Conclusion

We propose INFEX, a simple yet practical framework that mainly performs greedy selections while exploring according to a given schedule. Our theoretical analysis reveals that INFEX attains a polylogarithmic regret bound, whose growth rate with respect to T remains independent of the exploration schedule, provided that the exploration frequency exceeds the order of $\log T$. Empirical results further illustrate the strengths of INFEX, showing that judiciously timed exploration not only maintains robust theoretical performance guarantees but also delivers practical improvements in

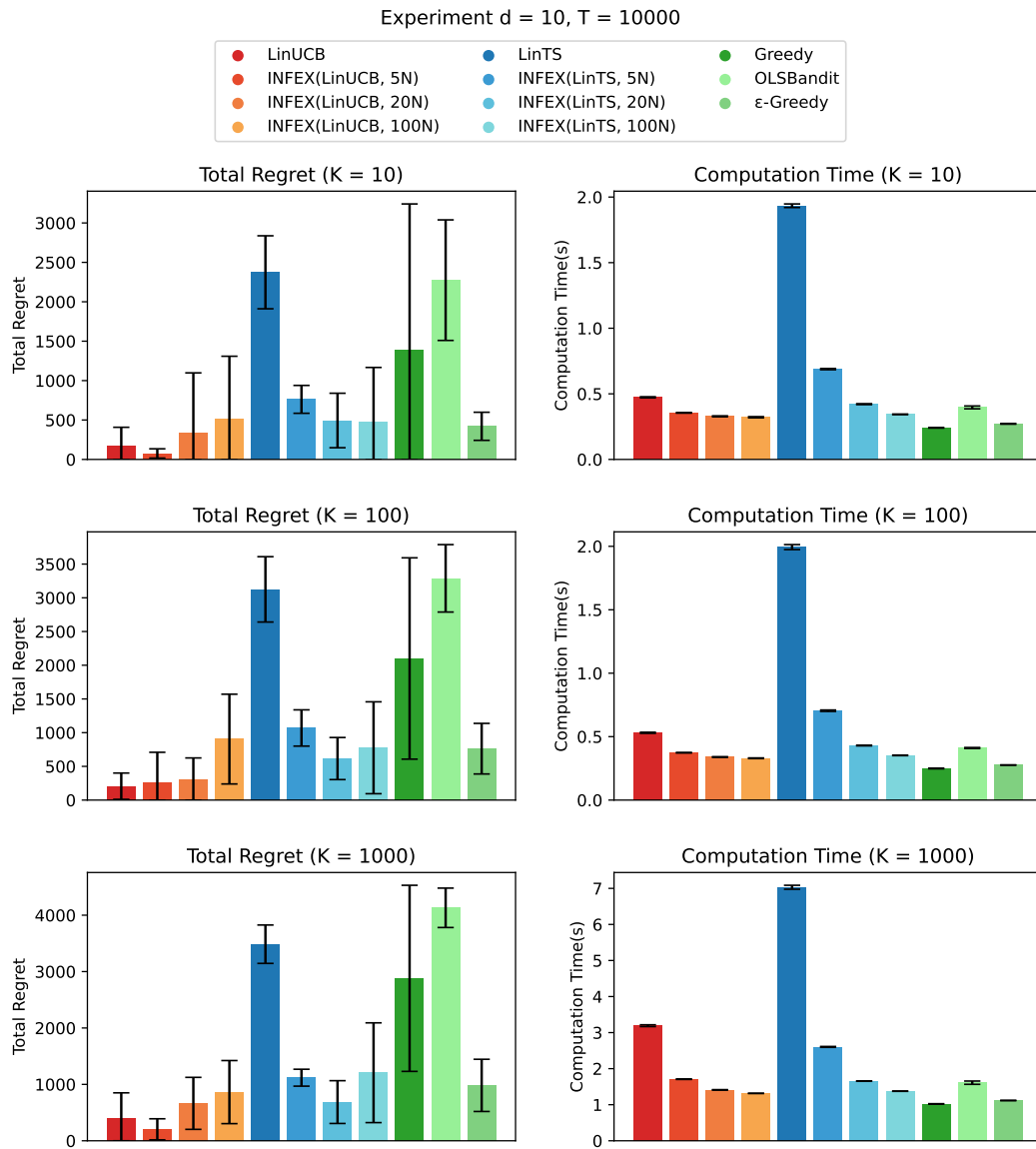


Figure 1: Comparison of total regret (left) and computation time (right) when $d = 10, T = 10000$, and $K = 10$ (top), $K = 100$ (middle), and $K = 1000$ (bottom).

terms of both regret and computational efficiency. While this work focuses specifically on linear bandit settings, we believe the framework and results serve as a foundation for broader exploration strategies, potentially enabling similar performance benefits in more complex and general function approximation scenarios. An exciting avenue for future research lies in extending our framework to accommodate these generalizations, further enhancing its applicability and impact.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-NR071853 and RS-2023-00222663), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263754), and by AI-Bio Research Grant through Seoul National University.

References

- Yasin Abbasi-Yadkori, András Antos, and Csaba Szepesvári. Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*, volume 92, page 236, 2009.
- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.
- Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning*, pages 3–11, 1999.
- Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 176–184. PMLR, PMLR, 2017.
- Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.
- Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.
- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.
- Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.
- Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.
- Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137, 1979.
- Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366, 2008.
- Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.
- Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.
- Osama A Hanna, Lin Yang, and Christina Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1791–1821. PMLR, 2023.
- Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.
- Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

- Matthieu Jedor, Jonathan Lou  dec, and Vianney Perchet. Be greedy in multi-armed bandits. *arXiv preprint arXiv:2101.01086*, 2021.
- Tianyuan Jin, Xianglin Yang, Xiaokui Xiao, and Pan Xu. Thompson sampling with less exploration is fast and optimal. In *International Conference on Machine Learning*, pages 15239–15261. PMLR, 2023.
- Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31, 2018.
- Seok-Jin Kim and Min-hwan Oh. Local anti-concentration class: Logarithmic regret for greedy linear contextual bandit. *Advances in Neural Information Processing Systems*, 37:77525–77592, 2024.
- John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.
- Tor Lattimore and Csaba Szepesv  ri. *Bandit algorithms*. Cambridge University Press, 2020.
- Harin Lee, Taehyun Hwang, and Min hwan Oh. Lasso bandit with compatibility condition on optimal arm. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-agnostic lasso bandit. In *International Conference on Machine Learning*, pages 8271–8280. PMLR, 2021.
- Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. 2016.
- Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. Greedy algorithm almost dominates in smoothed contextual bandits. *SIAM Journal on Computing*, 52(2): 487–524, 2023.
- Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.
- Vidyashankar Sivakumar, Steven Wu, and Arindam Banerjee. Structured linear contextual bandits: A sharp and geometric smoothed analysis. In *International Conference on Machine Learning*, pages 9026–9035. PMLR, 2020.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- Andrea Tirinzoni, Matteo Papini, Ahmed Touati, Alessandro Lazaric, and Matteo Pirodda. Scalable representation learning in linear contextual bandits with constant regret guarantees. *Advances in Neural Information Processing Systems*, 35:2307–2319, 2022.
- Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The abstract and introduction (Section 1) accurately reflect the paper's contribution and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitation of the work is discussed in Appendix E.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The full set of assumptions is presented in Section 2. The complete proofs of the theoretical results is provided in the Appendix A-D with a sketch of the proof provided in Section 4.4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide code for experiments with a random seed that produces the results in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit experiment code as supplemental material, which reproduces the experiment results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experiment setting and details are accurately and sufficiently described in Section 5 and Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The experiment results are presented with 1-standard deviation error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Information on the computer resources is provided in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The authors have reviewed the NeurIPS Code of Ethics and the research conducted in the paper conform, in every respect, with it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Proof of Theorem 1

In this section, we provide a detailed proof of Theorem 1. We supplement the proof by proving Proposition 1 in Appendix A.2 and verifying the bounds of $G_{\text{const}}(\tau_{\text{Alg}}, f)$ listed in Table 1 in Appendix A.3. In Appendix A.4, we prove Theorem 3. Proofs of technical lemmas are provided in Appendix C.

Throughout the proof, we denote $\tau_0 := f^{-1}(\tau_{\text{Alg}})$ for simplicity.

A.1 Proof of Theorem 1

Proof of Theorem 1. τ_{Alg} is set in the way described in Lemma 1 with $\text{Alg}' = \text{Alg}$, and the lemma guarantees that $\tau_{\text{Alg}} = \mathcal{O}(\frac{d^a}{\Delta^{b+1}} \log^c \frac{d}{\Delta})$. τ_1 is the constant defined in Proposition 1.

The total regret is decomposed into four parts, described in Eq. (1).

$$\begin{aligned} \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}(T) &\leq \mathcal{R}_{\text{Alg}}(f(T)) + 2S\tau_0 + \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0, \tau_0 + \tau_1) \\ &\quad + \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0 + \tau_1, T). \end{aligned} \quad (1)$$

The first term is the sum of the regret incurred by Alg. Since Alg is executed $f(T)$ times, this regret is bounded by $\mathcal{R}_{\text{Alg}}(f(T))$. The second part is the sum of the regret incurred by the greedy selections during the first τ_0 time steps. Since the maximum possible regret per time step is $2S$, we bound the sum by $2S\tau_0$. Note that this quantity is independent of T . Lastly, among the time steps that perform greedy selections, $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0, \tau_0 + \tau_1)$ is the sum of the regret incurred during the time steps between $\tau_0 + 1$ and $\tau_0 + \tau_1$, inclusively, and $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0 + \tau_1, T)$ is the sum of the regret incurred during the time steps between $\tau_0 + \tau_1 + 1$ and T , inclusively.

