On the Optimal Regret of Linear Contextual Bandit under Generalized Margin Condition

Xiaojie Mao*
School of Economics and Management, Tsinghua University, maoxj@sem.tsinghua.edu.cn

Zhiyuan Tang*
Naveen Jindal School of Management, University of Texas at Dallas, zhiyuan.tang@utdallas.edu

Yining Wang*
Naveen Jindal School of Management, University of Texas at Dallas, yining.wang@utdallas.edu

In this paper we study linear contextual bandit problems under a generalized margin condition. Unlike most existing works that focus primarily on explore-then-commit or greedy-type algorithms, we propose and analyze a successive elimination algorithm that carefully coordinates exploration and exploitation, improving existing algorithms and analysis in several aspects. First, our analysis shows that the algorithm achieves the *optimal* dependency on margin width and does so in an *adaptive* manner without prior knowledge of such margin parameters, neither of which is available in existing results. Second, we show that our algorithm's optimal dependency on margin parameters and context dimensionality applies to a wide range of problem scenarios under generalized margin conditions. In demonstrating such optimality, we construct novel lower bound instances including those based on non-Gaussian context distributions, which might be of independent interest for future research.

Key words: linear contextual bandits, rate–optimal policy, margin condition, adaptive algorithms.

1. Introduction

Contextual bandit is a fundamental question in operations management, with broad applications in precision medicine for healthcare management (Bastani and Bayati 2020), personalized recommendation and pricing in revenue management (Kallus and Udell 2020, Javanmard and Nazerzadeh 2019, Cohen et al. 2020), and contextual management of inventory systems in supply chain management (Besbes et al. 2023, Ding et al. 2024). In a conventional setup, a contextual bandit algorithm makes sequential decisions over T time periods, with access to changing *contextual information* (typically represented by a vector, containing information such as patients' vitals, consumers' personal information and historical purchase activities, or macroeconomic factors that would affect overall demand) at the beginning of each time period. The expected reward collected is characterized by *unknown* functions associated with the action taken, evaluated at the context vector at that particular time period.

* Authors are listed in alphabetical order.

Table 1 Summary of results.

Note: margin conditions are probability upper bounds on $\phi(h) = \mathbb{P}[|\Delta(x)| \le h]$ for $h > 0$, where $\Delta(x) = \langle x, \Delta_\beta \rangle$ and $\Delta_\beta = \beta^{(1)} - \beta^{(2)}$. Eigenvalue conditions are lower bounds $\underline{\lambda}$ on the minimum eigenvalue of the context covariance matrix restricted to regions of the form $\{x \in \mathbb{R}^d : \langle x, \psi \rangle \ge \rho\}$. In all big-O notations, only polynomial dependency on time horizon T, margin gap δ and context dimension d is traced. Adaptivity to δ means whether the algorithm needs to know δ in advance. Adaptivity to eigenvalues means whether the algorithm needs to know the eigenvalue lower bound $\underline{\lambda}$ in advance. GZ13 refers to Goldenshluger and Zeevi (2013); BB20 refers to Bastani and Bayati (2020); BBK21 refers to Bastani et al. (2021); LHO25 refers to Lee et al. (2025). ♯ These references studied high-dimensional context covariates; for convenience, we use d to represent sparsity levels. * It is pointed out in Goldenshluger and Zeevi (2013) that their margin and eigenvalue conditions together yield $\mathbb{P}[|\Delta(x)| \ge \kappa] = \Omega(1)$ with $\kappa = \delta$. An arm optimality condition with $\kappa > \delta$, however, is stronger and should be treated as an additional assumption because it cannot be implied by margin and eigenvalue conditions. * Although we limit our focus on the case of $\alpha \in (0, 1]$, our regret upper bound result in Theorem 1 can be easily extended to the case of $\alpha > 1$ to achieve an $\widetilde{\mathcal{O}}(d/\delta)$ regret rate.

Paper	Margin	Eigenvalue	Arm Opt.	Adapt. to δ	Adapt. to Eig.	Upper Bound	Lower Bound
GZ13	$\phi(h) \le h/\epsilon$ $\forall h \le \delta$	$\psi = \pm\Delta_\beta$ $\rho = \delta$	None	No	No	$\mathcal{O}\left(\frac{d^3}{\delta^2} + \frac{d^3 \log T}{\epsilon}\right)$	$\Omega(\log T)$
BB20♯	$\phi(h) \le h/\delta$	$\psi = \pm\Delta_\beta$ $\rho = \delta$	$\Pr[\lvert\Delta(x)\rvert \ge \kappa]$ $= \Omega(1)^*$	No	No	$\mathcal{O}\left(\frac{d^4}{\kappa^4} + \frac{d^2 \log^2 T}{\delta}\right)$	N/A
BBK21♯	$\phi(h) \le h/\delta$	$\forall \psi \in \mathbb{R}^d$ $\rho = \delta$	None	Yes	Yes	$\mathcal{O}\left(\frac{d \log T}{\delta}\right)$	N/A
LHO25♯	$\phi(h) \le (h/\delta)^\alpha$ $\alpha > 0^*$	$\psi = \pm\Delta_\beta$ $\rho = 0$	None	No	No	$\widetilde{\mathcal{O}}\left(\frac{d^{\frac{2\alpha+2}{\alpha}}}{\delta^2} + \frac{d^{\frac{1+\alpha}{\alpha}}}{\delta}\right)$ $+\mathcal{O}\left(\frac{d^{1+\alpha} T^{\frac{1-\alpha}{2}}}{\delta^\alpha}\right)$	N/A
This paper	$\phi(h) \le (h/\delta)^\alpha$ $\alpha \le 1^*$	$\psi = \pm\Delta_\beta$ $\rho = 0$	None	Yes	Yes	$\widetilde{\mathcal{O}}\left(T^{\frac{1-\alpha}{2}} d^{\frac{1+\alpha}{2}} \delta^{-\alpha}\right)$	$\Omega\left(T^{\frac{1-\alpha}{2}} d^{\frac{1+\alpha}{2}} \delta^{-\alpha}\right)$