By Proposition 1, we have $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0, \tau_0 + \tau_1) \leq \frac{7}{16}\Delta\tau_1$ and $\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0 + \tau_1, T) = \mathcal{O}(\alpha_T(\alpha_T + \log T)/\Delta)$. Denoting $\tilde{G}_{\text{const}} := 2S\tau_0 + \frac{7}{16}\Delta\tau_1$, we obtain that

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}(T) \leq \mathcal{R}_{\text{Alg}}(f(T)) + \tilde{G}_{\text{const}} + \mathcal{O}\left(\frac{\alpha_T(\alpha_T + \log T)}{\Delta}\right) \quad (2)$$

$$= \mathcal{R}_{\text{Alg}}(f(T)) + \tilde{G}_{\text{const}} + \mathcal{O}\left(\frac{(d \log T)^2}{\Delta}\right), \quad (3)$$

where we use Lemma 10 for the last equality. Eq. (3) shows that $\text{INFEX}(\text{Alg}, \mathcal{T}_e)$ achieves a poly-logarithmic regret bound added by a T -independent constant. We improve the bound on α_T using Lemma 4 and the derived regret bound. The growth rate of the logarithm of the cumulative regret is $\log(1 + \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}(T)) = \mathcal{O}(\log(\frac{d}{\Delta} \log T) + \log \tilde{G}_{\text{const}})$. Applying this fact to Lemma 4, we obtain that

$$\alpha_T = \mathcal{O}\left(\log T + d \log \log T + d \log \frac{d}{\Delta} + d \log \tilde{G}_{\text{const}}\right).$$

Plugging this bound into Eq. (2), we obtain that

$$\begin{aligned} \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}(T) &\leq \mathcal{R}_{\text{Alg}}(f(T)) + \tilde{G}_{\text{const}} \\ &\quad + \mathcal{O}\left(\frac{1}{\Delta} \left(\log T + d \log \log T + d \log \frac{d}{\Delta} + d \log \tilde{G}_{\text{const}}\right)^2\right) \\ &= \mathcal{R}_{\text{Alg}}(f(T)) + \tilde{G}_{\text{const}} + \mathcal{O}\left(\frac{1}{\Delta} \left(d \log \tilde{G}_{\text{const}}\right)^2\right) \\ &\quad + \mathcal{O}\left(\frac{1}{\Delta} \left(\log T + d \log \log T + d \log \frac{d}{\Delta}\right)^2\right), \end{aligned} \quad (4)$$

where the last equality holds since $(a + b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$. Therefore, there exists a constant $G_{\text{const}}(\tau_{\text{Alg}}, f) = \tilde{G}_{\text{const}} + \mathcal{O}\left(\frac{1}{\Delta} \left(d \log \tilde{G}_{\text{const}}\right)^2\right)$ and a function $G(T)$ in

$\mathcal{O}\left(\frac{1}{\Delta} \left(\log T + d \log \log T + d \log \frac{d}{\Delta}\right)^2\right)$ such that

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \tau_e)}(T) \leq \mathcal{R}_{\text{Alg}}(f(T)) + G_{\text{const}}(\tau_{\text{Alg}}, f) + G(T).$$

In Appendix A.3, we summarize how $G_{\text{const}}(\tau_{\text{Alg}}, f)$ is determined and provide its example bounds listed in Table 1. $\square$

A.2 Proof of Proposition 1

Proof of Proposition 1. By the sublinearity stated in Lemma 3, there exists a constant τ_1 that depends on $d, \Delta, \tau_{\text{Alg}}$, and f such that for all $T \geq \tau_1$,

$$\frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(\tau_0, T)} \frac{\beta_t^2}{f(t)} \leq \frac{7}{16} \Delta (T - \tau_0), \quad (5)$$

The first part of the proposition is trivial by the choice of τ_1 . Now, we prove the second part. Fix $T > \tau_0 + \tau_1$. While Lemma 3 only considers the optimal selections by Alg, we improve this result by showing that the number of optimal selections grows linearly in T and combining it with Lemma 2. Specifically, we show that $N_{\text{opt}}(T) \geq \frac{1}{8}(T - \tau_0)$. We consider two cases. First, suppose Alg is executed at more than half of the time steps between $\tau_0 + 1$ and T , that is, $|\mathcal{T}_e \cap \{\tau_0 + 1, \dots, T\}| \geq \frac{1}{2}(T - \tau_0)$. Then, $f(T) \geq \frac{1}{2}(T - \tau_0)$. Since at least a quarter of the selections made by Alg are optimal after time step $t = \tau_0$, it holds that

$$N_{\text{opt}}(T) \geq \frac{1}{4} f(T) \geq \frac{1}{8} (T - \tau_0).$$

Now, we suppose the opposite. Consider the case where Alg is executed at fewer than half of the time steps between $t = \tau_0 + 1$ and T . Then, $\frac{1}{2}(T - \tau_0) \leq |\mathcal{G}(\tau_0, T)|$. We bound the number of suboptimal selections during the time steps in $\mathcal{G}(\tau_0, T)$ as follows:

$$\begin{aligned} \sum_{t \in \mathcal{G}(\tau_0, T)} \Delta \mathbb{1}\{X_t \neq x^*\} &\leq \mathcal{R}_{\text{INFEX}(\text{Alg}, \tau_e)}^G(\tau_0, T) \\ &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(\tau_0, T)} \frac{\beta_t^2}{f(t)} \\ &\leq \frac{7}{16} \Delta (T - \tau_0) \\ &\leq \frac{7}{8} \Delta |\mathcal{G}(\tau_0, T)|, \end{aligned}$$

where the first inequality uses that the non-zero instantaneous regret is at least Δ , the second inequality applies Lemma 3, the third inequality follows from Eq. (5), and the last inequality uses that $\frac{1}{2}(T - \tau_0) \leq |\mathcal{G}(\tau_0, T)|$. Therefore, we conclude that the number of suboptimal selections at time steps in $\mathcal{G}(\tau_0, T)$ is at most $\frac{7}{8} |\mathcal{G}(\tau_0, T)|$. It follows that the number of optimal selections among the same set of time steps is at least $\frac{1}{8} |\mathcal{G}(\tau_0, T)|$. Since at least a quarter of the exploratory selections are optimal, we have

$$\begin{aligned} N_{\text{opt}}(T) &\geq \frac{1}{8} |\mathcal{G}(\tau_0, T)| + \frac{1}{4} f(T) \\ &\geq \frac{1}{8} |\mathcal{G}(\tau_0, T)| + \frac{1}{8} (f(T) - \tau_{\text{Alg}}) \\ &= \frac{1}{8} (T - \tau_0), \end{aligned}$$

where the last equality comes from that $|\mathcal{G}(\tau_0, T)|$ and $f(T) - \tau_{\text{Alg}}$ are the numbers of greedy selections and exploratory selections during time steps $t = \tau_0 + 1, \dots, T$ respectively and hence their sum is $T - \tau_0$. We have proved that $N_{\text{opt}}(T) \geq \frac{1}{8}(T - \tau_0)$ for both cases. Plugging this bound

into Lemma 2, we conclude that

$$\begin{aligned}
\frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau_0 + \tau_1, T)} \frac{\beta_t^2}{1 + N_{\text{opt}}(t-1)} &\leq \frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau_0 + \tau_1, T)} \frac{\beta_t^2}{\frac{1}{8}(t - \tau_0)} \\
&\leq \frac{16\beta_T^2}{\Delta} \sum_{t \in \mathcal{G}(\tau_0 + \tau_1, T)} \frac{1}{t - \tau_0} \\
&\leq \frac{16\beta_T^2}{\Delta} \int_{\tau_0 + \tau_1}^T \frac{1}{x - \tau_0} dx \\
&= \frac{16\beta_T^2(\log(T - \tau_0) - \log \tau_1)}{\Delta} \\
&\leq \frac{16\beta_T^2 \log T}{\Delta},
\end{aligned}$$

where the first inequality holds since $1 + N_{\text{opt}}(t-1) \geq 1 + \frac{1}{8}(t-1-\tau_0) \geq \frac{1}{8}(t-\tau_0)$, the second inequality uses that β_t is increasing, and the third inequality upper bounds the summation by an integral since $1/(t-\tau_0)$ is decreasing in t . The proof is completed by plugging this bound into Lemma 2.

$$\begin{aligned}
\mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau_0 + \tau_1, T) &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau_0 + \tau_1, T)} \frac{\beta_t^2}{1 + N_{\text{opt}}(t-1)} \\
&= \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{16\beta_T^2 \log T}{\Delta}.
\end{aligned}$$