In this paper, we focus on the *linear response bandit* model pioneered by the research of Goldenshluger and Zeevi (2013) and is subsequently followed up by numerous works such as Bastani et al. (2021), Bastani and Bayati (2020), He et al. (2022), Lee et al. (2025). In this model, the reward model for each action is a *linear function* of context variables, and certain eigenvalue and margin conditions are imposed to regulate context distributions, making them more amenable to online bandit learning. Sec. 2 of this paper gives mathematical details of the model setup and primary assumptions in this line of research.

Despite extensive studies, there are still several important questions remaining. One major shortcoming is that all existing algorithms in Bastani et al. (2021), Goldenshluger and Zeevi (2013), Bastani and Bayati (2020), Lee et al. (2025) are either explore-then-commit or myopic greedy algorithms that essentially separate exploration from exploitation. Such an approach, while still being asymptotically optimal in time horizon T, suffers from sub-optimal dependency and non-adaptivity of margin conditions, essentially requiring knowledge of margin condition parameters in advance before the algorithm starts. Furthermore, in Bastani et al. (2021), Goldenshluger and Zeevi (2013), Bastani and Bayati (2020), Lee et al. (2025), logarithmic regret is attained, which is in stark contrast to the $\widetilde{\mathcal{O}}(\sqrt{T})$ regret obtained in linear bandit without additional margin or eigenvalue constraints,

and it is unclear whether fundamental interpolation between logarithmic and square-root regret exists under weaker and perhaps more general margin conditions.

In this paper, we make important contributions addressing both the above major questions. Our contributions are summarized in Table 1 and the next subsection.

1.1. Our contributions

As shown in Table 1, existing works impose *margin conditions*, in addition to other assumptions on eigenvalues and action optimality gaps, that upper bound the probability of observing a context vector whose "margin" (i.e. difference in expected rewards between the two alternative actions) is small. More specifically, margin conditions can be roughly written as $\mathbb{P}[|\Delta(x)| \le h] \le (h/\delta)^\alpha$, which has two important parameters: δ, characterizing the "width" or "gap" of the margin on which the condition applies [1], and $\alpha \in (0, 1]$, characterizing the speed of probability decays within the margin gap when approaching the decision boundary.

Optimal and adaptive gap dependency. Our first contribution is to design an algorithm that achieves the *optimal* gap dependency $\widetilde{\mathcal{O}}(\delta^{-1})$ in the case of $\alpha = 1$, without requiring prior knowledge of δ. This improves existing $\widetilde{\mathcal{O}}(\delta^{-2})$ or $\widetilde{\mathcal{O}}(\delta^{-4})$ bounds from Goldenshluger and Zeevi (2013), Bastani and Bayati (2020), whose algorithms require knowing δ to properly set exploration durations. On the other hand, while Bastani et al. (2021) does achieve optimal δ dependency with a purely greedy algorithm, its analysis requires a substantially stronger eigenvalue condition (for *all* regions $\{x \in \mathbb{R}^d : \langle x, \psi \rangle \ge \delta\}$ with $\psi \in \mathbb{R}^d$ rather than a fixed region).

We achieve such optimal dependency and adaptivity of δ by a novel algorithm that uses successive elimination of arms for each context to balance exploration and exploitation. Our analysis, especially how it handles shifts of context distributions after each elimination, is significantly different from existing analyses of explore-then-commit or greedy algorithms.

We also remark the works of Hu et al. (2022), Gur et al. (2022) studied *non-parametric* contextual bandit problems with margin conditions, where the proposed algorithms also exhibit adaptivity to smoothness or margin parameters. It is important, however, to note that there are significant differences between non-parametric and parametric settings, such as the existence of curse of dimensionality and lower bounds on the probability density of context vectors. Additionally, lower bounds in non-parametric settings usually involve bump constructions which are not directly applicable for linear functions studied in this paper.

[1] The condition only applies to $|h| \le \delta$ because for other values of h the right-hand side is above 1, muting the condition.

Optimal regret characterization under generalized margin condition. We consider generalizing the margin condition by incorporating an exponent parameter $\alpha \in (0,1]$, with $\alpha = 1$ corresponding to the canonical existing condition and $\alpha \to 0^+$ reducing to linear contextual bandit without margin conditions. We show that the same successive elimination algorithm achieves cumulative regret $\widetilde{\mathcal{O}}(T^{\frac{1-\alpha}{2}} d^{\frac{1+\alpha}{2}} \delta^{-\alpha})$, which is minimax optimal in T, d and δ up to poly-logarithmic terms. Our results also cover $\widetilde{\mathcal{O}}(d/\delta)$ logarithmic regret when $\alpha = 1$ and $\widetilde{\mathcal{O}}(\sqrt{dT})$ when $\alpha = 0$, matching existing results in these two important special cases. For intermediate $\alpha \in (0,1)$ values, our results serve as interpolation between $\widetilde{\mathcal{O}}(\sqrt{T})$ and $O(\log T)$ regret.

To derive optimal regret characterization under generalized margin conditions, we propose novel lower bound arguments. More specifically, existing lower bounds such as that in He et al. (2022) use Normal distributions for instances of context vectors, which is *not* sufficient to capture general margin conditions when $\alpha < 1$. To address this, we use a mixture of Beta and normal distributions for contexts when $\alpha < 1$, and apply adjusted error metrics to accommodate our particular distributions. Our analysis also involves novel calculations of distribution distances, such as those connected with Cauchy distributions, that are drastically different from standard calculations with Normal distributions.

Additional improvements. In addition to main contributions above, our paper makes several additional improvements over existing results. For example, we derived the optimal regret dependency on context dimension d, which improves existing $O(d^3)$ or $O(d^4)$ upper bounds to $O(d)$ in the special case of $\alpha = 1$. Furthermore, we adopt eigenvalue conditions that are weaker than previous results, requiring only lower bounds on the minimum eigenvalue of context covariance matrix on the regions $\{x : \Delta(x) \geq 0\}$ and $\{x : \Delta(x) \leq 0\}$, rather than regions outside a strictly positive margin (e.g., $\Delta(x) \geq \delta$ and $\Delta(x) \leq -\delta$). Additionally, our algorithm requires no prior knowledge of the minimum eigenvalues, unlike many existing algorithms that need to use eigenvalue lower bounds to set lengths of exploration phases.

1.2. Additional Related works

Linear contextual bandit has been studied by Auer (2002), Abbasi-Yadkori et al. (2011) using UCB type methods, without additional margin or eigenvalue conditions. Their analysis, based on the principle of "optimism-in-face-of-uncertainty" (Abbasi-Yadkori et al. 2012), is quite different from ours. $\Omega(\sqrt{T})$ lower bounds for linear contextual bandit without margin conditions are proved in Bubeck et al. (2012). Their lower bound arguments are very different from ours because in our setting context vectors must be independently and identically distributed and satisfy margin/eigenvalue conditions. The works of Chen et al. (2024) also studied contextual bandits with sparse linear response models.

Auer et al. (2002) pioneered the study of contextual bandit with general reward functions, which was followed up by the important works of Agarwal et al. (2014), Simchi-Levi and Xu (2022), Foster and Rakhlin (2020). The works of Perchet and Rigollet (2013), Hu et al. (2022), Gur et al. (2022) studied contextual bandits with non-parametric reward functions, under function smoothness and generalized margin conditions. Goldenshluger and Zeevi (2009) studied a simple univariate contextual bandit problem under similar generalized margin conditions.

Recently, a growing body of literature has emerged that incorporates various structural assumptions into the framework considered in this paper, yielding a range of new results. For example, in the high-dimensional setting, Wang et al. (2024) investigates a related problem with generalized linear response functions; Duan et al. (2024) examines regret minimization, statistical inference, and their interplay under a similar formulation; and Xu and Bastani (2025) extends the problem to a multi-task learning framework. These efforts reflect the growing attention the problem has attracted in recent years. In this paper, we focus on the theoretical core of the problem and omit a comprehensive review of its various extensions.

2. Problem Setup

In this paper, we study a stylized two-armed contextual bandit problem. At the beginning of each time period $t \in [T]$, the algorithm observes a contextual vector $x_t \in \mathbb{R}^d$, where the contexts $x_1, x_2, \ldots, x_T$ are independently and identically distributed (i.i.d) from an unknown context distribution $\mathcal{P}_x$. After observing x_t, an action $a_t \in \{1,2\}$ is taken, and the algorithm subsequently receives a reward $Y_t = f^{(a_t)}(x_t) + \varepsilon_t$, where $f^{(a_t)}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is an unknown reward function associated with action a_t and ε_t are i.i.d. σ^2-subgaussian noises independent with x_t and a_t. In the linear response bandit problem, we assume

$$f^{(1)}(x) = \langle x, \beta^{(1)} \rangle, \quad f^{(2)}(x) = \langle x, \beta^{(2)} \rangle,$$

where $\beta^{(1)}, \beta^{(2)} \in \mathbb{R}^d$ are unknown d-dimensional linear model parameters.

A contextual bandit policy π that makes actions over consecutive time periods is *admissible* if it is non-anticipating. More specifically, let $\mathcal{F}_t$ be the σ-algebra representing the information available up to time t, which is generated by the distribution of $(x_1, a_1, y_1, x_2, a_2, y_2, \ldots, x_{t-1}, a_{t-1}, y_{t-1}, x_t)$. An admissible policy $\pi = \{\pi_t\}$ is a sequence of random functions $\pi_1, \pi_2, \ldots \pi_T$, such that for any $t \in [T]$, $\pi_t : \mathcal{X} \to \{1,2\}$ is $\mathcal{F}_t$ measurable. For any admissible policy π, we denote its cumulative regret over T time periods as

$$\text{Regret}_\pi(T) := \sum_{t=1}^{T} \left(\langle x_t, \beta^{(a_t^*)} \rangle - \langle x_t, \beta^{(a_t^\pi)} \rangle \right) \quad \text{where } a_t^* = \arg\max_{a \in \{1,2\}} \langle x_t, \beta^{(a)} \rangle$$

and a_t^π is the action prescribed by π_t for $t=1,\ldots,T$. Note that $\mathrm{Regret}_\pi(T)$ is a random variable and often its expectation or high-probability bounds are analyzed. The regret of a policy π characterizes the differences between this policy's rewards and those of the optimal policy in hindsight, with a smaller regret indicating a better algorithm/policy.

2.1. Main Assumptions

We first introduce two technical assumptions that are standard in the literature, assuming $\mathcal{P}_x$ has a light tail and $\beta^{(1)},\beta^{(2)}$ are bounded. Recall that $\mathcal{P}_x$ is the underlying distribution of contextual vectors and $\beta^{(1)},\beta^{(2)}$ are the linear model coefficients associated with the two actions.