□

A.3 Bounds on $G_{\text{const}}(\tau_{\text{Alg}}, f)$

In this subsection, we provide bounds on $G_{\text{const}}(\tau_{\text{Alg}}, f)$. The steps of determining $G_{\text{const}}(\tau_{\text{Alg}}, f)$ in the proofs of Theorem 1 can be summarized as follows. First, take τ_1 such that for all $T \geq \tau_1$, it holds that

$$\frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(\tau_0, T)} \frac{\beta_t^2}{f(t)} \leq \frac{7}{16} \Delta(T - \tau_0),$$

which exists by Lemma 3. Then, define $\tilde{G}_{\text{const}} := 2S\tau_0 + \frac{7}{16}\Delta\tau_1$. Lastly, take $G_{\text{const}}(\tau_{\text{Alg}}, f) = \tilde{G}_{\text{const}} + \mathcal{O}(\frac{1}{\Delta}(d \log \tilde{G}_{\text{const}})^2)$. The value of $\tau_0 = f^{-1}(\tau_{\text{Alg}})$ is determined once f and τ_{Alg} are determined. It remains to provide an upper bound for τ_1 . We define additional constants whose bounds are easier to obtain. Let $\tau_{1,1} \in \mathbb{N}$ be the least time step such that $\tau_{1,1} \geq \tau_0$ and for all $T \geq \tau_0 + \tau_{1,1}$, it holds that

$$\frac{4\alpha_T \beta_T^2}{\Delta} \leq \frac{1}{16} \Delta T.$$

Since $\alpha_T, \beta_T^2 = \mathcal{O}(d \log T)$, we infer that $\tau_{1,1} = \max\{\tau_0, \mathcal{O}((\frac{d}{\Delta} \log \frac{d}{\Delta})^2)\} = \mathcal{O}(\tau_0 + (\frac{d}{\Delta} \log \frac{d}{\Delta})^2)$. Define $\tau_{1,2} \in \mathbb{N}$ to be the least time step such that for all $T \geq \tau_0 + \tau_{1,2}$, it holds that

$$\frac{8}{\Delta} \sum_{t \in \mathcal{G}(\tau_0, T)} \frac{\beta_t^2}{f(t)} \leq \frac{1}{4} \Delta(T - \tau_0).$$

The scale of $\tau_{1,2}$ depends on $f(t)$. Putting together, we obtain that for all $T \geq \tau_0 + \max\{\tau_{1,1}, \tau_{1,2}\}$, it holds that

$$\begin{aligned}
\frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(\tau_0, T)} \frac{\beta_t^2}{f(t)} &\leq \frac{1}{16} \Delta T + \frac{1}{4} \Delta(T - \tau_0) \\
&\leq \frac{3}{8} \Delta(T - \tau_0) \\
&\leq \frac{7}{16} \Delta(T - \tau_0),
\end{aligned}$$

where we use $T \leq 2(T - \tau_0)$ for the second inequality, which is implied by $T \geq \tau_0 + \tau_{1,1} \geq 2\tau_0$. Since τ_1 is the least value that satisfies the property above, we have $\tau_1 \leq \max\{\tau_{1,1}, \tau_{1,2}\}$. Then, we obtain that $\tilde{G}_{\text{const}} = \mathcal{O}(\tau_0 + \Delta\tau_{1,2} + \frac{d^2}{\Delta} \log^2 \frac{d}{\Delta})$. Additionally, note that for some universal constant $C > 0$, we have $\frac{d^2}{\Delta} \log^2 x \leq x$ for all $x \geq \frac{C}{\Delta} (d \log \frac{d}{\Delta})^2$. Therefore, we have $\frac{d^2}{\Delta} \log^2 x = \mathcal{O}(x + \frac{d^2}{\Delta} \log^2 \frac{d}{\Delta})$. It implies that

$$\begin{aligned} \tilde{G}_{\text{const}} + \mathcal{O}\left(\frac{1}{\Delta}(d \log \tilde{G}_{\text{const}})^2\right) &= \tilde{G}_{\text{const}} + \mathcal{O}\left(\tilde{G}_{\text{const}} + \frac{d^2}{\Delta} \log^2 \frac{d}{\Delta}\right) \\ &= \mathcal{O}\left(\tau_0 + \Delta\tau_{1,2} + \frac{d^2}{\Delta} \log^2 \frac{d}{\Delta}\right). \end{aligned}$$

Combining with Eq. (4) in the proof of Theorem 1, we obtain that

$$\begin{aligned} \mathcal{R}_{\text{INFEX}(\text{Alg}, \tau_e)}(T) &\leq \mathcal{R}_{\text{Alg}}(f(T)) + \mathcal{O}\left(\tau_0 + \Delta\tau_{1,2} + \frac{d^2}{\Delta} \log^2 \frac{d}{\Delta}\right) \\ &\quad + \mathcal{O}\left(\frac{1}{\Delta}\left(\log T + d \log \log T + d \log \frac{d}{\Delta}\right)^2\right) \\ &= \mathcal{R}_{\text{Alg}}(f(T)) + \mathcal{O}(\tau_0 + \Delta\tau_{1,2}) \\ &\quad + \mathcal{O}\left(\frac{1}{\Delta}\left(\log T + d \log \log T + d \log \frac{d}{\Delta}\right)^2\right), \end{aligned}$$

where in the last equality, the $\mathcal{O}(\frac{d^2}{\Delta} \log^2 \frac{d}{\Delta})$ term in the second term is absorbed into the last $\mathcal{O}(\frac{1}{\Delta} (\log T + d \log \log T + d \log \frac{d}{\Delta})^2)$ term. Therefore, there exists $G_{\text{const}}(\tau_{\text{Alg}}, f)$ and $G(T)$ such that $G_{\text{const}}(\tau_{\text{Alg}}, f) = \mathcal{O}(\tau_0 + \Delta\tau_{1,2})$, $G(T) = \mathcal{O}(\frac{1}{\Delta} (\log T + d \log \log T + d \log \frac{d}{\Delta})^2)$, and

$$\mathcal{R}_{\text{INFEX}(\text{Alg}, \tau_e)}(T) \leq \mathcal{R}_{\text{Alg}}(f(T)) + G_{\text{const}}(\tau_{\text{Alg}}, f) + G(T).$$

It remains to bound τ_0 and $\tau_{1,2}$. Let $C_\beta > 0$ be a constant independent of d, Δ , and T that satisfies $\beta_T^2 \leq C_\beta d \log(1 + T)$ for all T , which exists by Lemma 10. Let $\tau'_{1,2}$ be the least time step such that for all $T \geq \tau'_{1,2}$, it holds that

$$\frac{32C_\beta d \log(1 + 2T)}{\Delta^2} \sum_{t=1}^T \frac{1}{\max\{f(t), 1\}} \leq T. \quad (6)$$

We show that $\tau_{1,2} \leq \max\{\tau_0, \tau'_{1,2}\}$. For all $T \geq \tau_0 + \max\{\tau_0, \tau'_{1,2}\}$, it holds that

$$\begin{aligned} \frac{8}{\Delta} \sum_{t \in \mathcal{G}(\tau_0, T)} \frac{\beta_t^2}{f(t)} &\leq \frac{8\beta_T^2}{\Delta} \sum_{t=\tau_0+1}^T \frac{1}{f(t)} \\ &\leq \frac{8\beta_T^2}{\Delta} \sum_{t=1}^{T-\tau_0} \frac{1}{\max\{f(t), 1\}} \\ &\leq \frac{8C_\beta d \log(1 + T)}{\Delta} \sum_{t=1}^{T-\tau_0} \frac{1}{\max\{f(t), 1\}} \\ &\leq \frac{8C_\beta d \log(1 + 2(T - \tau_0))}{\Delta} \sum_{t=1}^{T-\tau_0} \frac{1}{\max\{f(t), 1\}} \\ &\leq \frac{1}{4} \Delta(T - \tau_0), \end{aligned}$$

where the first inequality holds since β_t is increasing, the second inequality uses that $f(t)$ is increasing and $f(t) \geq 1$ for $t \geq \tau_0 + 1$, the third inequality holds by the definition of C_β , and the fourth inequality is due to $T \geq 2\tau_0$, and the last inequality holds by the definition of $\tau'_{1,2}$. Therefore, we deduce that $\tau_{1,2} \leq \max\{\tau_0, \tau'_{1,2}\}$. Then, we have that $G_{\text{const}}(\tau_{\text{Alg}}, f) = \mathcal{O}(\tau_0 + \Delta\tau_{1,2}) = \mathcal{O}(\tau_0 + \Delta\tau'_{1,2})$.

For some example functions f , we provide bounds on $G_{\text{const}}(\tau_{\text{Alg}}, f)$ by providing bounds on τ_0 and $\tau'_{1,2}$. We write $f(t) = \Omega(g(t))$ for a function $g(t)$ when there exist constants $C_1, C_2 > 0$ such that $f(t) \geq C_1 g(t) - C_2$ for all $t \in \mathbb{N}$.

Example 1. Suppose $f(t) = \lfloor t/m \rfloor$ for some $m \in \mathbb{N}$. This case corresponds to executing Alg with a fixed period of m . We have $f^{-1}(n) = mn$, so $\tau_0 = m\tau_{\text{Alg}}$. We now establish a bound on $\tau'_{1,2}$ that satisfies Eq. (6). We have

$$\begin{aligned} \sum_{t=1}^T \frac{1}{\max\{f(t), 1\}} &\leq m + \sum_{t=m+1}^T \frac{m}{t-m} \\ &\leq m(1 + \log T). \end{aligned}$$

Using elementary analysis, one can show that after some time step $\tau = \mathcal{O}(\frac{md}{\Delta^2} \log^2 \frac{md}{\Delta})$, it holds that $\frac{32C_\beta md}{\Delta^2} (1 + \log T) \log(1 + 2T) \leq T$ for all $T \geq \tau$, hence $\tau'_{1,2} = \mathcal{O}(\frac{md}{\Delta^2} \log^2 \frac{md}{\Delta})$ holds. Combining the bounds on τ_0 and $\tau'_{1,2}$, we obtain

$$G_{\text{const}}(\tau_{\text{Alg}}, f) = \mathcal{O}\left(m\tau_{\text{Alg}} + \frac{md}{\Delta} \log^2 \frac{md}{\Delta}\right).$$

Example 2. Suppose $f(t) = \Omega(t/(\log t)^r)$ for some constant $r \geq 0$. Then, $f^{-1}(n) = \mathcal{O}(n(\log n)^r)$. Also, we have

$$\begin{aligned} \sum_{t=1}^T \frac{1}{\max\{f(t), 1\}} &= \sum_{t=1}^T \mathcal{O}\left(\frac{(\log t)^r}{t}\right) \\ &= \mathcal{O}((\log T)^{r+1}). \end{aligned}$$