Assumption A1 (Centered and Sub-gaussian Covariates) *The distribution $\mathcal{P}_x$ is centered and a^2-subgaussian, meaning that $\mathbb{E}[x]=0$ and for any $v\in\mathbb{R}^d$, $\|v\|_2=1$, $t>0$, we have $\mathbb{E}[e^{tv^\top x}]\le e^{a^2t^2/2}$.*

Assumption A2 (Boundedness of Coefficients) *There exists an absolute constant b_0 such that $\|\beta^{(1)}\|_2\le b_0$, $\|\beta^{(2)}\|_2\le b_0$.*

The first assumption assumes that $\mathcal{P}_x$ is centered and subgaussian, which is without loss of generality because shifting context vectors by their means can be easily carried out in practice. A similar assumption has also been adopted in prior studies within the bandit literature (e.g., Chen et al. 2024). Note that Assumption A1 encompasses the stronger condition that $\|x\|_2$ is bounded almost surely as a special case, which is a slight relaxation of conditions adopted in Bastani et al. (2021), Goldenshluger and Zeevi (2013). Note also that Assumption A1 does *not* assume that $\mathcal{P}_x$ has a bounded probability density, which is essential for our generalized margin conditions, as our lower bound problem instances involve unbounded density (see e.g. Figure 1).

Next, we impose a condition restricting the behavior of certain covariance matrices associated with $\mathcal{P}_x$ and $\beta^{(1)},\beta^{(2)}$. Let $\Delta(x):=\langle x,\Delta_\beta\rangle$ for $\Delta_\beta=\beta^{(1)}-\beta^{(2)}$ denote the expected difference between rewards from both arms conditioned on contextual vector x. Let $\mathbf{\Sigma}:=\mathbb{E}_{x\sim\mathcal{P}_x}[xx^\top]$ denote the full covariance matrix of $\mathcal{P}_x$. We also define *partial* covariance matrices as $\mathbf{\Sigma}_1:=\mathbb{E}_{x\sim\mathcal{P}_x}[xx^\top\mathbb{1}\{\Delta(x)>0\}]$ and $\mathbf{\Sigma}_2:=\mathbb{E}_{x\sim\mathcal{P}_x}[xx^\top\mathbb{1}\{\Delta(x)<0\}]$, which are the context covariance matrices restricted on the regions where one arm dominates the other arm (in expected rewards).

Assumption A3 (Lower Bounds on Eigenvalues) *There exists a numerical constant $\underline{\lambda}>0$ such that $\lambda_{\min}(\mathbf{\Sigma}_1)\ge\underline{\lambda}$, $\lambda_{\min}(\mathbf{\Sigma}_2)\ge\underline{\lambda}$.*

Assumption A3 essentially ensures that the contextual vectors are "diverse" (in the sense that the underlying covariance matrix is *not* near singular along any direction), and such diversity

continues to hold when restricted to the optimality region of either one of the two arms. While such a diversity assumption is not required for all prior works on linear contextual bandit (e.g. the works of Abbasi-Yadkori et al. (2011), Rusmevichientong and Tsitsiklis (2010) allow for adversarial context vectors), the model setups that are the closest to ours do impose, sometimes even stronger or more restrictive conditions on context diversity. For example, in Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020), the "diversity outside margin" assumption is imposed, which requires the diversity of contexts on regions $\{x : \Delta(x) \geq \rho\}$ and $\{x : \Delta(x) \leq -\rho\}$ where ρ is some strictly positive parameter. In contrast, our Assumption A3 only requires diversity regions $\{x : \Delta(x) \geq 0\}$ and $\{x : \Delta(x) \leq 0\}$, which is considerably weaker when $\rho > 0$ is large (since our restricted regions are larger). In Bastani et al. (2021), a more stringent diversity assumption is imposed, requiring the contexts to be diverse in regions $\{x : \langle x, \psi \rangle > 0\}$ for *all* $\psi \in \mathbb{R}^d$, under which certain greedy algorithms are near optimal. In contrast, our Assumption A3 is only imposed for $\psi \in \{\Delta_\beta, -\Delta_\beta\}$, making it much weaker[2].

Our final assumption characterizes the concept of "margin" in linear contextual bandit. Unlike the previous assumptions where constants involved are typically numerical constants, in the definition of margin conditions below there are two parameters $\delta, \alpha > 0$ that we do *not* treat as numerical constants but track their dependency throughout our regret analysis.

Assumption A4 (Generalized Margin Condition) *A problem instance $\mathcal{P}_x$, $\beta^{(1)}$ and $\beta^{(2)}$ satisfies the generalized* margin condition *with parameters $\delta > 0$ and $\alpha \in (0, 1]$ if for every $h > 0$ it holds that* [3]

$$\mathbb{P}_{x \sim \mathcal{P}_x}\big[|\Delta(x)| < h\big] \leq (h/\delta)^\alpha. \tag{1}$$

The generalized margin condition (Assumption A4), on the other hand, ensures that the probability of a data point x being close to the decision boundary (where $|\Delta(x)|$ is small) is controlled. Specifically, the closer a point is to the boundary, the less likely it is to occur, following at most a polynomial rate characterized by parameters δ and α. The parameter δ characterizes the "width" of the margin and the parameter α characterizes the speed at which the probabilities of contexts in the margin region should decrease to zero. The case $\alpha = 0$ imposes essentially no restriction on the context distribution, while $\alpha = 1$ recovers the classical margin condition commonly assumed in prior work (Goldenshluger and Zeevi 2013, Bastani and Bayati 2020). To capture a broader range of margin behaviors, we adopt a more flexible formulation inspired by Lee et al. (2025), which accommodates all $\alpha \in (0, 1]$ and interpolates between these two extremes.

[2] Several recent works have proposed additional conditions under which greedy-type algorithms are sufficient to achieve sublinear regret; see, e.g., Oh et al. (2021), Kim and Oh (2024), Ren and Zhou (2024). Nevertheless, our assumptions are strictly more general and cover many scenarios where all these conditions fail to apply.

[3] Note that in (1) if $h \geq \delta$ the right-hand side is greater than or equal to one, essentially muting the condition.

Remark 1 *We focus on the case $\alpha \in (0,1]$, rather than the more general $\alpha > 0$ considered in, e.g., Lee et al. (2025), as most algorithms—including ours—achieve the same regret rate for all $\alpha \geq 1$. In particular, our upper bound in Theorem 1 naturally extends to the regime $\alpha > 1$, yielding a regret rate of $\widetilde{\mathcal{O}}(d/\delta)$. This includes the special case $\alpha = +\infty$, which corresponds to the well-separated or minimum-gap setting studied in Abbasi-Yadkori et al. (2011), Goldenshluger and Zeevi (2013). The question of whether our regret rate is minimax optimal for general $\alpha > 1$ remains open in the linear bandit literature.*

3. Proposed Algorithm: Successive Elimination

In this section, we propose our algorithm for the problem introduced above. The algorithm is outlined in Algorithm 1.

Algorithm 1 Linear Bandit with Successive Elimination

input: Parameter ω_0

1: $\widehat{\beta}_0^{(i)} \leftarrow 0$, $\mathbf{V}_0^{(i)} \leftarrow \mathbf{0}$, $\rho_0^{(i)} \leftarrow +\infty$, for $i = 1, 2$;
2: **for** each time period $t \in [T]$ **do**
3: Observe context vector $x_t \in \mathbb{R}^d$;
4: Calculate $L_t^{(i)}(x_t) \leftarrow \langle x_t, \widehat{\beta}_{t-1}^{(i)} \rangle - \rho_{t-1}^{(i)}\omega_0$, $U_t^{(i)}(x_t) \leftarrow \langle x_t, \widehat{\beta}_{t-1}^{(i)} \rangle + \rho_{t-1}^{(i)}\omega_0$ for $i = 1, 2$;
5: **if** $[L_t^{(1)}(x_t), U_t^{(1)}(x_t)] \cap [L_t^{(2)}(x_t), U_t^{(2)}(x_t)] = \emptyset$ **then**
6: Take greedy actions $a_t = 1$ if $L_t^{(1)}(x_t) > U_t^{(2)}(x_t)$ and vice versa;
7: **else**
8: Take a random action $a_t \sim \text{Uniform}(\{1, 2\})$;
9: Update $\mathbf{V}_t^{(a_t)} \leftarrow \mathbf{V}_{t-1}^{(a_t)} + x_t x_t^\top$; $\rho_t^{(i)} \leftarrow \rho_{t-1}^{(i)}$ for $i = 1, 2$;
10: Observe reward Y_t;
11: **if** $\min\left\{\lambda_{\min}\big(\mathbf{V}_{t-1}^{(1)}\big), \lambda_{\min}\big(\mathbf{V}_{t-1}^{(2)}\big)\right\} > 0$ **then**
12: Update estimates $\widehat{\beta}_t^{(i)} = \arg\min_{\beta \in \mathbb{R}^d} \sum_{\tau \in [t]} (Y_\tau - \langle x_\tau, \beta \rangle)^2 \mathbb{1}\{a_\tau = i\}$ for $i = 1, 2$;
13: Update $\rho_t^{(i)} \leftarrow \sqrt{\lambda_{\max}(\mathbf{V}_t^{(i)})/\lambda_{\min}\big(\mathbf{V}_t^{(i)}\big)}$ for $i = 1, 2$.

At a high level, the algorithm maintains a candidate set of arms for each context and will eliminate the "bad" arm if there is convincing evidence. Initially, the confidence radius ρ_0 is set to infinity (Line 1), resulting in purely random exploration until sufficient data is gathered to ensure the cumulative Gram matrices $\mathbf{V}_t^{(i)}$ for $i = 1, 2$ are non-singular, so that the ordinary least squares (OLS) estimators in Line 12 are well-defined. After that, the algorithm conducts successive elimination based on the OLS estimators.

At each time period t, upon observing the context vector x_t, the algorithm uses estimators $\widehat{\beta}_{t-1}^{(1)}, \widehat{\beta}_{t-1}^{(1)}$ and confidence parameter ρ_{t-1} computed from previous observations, to construct confidence intervals for the rewards of both arms (Line 4). Specifically, the confidence intervals are scaled by a predefined factor ω_0, which determines the width of these intervals and thus controls the exploration intensity. If one confidence interval strictly dominates the other—i.e., the lower bound of one's interval exceeds the upper bound of the other interval—the algorithm eliminates the inferior arm and selects the dominant arm (Line 6). If no such dominance is observed, the algorithm selects an arm uniformly at random (Line 8). At the end of time period t, the algorithm calculates estimators for both parameters $\widehat{\beta}_t^{(1)}$ and $\widehat{\beta}_t^{(2)}$ using all previously collected samples via OLS regression (Line 12). Correspondingly, the algorithm recalculates the confidence radius $\rho_{t-1}^{(i)}$ by the eigenvalues of the cumulative Gram matrices $\mathbf{V}_t^{(i)}$ for $i = 1, 2$ at the end of time period t (Line 13).