$\tau'_{1,2}$ is the first time step such that $\mathcal{O}(\frac{d}{\Delta^2} (\log T)^{r+2}) \leq T$ for all $T \geq \tau'_{1,2}$, and we can derive that $\tau'_{1,2} = \mathcal{O}(\frac{d}{\Delta^2} (\log \frac{d}{\Delta})^{r+2})$. Therefore, we conclude that

$$G_{\text{const}}(\tau_{\text{Alg}}, f) = \mathcal{O}\left(\tau_{\text{Alg}} (\log \tau_{\text{Alg}})^r + \frac{d}{\Delta} \left(\log \frac{d}{\Delta}\right)^{r+2}\right).$$

Example 3. Let $f(t) = \Omega(t^r)$ for some constant $r \in (0, 1)$. Then, $f^{-1}(n) = \mathcal{O}(n^{1/r})$. We have

$$\begin{aligned} \sum_{t=1}^T \frac{1}{\max\{f(t), 1\}} &\leq \sum_{t=1}^T \mathcal{O}\left(\frac{1}{t^r}\right) \\ &= \mathcal{O}(T^{1-r}). \end{aligned}$$

For a constant $C > 0$, $CT^{1-r} \log T \leq T$ is equivalent to $(C \log T)^{1/r} \leq T$, and this inequality holds for all $T \geq \tau$ with $\tau = \mathcal{O}((C \log C)^{1/r})$. Therefore, we have that for $\tau'_{1,2} = \mathcal{O}((\frac{d}{\Delta^2} \log \frac{d}{\Delta})^{1/r})$, it holds that $\mathcal{O}(\frac{d}{\Delta^2} T^{1-r} \log T) \leq T$ for all $T \geq \tau'_{1,2}$. Therefore, we conclude that

$$G_{\text{const}}(\tau_{\text{Alg}}, f) = \mathcal{O}\left(\tau_{\text{Alg}}^{\frac{1}{r}} + \frac{1}{\Delta^{\frac{2}{r}-1}} \left(d \log \frac{d}{\Delta}\right)^{\frac{1}{r}}\right).$$

Example 4. Let $f(t) = \Omega((\log t)^r)$ for some constant $r > 1$. Then, $f^{-1}(n) = e^{\mathcal{O}(n^{1/r})}$. We have

$$\begin{aligned} \sum_{t=1}^T \frac{1}{\max\{f(t), 1\}} &= \sum_{t=1}^T \mathcal{O}\left(\frac{1}{(\log t)^r}\right) \\ &= \mathcal{O}\left(\frac{T}{(\log T)^r}\right). \end{aligned}$$

Then, $\tau'_{1,2}$ must satisfy $\frac{CdT}{\Delta^2 (\log T)^{r-1}} \leq T$ for some constant $C > 0$, or equivalently, $\frac{Cd}{\Delta^2} \leq (\log T)^{r-1}$. We see that $\tau'_{1,2} = \exp(\mathcal{O}((d/\Delta^2)^{1/(r-1)}))$. Therefore, we conclude that

$$G_{\text{const}}(\tau_{\text{Alg}}, f) = \exp\left(\mathcal{O}\left(\tau_{\text{Alg}}^{\frac{1}{r}}\right)\right) + \Delta \exp\left(\mathcal{O}\left((d/\Delta^2)^{\frac{1}{r-1}}\right)\right).$$

A.4 Proof of Theorem 3

Proof of Theorem 3. For simplicity, we write $\pi := \text{INFEX}(\text{Alg}, \mathcal{T}_e)$. We analyze the performance of π under two linear bandit instances. Let $\Delta > 0$ be a fixed constant whose value is chosen later. We define the arm set as $\mathcal{X} = \{e_1, \mathbf{0}_d\}$, where $e_1 \in \mathbb{R}^d$ is the first standard basis vector and $\mathbf{0}_d \in \mathbb{R}^d$ is the zero vector. This instance can be viewed as the one-armed bandit setting since the agent is aware that the second arm has reward 0. The true parameter vectors are defined as $\theta_1 = (-\Delta, 0, \dots, 0)$ and $\theta_2 = (\Delta, 0, \dots, 0)$. In the first instance, $(\mathcal{X}, \theta_1)$, the expected reward of the first arm is $-\Delta$, while the second arm yields a reward of 0. Thus, the second arm is the optimal arm. Conversely, in the second instance, $(\mathcal{X}, \theta_2)$, the first arm yields an expected reward of Δ and is the optimal arm. We assume that i.i.d. unit Gaussian noise is added to the observed reward.

Fix $T \in \mathbb{N}$. Let $N_1(T)$ and $N_2(T)$ be the number of times the first and second arms are selected up to time T , respectively. We define $\mathbb{P}_1$ to be the probability distribution over the trajectory induced by policy π interacting with instance $(\mathcal{X}, \theta_1)$ for T time steps, and define $\mathbb{P}_2$ similarly for the second instance $(\mathcal{X}, \theta_2)$.

Let $D_{\text{KL}}(\cdot, \cdot)$ be the KL-divergence between two probability measures. By Lemma 15.1 in Lattimore and Szepesvári [2020], we have that

$$D_{\text{KL}}(\mathbb{P}_1, \mathbb{P}_2) = 4\Delta^2 \mathbb{E}_1[N_1(T)].$$

Let $A := \{N_1(T) < T/2\}$ be the event that the first arm is selected less than $T/2$ times. By Lemma 12, we obtain that

$$\mathbb{P}_1(A) + \mathbb{P}_2(A^c) \geq \frac{1}{2} \exp(-D_{\text{KL}}(\mathbb{P}_1, \mathbb{P}_2)).$$

Under the first instance, we have $\mathcal{R}_\pi(T) = \Delta N_2(T)$. Using Markov's inequality, we obtain that $\mathbb{E}_1[N_2(T)] \geq \frac{T}{2} \mathbb{P}_1(N_2(T) \geq \frac{T}{2}) = \frac{T}{2} \mathbb{P}_1(N_1(T) < \frac{T}{2}) = \frac{T}{2} \mathbb{P}_1(A)$, which implies that $\mathbb{E}_1[\mathcal{R}_\pi(T)] \geq \frac{\Delta T}{2} \mathbb{P}_1(A)$. Using a similar argument, we also derive that $\mathbb{E}_2[\mathcal{R}_\pi(T)] \geq \frac{\Delta T}{2} \mathbb{P}_2(A^c)$. Combining everything, we conclude that

$$\begin{aligned} \mathbb{E}_1[\mathcal{R}_\pi(T)] + \mathbb{E}_2[\mathcal{R}_\pi(T)] &\geq \frac{\Delta T}{2} (\mathbb{P}_1(A) + \mathbb{P}_2(A^c)) \\ &\geq \frac{\Delta T}{4} \exp(-D_{\text{KL}}(\mathbb{P}_1, \mathbb{P}_2)) \\ &= \frac{\Delta T}{4} \exp(-4\Delta^2 \mathbb{E}_1[N_1(T)]). \end{aligned} \quad (7)$$

Now, we show that $\mathbb{E}_1[N_1(T)]$ increases too slowly when $f(t) \neq \omega(\log t)$. First, we show that the expected number of greedy selections of the first arm under the first instance is at most a constant. Let $\hat{\mu}_1(T)$ be the empirical mean of the first arm after T time steps. The greedy selection chooses the first arm only if $\hat{\mu}_1(T) \geq 0$. We bound the expected number of the averages of a Gaussian random walk exceeding Δ by the following lemma, whose proof is provided in Appendix C.5:

Lemma 5. *Let $Z_1, Z_2, \dots$ be a sequence of i.i.d. samples of the unit Gaussian distribution and $S_n = \sum_{t=1}^n Z_t$ be its partial sum. Then, for any constant $c > 0$, the expected number of indices n such that S_n/n exceeds c is at most $\frac{1}{2c^2}$, that is, $\mathbb{E}[\sum_{t=1}^\infty \mathbb{1}\{\frac{S_n}{n} \geq c\}] \leq \frac{1}{2c^2}$.*

For $\hat{\mu}_1(T) \geq 0$ to hold, the average of the noises added to the random rewards of the first arm must be greater than Δ . Using Lemma 5, we infer that

$$\mathbb{E}_1 \left[\sum_{t=1}^\infty \mathbb{1}\{X_t = e_1, \hat{\mu}_1(T) \geq 0\} \right] \leq \frac{1}{2\Delta^2}.$$

Therefore, the expected number of suboptimal greedy selections is at most $\frac{1}{2\Delta^2}$.