Now we compare Algorithm 1 with existing approaches in the literature. Algorithms addressing the problem formulated in this paper can be broadly categorized into two types: greedy algorithms and non-greedy algorithms. Greedy algorithms prioritize immediate rewards, selecting the arm that appears optimal at each time step based on the current information, whereas non-greedy algorithms balance exploitation and exploration, accounting for the uncertainty in the current information and exploring seemingly suboptimal arms to acquire more information for better future decision-making. Our algorithm belongs to the non-greedy category, as it explores both arms when lacking confident evidence for which arm is the best.

Broadly speaking, non-greedy algorithms, including the one proposed in this paper, offer greater general applicability compared to greedy ones. Greedy approaches, such as those presented in Bastani et al. (2021) and Oh et al. (2021), achieve optimal regret rates only when contexts exhibit sufficient diversity across all directions. This can often be overly stringent and difficult to satisfy in practice. In contrast, non-greedy algorithms are capable of attaining optimal error rates under more relaxed assumptions, enhancing their robustness across a wider range of settings.

We next discuss the differences between our algorithm and existing non-greedy algorithms. The primary advantage of our algorithm lies in its **gap-adaptive** design, which eliminates the need to specify the gap parameter δ as an algorithm input. In contrast, existing non-greedy bandit algorithms require δ as a predefined input because they rely on random exploration for $\mathcal{O}(1/\delta^2)$ rounds, such that estimates within δ precision can be obtained, after which the margin condition can be leveraged. When δ is misspecified, the performance of such algorithms may degrade due to their non-adaptivity to δ. Our algorithm, in contrast, only requires minimal knowledge of the problem parameters. In particular, in Theorem 1, we will show that the knowledge of δ and $\underline{\lambda}$ is not required to achieve a favourable regret upper bound. This adaptive character is enabled by the dynamic adjustment of the exploration intensity according to the data collected on-the-fly.

4. Main Result: Regret Upper Bound

We present the following theorem as the first main result of this paper, establishing upper bounds on the worst-case expected cumulative regret of Algorithm 1 under different margin condition parameters.

Theorem 1 (Regret Upper Bound) *Suppose Assumptions A1 to A4 hold. Set in Algorithm 1* $\omega_0 = 2a\sigma\sqrt{3(d+\log T)\log(2T^4)}$. *The expected regret of the policy* π *produced by Algorithm 1 then satisfies*

$$\mathbb{E}\left[\mathrm{Regret}_\pi(T)\right] = \widetilde{\mathcal{O}}\left(\min\left\{\sqrt{dT}, T^{\frac{1-\alpha}{2}} d^{\frac{1+\alpha}{2}} \delta^{-\alpha}\right\}\right),$$

where in the $\widetilde{O}(\cdot)$ *notation we omit numerical constants as well as dependency on* $a, b_0, \underline{\lambda}$ *in Assumptions.*

The upper bound in Theorem 1 tracks dependency on three parameters: d, T, and δ. In the following discussion, we analyze how the regret upper bound scales with respect to each of these parameters.

First, considering only dependency on time horizon T, our result reveals that $\mathbb{E}\left[\mathrm{Regret}_\pi(T)\right] = \widetilde{\mathcal{O}}\left(T^{(1-\alpha)/2}\right)$. It shows that, as α approaches 1, the cumulative regret decreases, which is intuitive because a larger α value indicates a faster decrease in probability of near-margin context samples, implying that the arms are indistinguishable at fewer contexts and therefore the inherent difficulty of the problem decreases. Conversely, when α approaches 0, the regret rate converges to $\widetilde{\mathcal{O}}(\sqrt{T})$, corresponding to the hardest case in which the margin condition is mute (when $\alpha = 0$, the right-hand side of (1) is one and the margin condition is satisfied for all $\mathcal{P}_x$). Our results generalize existing results in Bastani et al. (2021), Goldenshluger and Zeevi (2013) that obtained $O(\log T)$ regret in the special case of $\alpha = 1$ of the generalized margin condition.

Second, considering only dependency on covariate dimension d, our result shows that $\mathbb{E}[\mathrm{Regret}_\pi(T)] = \widetilde{\mathcal{O}}(d^{(1+\alpha)/2})$. This improves the $O(d^2)$ order established in Bastani and Bayati (2020) and the $O(d^3)$ rate in Bastani et al. (2021). Such improvement is made possible by our analysis that carefully exploits the centrality and sub-Gaussianity of $\mathcal{P}_x$, so that additional d factors from conventional Cauchy-Schwarz inequality arguments could be avoided. Our dependency on the covariate dimension is *tight*, as confirmed by our regret lower bound presented in section 5. We note that there is already empirical evidence suggesting that the dependency in previous results on dimension d is conservative; see, e.g., Table 1 in Goldenshluger and Zeevi (2013). Our algorithm goes beyond this empirical observation by providing a rigorous guarantee that achieves the optimal dependency on d.

Third, considering only dependency on gap parameter δ, in the margin condition, our result shows that $\mathbb{E}[\text{Regret}_\pi(T)] = \mathcal{O}(\delta^{-\alpha})$ which is the first such result in the literature with correct exponent on δ for all $\alpha \in (0, 1]$. For the special case of $\alpha = 1$, Bastani and Bayati (2020), Goldenshluger and Zeevi (2013), Lee et al. (2025) obtained *sub-optimal* upper bounds of $\mathcal{O}(\delta^{-2})$, with algorithms that are *not* adaptive to δ and require prior knowledge of δ to set up the correct number of exploration rounds. The work of Bastani et al. (2021), on the other hand, does achieve the correct $\mathcal{O}(\delta^{-1})$ scaling but only under substantially more restrictive eigenvalue assumptions that would allow a simple greedy algorithm to work.