Therefore, we have $\mathbb{E}[N_1(T)] \leq \frac{1}{2\Delta^2} + f(T)$ since there are at most $\frac{1}{2\Delta^2}$ suboptimal greedy selections and $f(T)$ exploratory selections. By $f(t) \neq \omega(\log t)$, there exists a constant $C > 0$ and infinitely many $T \in \mathbb{N}$ such that $f(T) \leq C \log T$. We conclude that for infinitely many T , we have $\mathbb{E}[N_1(T)] \leq \frac{1}{2\Delta^2} + C \log T$. Plugging this bound into Eq. (7), we obtain that for infinitely many

Algorithm 2 Linear Thompson Sampling

```
1: Input : Sampling distribution  $\mathcal{D}^{\text{TS}}$ 
2: Initialize  $V_0 = I_d$ 
3: for  $t = 1, 2, \dots, T$  do
4:   Compute ridge estimator  $\hat{\theta}_{t-1} = V_{t-1}^{-1} \sum_{i=1}^{t-1} X_i Y_i$ 
5:   Sample  $\tilde{\eta}_t \sim \mathcal{D}^{\text{TS}}$ 
6:   Compute perturbed parameter  $\tilde{\theta}_t = \hat{\theta}_{t-1} + \beta_{t-1} V_{t-1}^{-1/2} \tilde{\eta}_t$ 
7:   Choose  $X_t = \operatorname{argmax}_{x \in \mathcal{X}} x^\top \tilde{\theta}_t$  and observe  $Y_t$ 
8:   Update  $V_t = V_{t-1} + X_t X_t^\top$ 
9: end for
```

$T \in \mathbb{N}$, it holds that

$$\begin{aligned} \mathbb{E}_1[\mathcal{R}_\pi(T)] + \mathbb{E}_2[\mathcal{R}_\pi(T)] &\geq \frac{\Delta T}{4} \exp\left(-4\Delta^2 \left(\frac{1}{2\Delta^2} + C \log T\right)\right) \\ &= \frac{\Delta}{4e^2} T^{1-4\Delta^2 C}. \end{aligned}$$

It implies that either $\mathbb{E}_1[\mathcal{R}_\pi(T)]$ or $\mathbb{E}_2[\mathcal{R}_\pi(T)]$ exceeds $\frac{\Delta}{8e^2} T^{1-4\Delta^2 C}$. The proof is completed by taking $\Delta = \sqrt{\varepsilon/4C}$ and $c(f, \varepsilon) = \frac{\Delta}{8e^2}$. $\square$

Remark 3. In the proof of Theorem 3, we show that $\mathbb{E}_1[N_1(T)] \leq (\frac{1}{2\Delta^2} + C \log T)$ and $\mathbb{E}_1[\mathcal{R}_\pi(T)] = \Delta \mathbb{E}_1[N_1(T)]$, so INFEX attains polylogarithmic regret for the first instance. Therefore, we can conclude that the instance that INFEX incurs almost linear regret is the second instance $(\mathcal{X}, \theta_2)$.

B Instance-Dependent Regret Analysis of Linear Thompson Sampling

In this section, we provide an instance-dependent polylogarithmic regret bound of LinTS [Agrawal and Goyal, 2013, Abeille and Lazaric, 2017]. For completeness, we present the algorithm in Algorithm 2, where we use the version by Abeille and Lazaric [2017].

The input of the algorithm, $\mathcal{D}^{\text{TS}}$, is a distribution over $\mathbb{R}^d$. We pose two conditions on the sampling distribution as in Abeille and Lazaric [2017].

1. (anti-concentration) There exists a positive probability p such that for any $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$,

$$\mathbb{P}_{\eta \sim \mathcal{D}^{\text{TS}}}(u^\top \eta \geq 1) \geq p.$$

2. (concentration) There exists positive constants c, c' such that for all $u \in \mathbb{R}^d$ with $\|u\|_2 = 1$ and $\delta \in (0, 1]$,

$$\mathbb{P}_{\eta \sim \mathcal{D}^{\text{TS}}} \left(|u^\top \eta| \leq \sqrt{c \log \frac{c'}{\delta}} \right) \geq 1 - \delta.$$

The first condition comes directly from Abeille and Lazaric [2017]. We slightly strengthen the second condition to derive a tighter bound when $\log K \ll d$. The original condition in Abeille and Lazaric [2017] poses that $\mathbb{P}_{\eta \sim \mathcal{D}^{\text{TS}}} \left(\|\eta\|_2 \leq \sqrt{cd \log(c'd/\delta)} \right) \geq 1 - \delta$. Our strengthened condition implies the original condition by taking u to be the vectors of the standard basis and taking the union bound. The strengthened condition holds for all the distributions discussed in Abeille and Lazaric [2017], including the multivariate Gaussian distribution and spherical distribution.

Now, assuming that the conditions are true, we prove Theorem 2.

Proof of Theorem 2. Let $\gamma_t := \beta_t(\delta) \min \left\{ \sqrt{cd \log(2c'dt^2/\delta)}, \sqrt{c \log(2c'Kt^2/\delta)} \right\}$. Our choice of γ_t slightly differs from Abeille and Lazaric [2017]; they choose it to be the first term in the minimum

instead of taking the minimum over the two values. We show that their analysis still applies even with this refined value of γ_t . Suppose $\gamma_t = \beta_t(\delta) \sqrt{c \log(2c' K t^2 / \delta)}$. By the concentration condition on $\mathcal{D}^{\text{TS}}$, for any $x \in \mathbb{R}^d$, it holds that

$$\mathbb{P}_{t-1} \left(x^\top (\beta_{t-1}(\delta) V_{t-1}^{-1/2} \tilde{\eta}_t) \leq \beta_{t-1}(\delta) \|x\|_{V_{t-1}^{-1/2}} \sqrt{c \log(2c' K t^2 / \delta)} \right) \geq 1 - \frac{\delta}{2t^2}.$$

Taking the union bound over $x \in \mathcal{X}$, we obtain

$$\begin{aligned} & \mathbb{P}_{t-1} \left(\forall x \in \mathcal{X}, x^\top (\beta_{t-1}(\delta) V_{t-1}^{-1/2} \tilde{\eta}_t) \leq \beta_{t-1}(\delta) \|x\|_{V_{t-1}^{-1/2}} \sqrt{c \log(2c' K t^2 / \delta)} \right) \\ &= \mathbb{P}_{t-1} \left(\forall x \in \mathcal{X}, x^\top (\beta_{t-1}(\delta) V_{t-1}^{-1/2} \tilde{\eta}_t) \leq \gamma_t \|x\|_{V_{t-1}^{-1/2}} \right) \\ &\geq 1 - \frac{\delta}{2t^2}. \end{aligned}$$

This probabilistic inequality is the only property γ_t must satisfy in the analysis of Abeille and Lazaric [2017], therefore the results in their paper hold for this refined value of γ_t .

We first decompose the instantaneous regret of LinTS as follows:

$$\begin{aligned} \text{reg}_t &= x^{*\top} \theta^* - X_t^\top \theta^* \\ &= \underbrace{x^{*\top} \theta^* - X_t^\top \tilde{\theta}_{t-1}}_{R_t^{\text{TS}}} + \underbrace{X_t^\top \tilde{\theta}_{t-1} - X_t^\top \theta^*}_{R_t^{\text{RLS}}}. \end{aligned}$$

Following the proof of Abeille and Lazaric [2017], we obtain that $R_t^{\text{TS}} \leq \frac{4\gamma_t}{p} \mathbb{E}_{t-1} \left[\|X_t\|_{V_{t-1}^{-1}} \right]$ and $R_t^{\text{RLS}} \leq \beta_t(\delta) \|X_t\|_{V_{t-1}^{-1}}$. By the definition of the minimum gap Δ , we have either $\text{reg}_t = 0$ or $\text{reg}_t \geq \Delta$, which implies that $\text{reg}_t \leq \frac{\text{reg}_t^2}{\Delta}$. Therefore, we derive the following bound on reg_t .

$$\begin{aligned} \text{reg}_t &\leq \frac{\text{reg}_t^2}{\Delta} \\ &= \frac{(R_t^{\text{TS}} + R_t^{\text{RLS}})^2}{\Delta} \\ &\leq \frac{2(R_t^{\text{TS}})^2 + 2(R_t^{\text{RLS}})^2}{\Delta} \\ &\leq \frac{2}{\Delta} \left(\frac{16\gamma_t^2}{p^2} \mathbb{E}_{t-1} \left[\|X_t\|_{V_{t-1}^{-1}}^2 \right] + \beta_t(\delta)^2 \|X_t\|_{V_{t-1}^{-1}}^2 \right) \\ &\leq \frac{2}{\Delta} \left(\frac{16\gamma_t^2}{p^2} \mathbb{E}_{t-1} \left[\|X_t\|_{V_{t-1}^{-1}}^2 \right] + \beta_t(\delta)^2 \|X_t\|_{V_{t-1}^{-1}}^2 \right), \end{aligned}$$

where the second inequality uses that $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$, and the last inequality is due to Jensen's inequality. We bound $\sum_{t=1}^T \mathbb{E}_{t-1} [\|X_t\|_{V_{t-1}^{-1}}^2]$ using the following lemma that provides a lower bound for a sum of nonnegative random variables. Its proof is provided in Appendix C.6.

Lemma 6. *Let $\{X_t\}_{t=1}^\infty$ be a sequence of real-valued random variables adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Suppose $0 \leq X_t \leq 1$ for all t . For any $\delta \in (0, 1]$, the following inequality holds for all $n \in \mathbb{N}$ with probability at least $1 - \delta$:*

$$\sum_{t=1}^n \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \leq 2 \sum_{t=1}^n X_t + 2 \log \frac{1}{\delta}.$$

Applying Lemma 6 on $\{\|X_t\|_{V_{t-1}^{-1}}^2\}_t$, we derive that with probability at least $1 - \delta$, it holds that

$$\sum_{t=1}^T \mathbb{E}_{t-1} \left[\|X_t\|_{V_{t-1}^{-1}}^2 \right] \leq 2 \sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}}^2 + 2 \log \frac{1}{\delta}.$$

for all $T \in \mathbb{N}$. Therefore, the cumulative regret of LinTS is bounded as follows:

$$\begin{aligned}\mathcal{R}_{\text{LinTS}}(T) &\leq \sum_{t=1}^T \frac{2}{\Delta} \left(\frac{16\gamma_t^2}{p^2} \mathbb{E}_{t-1} [\|X_t\|_{V_{t-1}^{-1}}^2] + \beta_t(\delta)^2 \|X_t\|_{V_{t-1}^{-1}}^2 \right) \\ &\leq \frac{2}{\Delta} \left(\left(\frac{32\gamma_T^2}{p^2} + \beta_T(\delta)^2 \right) \sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}}^2 + \frac{32\gamma_T^2}{p^2} \log \frac{1}{\delta} \right) \\ &\leq \frac{4}{\Delta} \left(\left(\frac{32\gamma_T^2}{p^2} + \beta_T(\delta)^2 \right) \alpha_T + \frac{16\gamma_T^2}{p^2} \log \frac{1}{\delta} \right),\end{aligned}$$

where the third inequality applies Lemma 11. Finally, plugging in $\beta_T(\delta)^2 = \mathcal{O}(\alpha_T + \log \frac{1}{\delta})$ and $\gamma_T^2 = \mathcal{O}(\min\{d \log \frac{dT}{\delta}, \log \frac{KT}{\delta}\}(\alpha_T + \log \frac{1}{\delta}))$ proves the theorem. $\square$

C Proofs of Technical Lemmas

In this section, we provide proofs of Lemmas 1 to 7.

C.1 Proof of Lemma 1

Proof of Lemma 1. Take $\tau_{\text{Alg}'}$ to be the least positive integer that satisfies

$$\frac{Cd^a}{\Delta^b} \log^c T \leq \frac{3}{4} \Delta T$$

for all $T \geq \tau_{\text{Alg}'}$, which exists since $\lim_{T \rightarrow \infty} \frac{\log^c T}{T} = 0$. Elementary analysis shows that $\tau_{\text{Alg}'} = \mathcal{O}(\frac{d^a}{\Delta^{b+1}} \log^c \frac{d}{\Delta})$. Let $N_{\text{sub}}(T)$ be the number of suboptimal selections made by Alg' up to time step T . Since a suboptimal selection incurs at least Δ regret, we have $\Delta N_{\text{sub}}(T) \leq \mathcal{R}_{\text{Alg}'}(T)$. It implies that for any $T \geq \tau_{\text{Alg}'}$, we have $\Delta N_{\text{sub}}(T) \leq \frac{3}{4} \Delta T$, or equivalently, $N_{\text{opt}}(T) \geq \frac{1}{4} T$, which proves the lemma. $\square$

C.2 Proof of Lemma 2

In this subsection, we prove Lemma 2. To do so, we show that the estimation error of the optimal reward $x^{*\top} \theta^*$ scales with $\frac{1}{\sqrt{N_{\text{opt}}(t)}}$, where we need the following technical lemma. Its proof is deferred to Appendix C.7.

Lemma 7. *We have that for all $t \in \mathbb{N}$,*

$$\|x^*\|_{V_t^{-1}}^2 \leq \frac{1}{1 + N_{\text{opt}}(t)}.$$

Now, we prove Lemma 2.

Proof of Lemma 2. The instantaneous regret of a greedy selection can be bounded as follows:

$$\begin{aligned}\text{reg}_t &= x^{*\top} \theta^* - X_t^\top \theta^* \\ &\leq x^{*\top} \theta^* - x^{*\top} \hat{\theta}_{t-1} + X_t^\top \hat{\theta}_{t-1} - X_t^\top \theta^* \\ &= x^{*\top} (\theta^* - \hat{\theta}_{t-1}) + X_t^\top (\hat{\theta}_{t-1} - \theta^*) \\ &\leq (\|x^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}}) \|\theta^* - \hat{\theta}_{t-1}\|_{V_{t-1}} \\ &\leq \beta_{t-1} (\|x^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}}),\end{aligned}$$

where the first inequality uses that $x^{*\top} \hat{\theta}_{t-1} \leq X_t^\top \hat{\theta}_{t-1}$ when X_t is chosen greedily, the second inequality is due to the Cauchy-Schwarz inequality, and the last inequality comes from Lemma 9.

By the definition of the minimum gap Δ , we have either $\text{reg}_t = 0$ or $\text{reg}_t \geq \Delta$, which implies that $\text{reg}_t \leq \frac{\text{reg}_t^2}{\Delta}$. Then, we obtain that

$$\begin{aligned} \text{reg}_t &\leq \frac{\text{reg}_t^2}{\Delta} \\ &\leq \frac{\beta_{t-1}^2 \left(\|x^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}} \right)^2}{\Delta} \\ &\leq \frac{2\beta_{t-1}^2 \left(\|x^*\|_{V_{t-1}^{-1}}^2 + \|X_t\|_{V_{t-1}^{-1}}^2 \right)}{\Delta}, \end{aligned}$$

where the last inequality uses that $(a + b)^2 \leq 2(a^2 + b^2)$ for any $a, b \in \mathbb{R}$. Taking the sum of instantaneous regret for $t \in \mathcal{G}(\tau, T)$, we proceed as follows:

$$\begin{aligned} \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(\tau, T) &= \sum_{t \in \mathcal{G}(\tau, T)} \text{reg}_t \\ &\leq \sum_{t \in \mathcal{G}(\tau, T)} \frac{2\beta_{t-1}^2 \left(\|x^*\|_{V_{t-1}^{-1}}^2 + \|X_t\|_{V_{t-1}^{-1}}^2 \right)}{\Delta} \\ &\leq \frac{2\beta_T^2}{\Delta} \sum_{t \in \mathcal{G}(\tau, T)} \|X_t\|_{V_{t-1}^{-1}}^2 + \frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau, T)} \beta_{t-1}^2 \|x^*\|_{V_{t-1}^{-1}}^2 \\ &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau, T)} \beta_{t-1}^2 \|x^*\|_{V_{t-1}^{-1}}^2 \\ &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{2}{\Delta} \sum_{t \in \mathcal{G}(\tau, T)} \frac{\beta_{t-1}^2}{1 + N_{\text{opt}}(t-1)}, \end{aligned}$$

where the third inequality is due to Lemma 11 and the last inequality applies Lemma 7. $\square$

C.3 Proof of Lemma 3

Proof of Lemma 3. By the choice of τ_{Alg} , at least a quarter of the selections by Alg are optimal when $f(t) \geq \tau_{\text{Alg}}$, or equivalently, $t \geq f^{-1}(\tau_{\text{Alg}})$. It implies that $N_{\text{opt}}(t) \geq \frac{1}{4}f(t)$. Then, it holds that $1 + N_{\text{opt}}(t-1) \geq 1 + \frac{1}{4}f(t-1) \geq 1 + \frac{1}{4}(f(t) - 1) \geq \frac{1}{4}f(t)$. Plugging this bound into Lemma 2, we conclude that

$$\begin{aligned} \mathcal{R}_{\text{INFEX}(\text{Alg}, \mathcal{T}_e)}^G(f^{-1}(\tau_{\text{Alg}}), T) &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{2}{\Delta} \sum_{t \in \mathcal{G}(f^{-1}(\tau_{\text{Alg}}), T)} \frac{\beta_{t-1}^2}{1 + N_{\text{opt}}(t-1)} \\ &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(f^{-1}(\tau_{\text{Alg}}), T)} \frac{\beta_{t-1}^2}{f(t)} \\ &\leq \frac{4\alpha_T \beta_T^2}{\Delta} + \frac{8}{\Delta} \sum_{t \in \mathcal{G}(f^{-1}(\tau_{\text{Alg}}), T)} \frac{\beta_t^2}{f(t)}. \end{aligned}$$

Now, we show that this quantity is sublinear in T . By Lemma 10, we have $\alpha_T, \beta_T^2 = \mathcal{O}(d \log T)$, so $\frac{4\alpha_T \beta_T^2}{\Delta}$ is sublinear in T . By $f(t) = \omega(\log t)$ and $\beta_t^2 = \mathcal{O}(d \log T)$, we have $\lim_{t \rightarrow \infty} \frac{\beta_t^2}{f(t)} = 0$, which implies that $\sum_{t \in \mathcal{G}(f^{-1}(\tau_{\text{Alg}}), T)} \frac{\beta_t^2}{f(t)}$ is sublinear in T . $\square$

C.4 Proof of Lemma 4

Proof of Lemma 4. We decompose V_T as follows:

$$\begin{aligned} V_T &= I_d + \sum_{t=1}^T X_t X_t^\top \\ &= I_d + \sum_{t=1}^T \mathbb{1}\{X_t = x^*\} X_t X_t^\top + \sum_{t=1}^T \mathbb{1}\{X_t \neq x^*\} X_t X_t^\top \\ &= I_d + N_{\text{opt}}(T) x^* x^{*\top} + \sum_{t=1}^T \mathbb{1}\{X_t \neq x^*\} X_t X_t^\top \\ &=: A + B, \end{aligned}$$

where we define $A := I_d + N_{\text{opt}}(T) x^* x^{*\top}$ and $B := \sum_{t=1}^T \mathbb{1}\{X_t \neq x^*\} X_t X_t^\top$. The eigenvalues of A are $1 + N_{\text{opt}}(T) \|x^*\|_2, 1, \dots, 1$. Let $b_1 \geq b_2 \geq \dots \geq b_d$ be the eigenvalues of B . Finally, let $v_1 \geq v_2 \geq \dots \geq v_d$ be the eigenvalues of V_T . By Lemma 13, we have

$$v_1 \leq (1 + N_{\text{opt}}(T) \|x^*\|_2) + b_1$$

and

$$v_i \leq \lambda_2(A) + b_{i-1} = 1 + b_{i-1}$$

for $i = 2, \dots, d$. Let $N_{\text{sub}}(T) := T - N_{\text{opt}}(T)$ be the number of suboptimal arm selections up to time T . Then, we have $b_1 \leq \text{tr}(B) \leq N_{\text{sub}}(T)$, so we infer that

$$v_1 \leq (1 + N_{\text{opt}}(T) \|x^*\|_2) + b_1 \leq 1 + N_{\text{opt}}(T) \|x^*\|_2 + N_{\text{sub}}(T) \leq 1 + T.$$

and

$$\begin{aligned} \prod_{i=2}^d v_i &\leq \prod_{i=2}^d (1 + b_{i-1}) \\ &\leq \left(\frac{\sum_{i=2}^d (1 + b_{i-1})}{d-1} \right)^{d-1} \\ &\leq \left(1 + \frac{\text{tr}(B)}{d-1} \right)^{d-1} \\ &\leq \left(1 + \frac{N_{\text{sub}}(T)}{d-1} \right)^{d-1}, \end{aligned}$$

where the second inequality is the AM-GM inequality. Then, we have

$$\begin{aligned} \alpha_T &= \log \frac{\det V_T}{\det V_0} \\ &= \sum_{i=1}^d \log v_i \\ &\leq \log(1 + T) + (d-1) \log \left(1 + \frac{N_{\text{sub}}(T)}{(d-1)} \right). \end{aligned}$$