Finally, we remark that the $\min\{(dT)^{1/2}, T^{(1-\alpha)/2} d^{(1+\alpha)/2} \delta^{-\alpha}\}$ upper bound shows how the relationship between horizion length T, covariate number d and margin paramter δ influences the regret rate. Specifically, when $T \preceq d\delta^{-2}$, the regret is dominated by the term $\sqrt{dT}$; otherwise, the term $T^{(1-\alpha)/2} d^{(1+\alpha)/2} \delta^{-\alpha}$ becomes dominant. We note the latter case is of particular interest, as long-term performance is typically the primary concern in sequential decision-making problems.

Theorem 2 later in this paper demonstrates that our upper bound is minimax optimal with respect to all three parameters, T, d, and δ – constituting the first such result in the literature.

In the remainder of this section we prove Theorem 1. We first present an important technical lemma for the validity of confidence intervals constructed in Algorithm 1. Afterwards, regret cumulated in each step of the algorithm is upper bounded by using the confidence interval width and the margin condition.

4.1. A Technical Lemma on Confidence Intervals

We first present an important lemma in Lemma 1. This lemma shows that the confidence intervals in Algorithm 1 are valid after a certain time step t_0 when choosing the radius parameter ω_0 in Theorem 1.

Lemma 1 (High Probability Bound of Confidence Interval) *Suppose that the assumptions and conditions listed in Theorem 1 hold. Then for any $t > t_0$, $\mathbb{P}[\mathcal{G}_t] \geq 1 - 10T^{-2}$ where*

$$\mathcal{G}_t := \left\{ \left| \langle x_t, \widehat{\beta}_{t-1}^{(i)} - \beta^{(1)} \rangle \right| \leq \rho_{t-1}^{(i)} \omega_0 \text{ and } \rho_{t-1}^{(i)} \leq \frac{\sqrt{8\underline{\lambda} + 16a^2}}{\underline{\lambda}\sqrt{t-1}}, \text{ for } i = 1, 2 \right\} \tag{2}$$

and

$$t_0 := \left\lceil \left(d + 3\log T\right) \max\left\{ \frac{256a^4}{\underline{\lambda}^2}, \frac{16a^2}{\underline{\lambda}} \right\} \right\rceil. \tag{3}$$

Lemma 1 establishes that our confidence intervals are valid and the interval width is well bounded with high probability . Building on this, we will later demonstrate that the regret rate derived from these confidence intervals is minimax optimal, implying that they are not only valid but also

tight—that is, they achieve minimal width. This property is crucial in our regret analysis because when the true rewards of both arms fall within their respective confidence intervals, the incurred regret is on the order of interval widths. Therefore, tighter confidence intervals directly lead to a lower regret, provided their validity is maintained.

The proof of Lemma 1 primarily relies on the analysis of ℓ_2 error of estimators $\widehat{\beta}_t^{(1)}$ and $\widehat{\beta}_t^{(2)}$ for each $t \in [T]$. A key step in this analysis is ensuring that the minimum eigenvalues of the cumulative Gram matrices $\mathbf{V}_t^{(i)}$ corresponding to both arms $i = 1, 2$ grow linearly with t. However, directly analyzing the full cumulative Gram matrix is challenging due to the strong dependency structure among samples, which arises from the adaptive decision-making process. To address this, we focus on a subset of samples that are actually independent of each other, with this independence ensured by the randomization step in our algorithm. The complete proof of Lemma 1 is given in Appendix A.

4.2. Proof of Theorem 1

Fix an index $t > t_0, t \in [T]$. We already determined our policy π and thus we use a_t for a_t^π in this proof. Let $r_t = |\Delta(x_t)| \mathbb{1}\{a_t \neq a_t^*\}$ and $\rho_{t-1} = \max\{\rho_{t-1}^{(1)}, \rho_{t-1}^{(2)}\}$. First note that, if $a_t = 2$, then we must have $L_t^{(1)}(x_t) - U_t^{(2)}(x_t) \leq 0$ from definition of our algorithm. This implies

$$\begin{aligned}
&\Delta(x_t)\mathbb{1}\{a_t^* = 1, a_t = 2\} \leq (\langle x_t, \beta^{(1)} - \beta^{(2)} \rangle + U_t^{(2)}(x_t) - L_t^{(1)}(x_t))\mathbb{1}\{a_t^* = 1\} \\
&\leq \left|\langle x_t, \widehat{\beta}_{t-1}^{(2)} \rangle - \langle x_t, \beta^{(2)} \rangle\right| + \left|\langle x_t, \beta^{(1)} \rangle - \langle x_t, \widehat{\beta}_{t-1}^{(1)} \rangle\right| + 2 \cdot \rho_{t-1}\omega_0 =: \gamma_t^{(1)}(x_t) + \gamma_t^{(2)}(x_t) + 2\rho_{t-1}\omega_0,
\end{aligned}$$

where $\gamma_t^{(i)}(x) = |\langle x, \widehat{\beta}_{t-1}^{(i)} - \beta^{(i)} \rangle|$ for $i = 1, 2$. Similarly we have $-\Delta(x_t)\mathbb{1}\{a_t^* = 2, a_t = 1\} \leq \gamma_t^{(2)}(x_t) + \gamma_t^{(1)}(x_t) + 2\rho_{t-1}\omega_0$. Therefore we have

$$|\Delta(x_t)| \mathbb{1}\{a_t \neq a_t^*\} \leq 2\left(\gamma_t^{(2)}(x_t) + \gamma_t^{(1)}(x_t) + 2\rho_{t-1}\omega_0\right). \tag{4}$$