Since a suboptimal selection incurs at least Δ regret, we have that $\Delta N_{\text{sub}}(T) \leq \mathcal{R}_{\text{Alg}'}(T)$, or equivalently, $N_{\text{sub}}(T) \leq \frac{1}{\Delta} \mathcal{R}_{\text{Alg}'}(T)$. Plugging in this bound completes the proof. $\square$

C.5 Proof of Lemma 5

Proof of Lemma 5. Let $\Phi(\cdot)$ be the cumulative density function of the standard Gaussian distribution. Since the distribution of S_n/n follows the Gaussian distribution with mean 0 and variance $1/n$, we have $\mathbb{P}(\frac{S_n}{n} \geq c) = 1 - \Phi(c\sqrt{n})$. Then, we have that

$$\begin{aligned} \mathbb{E} \left[\sum_{n=1}^{\infty} \mathbb{1} \left\{ \frac{S_n}{n} \geq c \right\} \right] &= \sum_{n=1}^{\infty} \mathbb{E} \left[\mathbb{1} \left\{ \frac{S_n}{n} \geq c \right\} \right] \\ &= \sum_{n=1}^{\infty} (1 - \Phi(c\sqrt{n})). \end{aligned}$$

Since $1 - \Phi(c\sqrt{n})$ is a decreasing function with respect to n , we can upper bound the summation by an integral and conclude as follows:

$$\begin{aligned} \sum_{n=1}^{\infty} (1 - \Phi(c\sqrt{n})) &\leq \int_0^{\infty} 1 - \Phi(c\sqrt{t}) dt \\ &= \int_0^{\infty} \int_{c\sqrt{t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx dt \\ &= \int_0^{\infty} \int_0^{\left(\frac{x}{c}\right)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dt dx \\ &= \int_0^{\infty} \left(\frac{x}{c}\right)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{2c^2}, \end{aligned}$$

where the first equality plugs in the probability density function of the Gaussian distribution and the second equality interchanges the order of the integral, which is justified by Fubini's theorem since the integrand is continuous and positive. $\square$

C.6 Proof of Lemma 6

Proof of Lemma 6. For simplicity, denote $\mathbb{E}[\cdot | \mathcal{F}_{t-1}]$ by $\mathbb{E}_{t-1}[\cdot]$. By $e^x \leq 1 + x + \frac{x^2}{2}$ for all $x \leq 0$, we have that

$$\begin{aligned} \mathbb{E}_{t-1}[e^{-X_t}] &\leq \mathbb{E}_{t-1}\left[1 - X_t + \frac{1}{2}X_t^2\right] \\ &= 1 - \mathbb{E}_{t-1}[X_t] + \frac{1}{2}\mathbb{E}_{t-1}[X_t^2] \\ &\leq 1 - \frac{1}{2}\mathbb{E}_{t-1}[X_t] \\ &\leq e^{-\frac{1}{2}\mathbb{E}_{t-1}[X_t]}, \end{aligned}$$

where the second inequality uses that $X_t \geq 0$ and $X_t^2 \leq X_t$ when $0 \leq X_t \leq 1$ and the last inequality holds since $1 + x \leq e^x$ for all $x \in \mathbb{R}$. Then, $M_n := \exp\left(\sum_{t=1}^n \left(-X_t + \frac{1}{2}\mathbb{E}_{t-1}[X_t]\right)\right)$ is a supermartingale. By Ville's maximal inequality, we have that $\mathbb{P}(\exists n \in \mathbb{N} : M_n \geq \frac{1}{\delta}) \leq \delta$. Taking the logarithm and rearranging the terms leads to the following conclusion:

$$\mathbb{P}\left(\exists n \in \mathbb{N} : \sum_{t=1}^n \mathbb{E}_{t-1}[X_t] \geq 2 \sum_{t=1}^n X_t + 2 \log \frac{1}{\delta}\right) \leq \delta.$$

$\square$

C.7 Proof of Lemma 7

We prove Lemma 7 by proving the following more general lemma.

Lemma 8. For $\lambda, n > 0$ and $x \in \mathbb{R}^d$, let V be a symmetric matrix with $V \succeq \lambda I_d + nxx^\top$. Then, $\|x\|_{V^{-1}}^2 \leq \frac{1}{\lambda+n}$.

Proof. It is sufficient to consider the case $V = \lambda I_d + nxx^\top$ only since $\|x\|_{V^{-1}}^2 \leq \|x\|_{(\lambda I_d + nxx^\top)^{-1}}^2$. In this case, we have

$$\begin{aligned} Vx &= \lambda x + nxx^\top x \\ &= (\lambda + n\|x\|_2^2)x. \end{aligned}$$

Multiply $x^\top V^{-1}$ on the left to both sides and obtain

$$\|x\|_2^2 = (\lambda + n\|x\|_2^2) \|x\|_{V^{-1}}^2.$$

By reordering the terms, we obtain that

$$\|x\|_{V^{-1}}^2 = \frac{\|x\|_2^2}{\lambda + n\|x\|_2^2} \leq \frac{1}{\lambda + n},$$

completing the proof. $\square$

D Auxiliary Lemmas

Recall that $\alpha_T = \log \frac{\det V_T}{\det V_0}$ and $\beta_T(\delta) = \sigma \sqrt{\alpha_T + 2 \log(1/\delta)} + S$.

Lemma 9 (Theorem 2 in Abbasi-Yadkori et al. [2011]). *With probability at least $1 - \delta$, $\|\theta^* - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta)$ holds for all $t \geq 0$.*

Lemma 10 (Lemma 10 in Abbasi-Yadkori et al. [2011]). *It holds that $\alpha_T \leq d \log(1 + \frac{T}{d})$.*

Lemma 11 (Lemma 11 in Abbasi-Yadkori et al. [2011]). *For any sequence of $X_1, \dots, X_T$ with $X_t \in \mathbb{B}^d$ for all $t = 1, \dots, T$, we have $\sum_{t=1}^T \|X_t\|_{V_{t-1}^{-1}}^2 \leq 2\alpha_T$.*

Lemma 12 (Bretagnolle-Huber inequality [Bretagnolle and Huber, 1979], Theorem 14.2 in Lattimore and Szepesvári [2020]). *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Let $D_{KL}(\mathbb{P}, \mathbb{Q}) := \int \log \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{P}$ be the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{Q}$. Then, for any event $A \in \mathcal{F}$, it holds that*

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2} \exp(D_{KL}(\mathbb{P}, \mathbb{Q})).$$

Lemma 13 (Weyl's inequality [Weyl, 1912]). *For a Hermitian matrix $A \in \mathbb{C}^{d \times d}$, let $\lambda_1(A) \geq \dots \geq \lambda_d(A)$ be its eigenvalues sorted from large to small. For two Hermitian matrices $A, B \in \mathbb{C}^{d \times d}$ and any $1 \leq i, j \leq d$ with $i + j - 1 \leq d$, it holds that*

$$\lambda_{i+j-1}(A + B) \leq \lambda_i(A) + \lambda_j(B).$$

E Extension to Time-Varying Features

In this section, we discuss the possibility of relaxing the assumption of requiring a finite and fixed arm set.

Previous literature on greedy bandit algorithms [Bastani et al., 2021, Kannan et al., 2018, Sivakumar et al., 2020, Raghavan et al., 2023, Kim and Oh, 2024] has established the effectiveness of purely greedy selections under certain favorable context distributions, specifically when features are drawn i.i.d. from distributions with suitable diversity conditions. Under such conditions, the regret contributions from the base exploratory algorithm and greedy selections can be analyzed separately. Moreover, since our analysis primarily assumes a fixed optimal arm x^* , the theoretical results provided in Theorem 1 readily extend to contexts where the optimal arm remains invariant.

However, an important and open challenge remains: extending the performance guarantees of INFEX to scenarios involving dynamically varying optimal arms. Addressing these more general cases is non-trivial, as our current analysis relies on the property that estimation errors of $x^{*\top} \theta^*$ diminish when the optimal arm is selected frequently. This property becomes less straightforward to guarantee when the optimal arm itself is random or time-varying. Notably, pointwise guarantees for linear regression with random design require additional distributional assumptions [Hsu et al., 2012], suggesting that bounding the estimation error of a random optimal arm without assumptions may be infeasible.

Meanwhile, Hanna et al. [2023] propose a reduction technique that enables linear bandit algorithms to address linear contextual bandit problems when the arm set is sampled i.i.d. from a fixed distribution. Their results, however, focus on worst-case $\mathcal{O}(\sqrt{T})$ -type regret, which is suboptimal in our context where instance-dependent polylogarithmic regret is desired. Additionally, while a greedy selection chooses the same arm irrespective of this reduction, the parameter update involves a mismatch: the observed reward Y_t from the selected arm X_t is attributed to a potentially different predetermined vector X'_t . Despite these challenges, the approach by Hanna et al. [2023] underscores the feasibility of adapting linear bandit methods to contextual scenarios, suggesting promising directions for extending our results in future work.

F Additional Experiments

We provide additional experimental results for different values of d omitted in Section 5. Except for the difference in the ambient dimension d , the generation of the problem instances and the algorithms is identical to those described in Section 5. Figure 2 presents the result when $d = 20$ and Figure 3 presents the result when $d = 40$. We observe the same trends as in the case where $d = 10$. Even for larger d , INFEX consistently demonstrates efficiency in both regret and computational time.

All hyperparameters of the algorithms are set to their theoretical values. Both LinUCB and LinTS require the confidence radius β_t . We explicitly compute the value of $\log \frac{\det V_t}{\det V_0}$ using rank-one update [Abbasi-Yadkori et al., 2011] instead of using its upper bound $d \log T$, so that the base algorithms achieve the regret bounds of Theorem 5 in Abbasi-Yadkori et al. [2011] and Theorem 2. We expect that the regret reduction achieved by INFEX would have been even more significant if the base algorithm had used a crude upper bound for the confidence radius.

The experiments are conducted on a computing cluster with twenty Intel(R) Xeon(R) Silver 4214R CPUs, and three of them are used for the experiments. The total runtime of the entire experiment is approximately one hour.

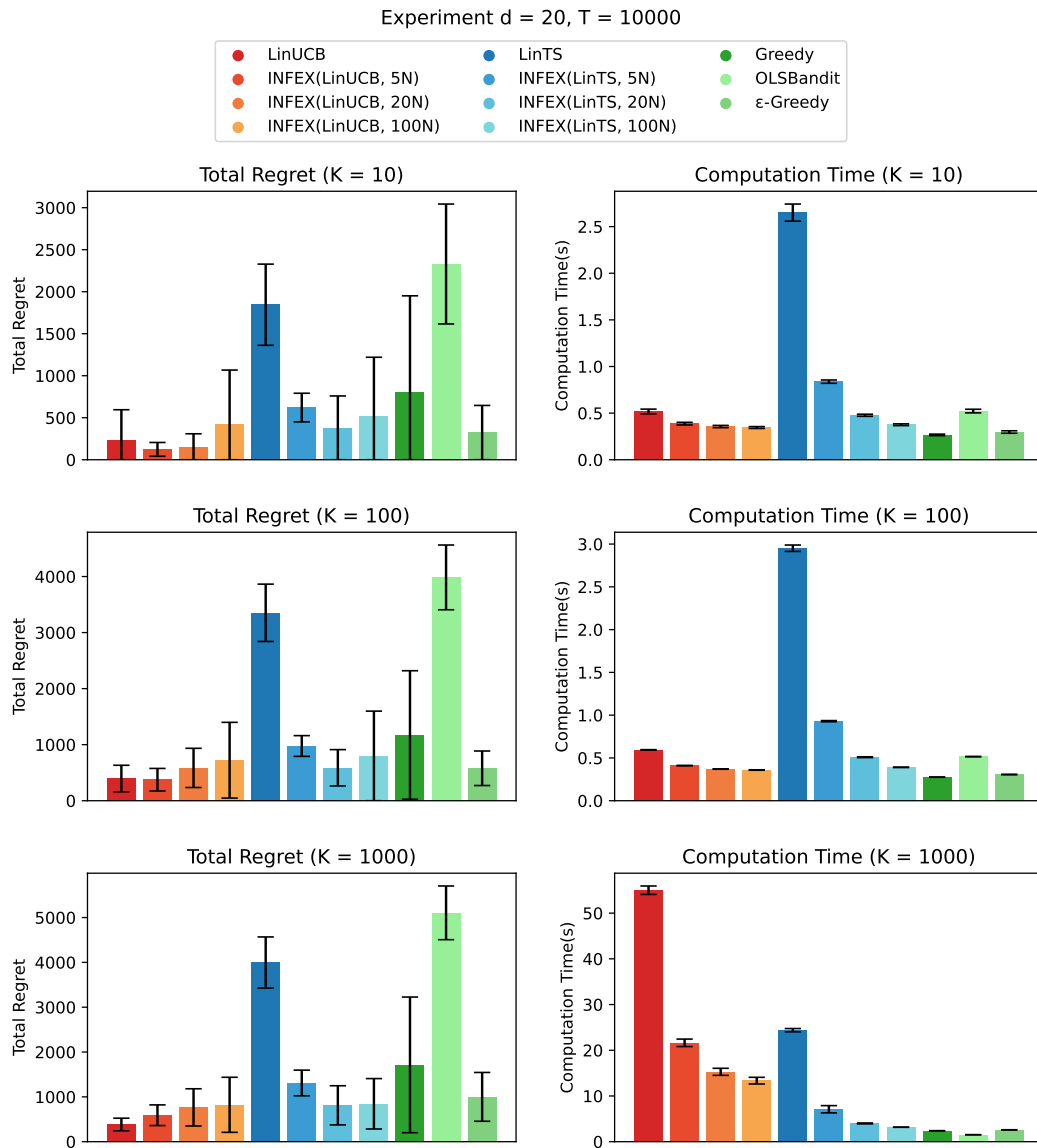


Figure 2: Comparison of total regret (left) and computation time (right) when $d = 20$, $T = 10000$, and $K = 10$ (top), $K = 100$ (middle), and $K = 1000$ (bottom).

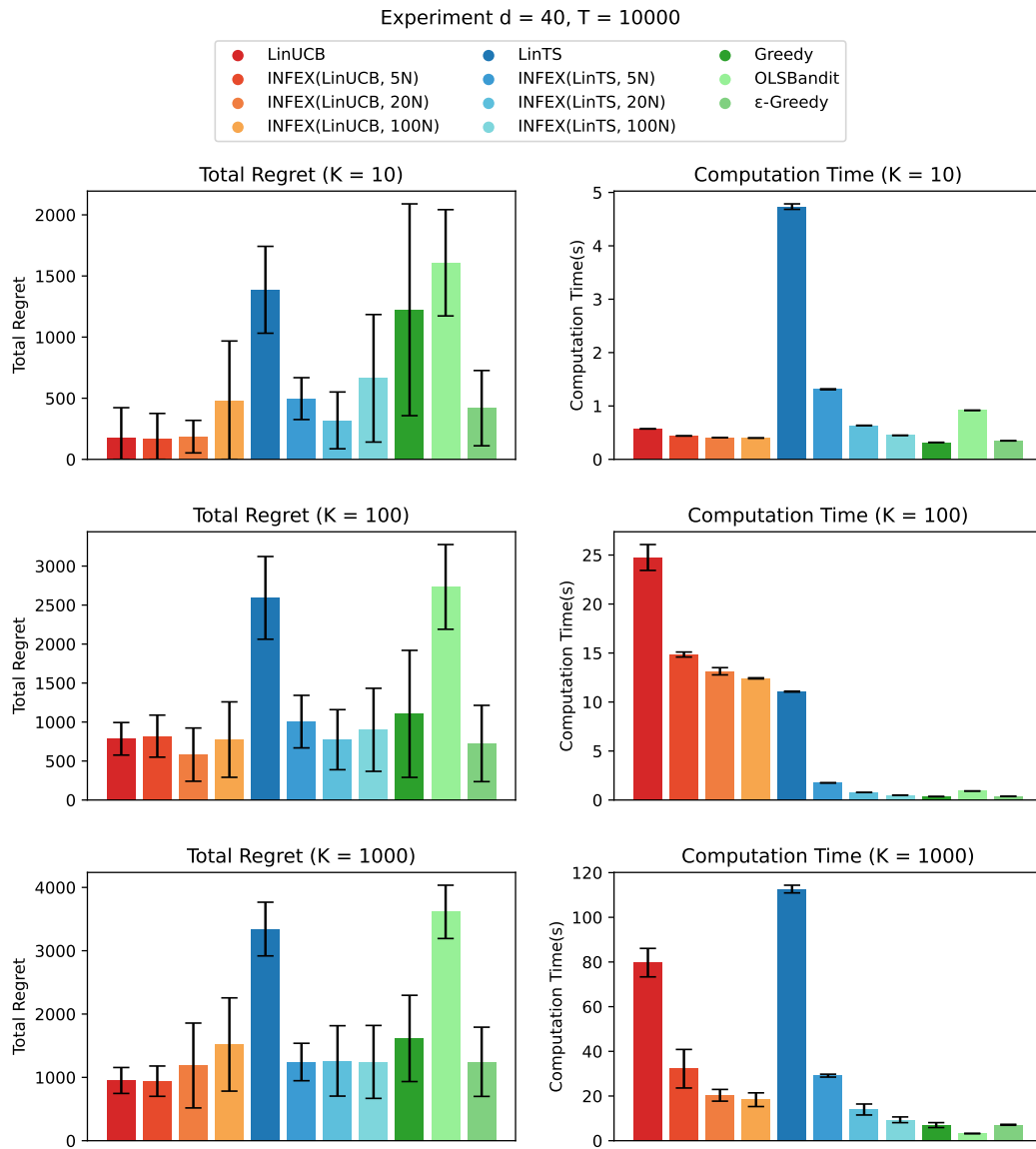


Figure 3: Comparison of total regret (left) and computation time (right) when $d = 40, T = 10000$, and $K = 10$ (top), $K = 100$ (middle), and $K = 1000$ (bottom).