Now recall the "good" event $\mathcal{G}_t$ we defined in (2). We decompose the total regret into

$$\mathbb{E}\left[\text{Regret}(T)\right] = \mathbb{E}\left[\sum\nolimits_{t \in [t_0]} r_t\right] + \mathbb{E}\left[\sum\nolimits_{t=t_0+1}^{T} r_t \mathbb{1}\{\mathcal{G}_t^c\}\right] + \mathbb{E}\left[\sum\nolimits_{t=t_0+1}^{T} r_t \mathbb{1}\{\mathcal{G}_t\}\right]. \tag{5}$$

For the third term on the right-hand side of (5), note that when $\mathcal{G}_t$ holds, (4) gives us $r_t = |\Delta(x_t)| \mathbb{1}\{a_t \neq a_t^*\} \leq 8\rho_{t-1}\omega_0 \leq 8\omega_0\sqrt{8\underline{\lambda} + 16a^2}/(\underline{\lambda}\sqrt{t-1})$. This implies that when $a_t \neq a_t^*$ and $\mathcal{G}_t$ holds, $|\Delta(x_t)| \leq 8\omega_0\sqrt{8\underline{\lambda} + 16a^2}/(\underline{\lambda}\sqrt{t-1})$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\sum\nolimits_{t=t_0+1}^{T} r_t \mathbb{1}\{\mathcal{G}_t\}\right] \leq \mathbb{E}\left[\sum\nolimits_{t=t_0+1}^{T} 8\omega_0 \frac{\sqrt{8\underline{\lambda} + 16a^2}}{\underline{\lambda}\sqrt{t-1}} \mathbb{1}\left\{|\Delta(x_t)| \leq 8\omega_0 \frac{\sqrt{8\underline{\lambda} + 16a^2}}{\underline{\lambda}\sqrt{t-1}}\right\}\right] \\
&\leq \sum_{t=t_0+1}^{T} 8\omega_0 \frac{\sqrt{8\underline{\lambda} + 16a^2}}{\underline{\lambda}\sqrt{t-1}} \min\left\{\left(\frac{8\omega_0\sqrt{8\underline{\lambda} + 16a^2}}{\delta\underline{\lambda}\sqrt{t-1}}\right)^{\alpha}, 1\right\} = \widetilde{\mathcal{O}}\left(\min\left\{\omega_0^{1+\alpha} T^{(1-\alpha)/2} \delta^{-\alpha}, \omega_0\sqrt{T}\right\}\right),
\end{aligned} \tag{6}$$

where the second inequality is given by the margin condition in Assumption A4 and the last inequality is given by the calculation of the power sum. For the term $\mathbb{E}\left[\sum_{t=t_0+1}^{T} r_t \mathbb{1}\{\mathcal{G}_t^c\}\right]$, by the Cauchy-Schwarz inequality and Lemma 1 we have

$$\mathbb{E}\left[\sum_{t=t_0+1}^{T} r_t \mathbb{1}\{\mathcal{G}_t^c\}\right] \leq \sum_{t=t_0+1}^{T} \sqrt{\mathbb{E}\left[r_t^2\right]}\sqrt{\mathbb{P}\left[\mathcal{G}_t^c\right]} \leq \sum_{t=t_0+1}^{T} 2ab_0\sqrt{10T^{-2}} \leq 2\sqrt{10}ab_0. \tag{7}$$

Similarly, for the term $\mathbb{E}[\sum_{t\in[t_0]} r_t]$ in (5), we have

$$\mathbb{E}\left[\sum_{t\in[t_0]} r_t\right] \leq t_0 \max_{t\in[t_0]} \sqrt{\mathbb{E}[r_t^2]} \leq 2ab_0t_0. \tag{8}$$

Therefore, by combining (5),(6),(7) and (8), the total regret can be upper bounded by

$$\begin{aligned}\mathbb{E}\left[\mathrm{Regret}(T)\right] &\leq 2ab_0t_0 + 2\sqrt{10}ab_0 + \widetilde{\mathcal{O}}\left(\min\left\{\omega_0^{1+\alpha}T^{(1-\alpha)/2}\delta^{-\alpha}, \omega_0\sqrt{T}\right\}\right) \\ &= \widetilde{\mathcal{O}}\left(t_0 + \min\left\{\omega_0^{1+\alpha}T^{(1-\alpha)/2}\delta^{-\alpha}, \omega_0\sqrt{T}\right\}\right).\end{aligned}$$

By the value choices in the Theorem statement we have

$$t_0 = \lceil (d+3\log T)\max\{256a^4/\underline{\lambda}^2, 16a^2/\underline{\lambda}\}\rceil = \widetilde{\mathcal{O}}(d), \omega_0 = 2a\sigma\sqrt{3(d+\log T)\log(2T^4)} = \widetilde{\mathcal{O}}(\sqrt{d}),$$

we can finally conclude that $\mathbb{E}\left[\mathrm{Regret}_\pi(T)\right] = \widetilde{\mathcal{O}}(\min\{\sqrt{dT}, T^{\frac{1-\alpha}{2}}d^{\frac{1+\alpha}{2}}\delta^{-\alpha}\})$. This completes the proof.

5. Main Result: Regret Lower Bound

In this section, we establish the worst-case regret lower bound for any admissible bandit policy and show that it matches our regret upper bound in Theorem 1. Our main result is the following, which is a direct result of Lemma 2 given later in this section.

Theorem 2 *Suppose $d \geq 9$ and $T = \Omega(d\delta^{-2})$. Then for any $\alpha \in (0,1]$ and $\delta \in (0,1)$, the worst-case regret of any admissible policy is lower bounded by $\Omega(T^{\frac{1-\alpha}{2}}d^{\frac{1+\alpha}{2}}\delta^{-\alpha})$.*

Comparing Theorem 2 with the regret upper bounds in Theorem 1, we observe that both bounds match in polynomial dependency on all three major problem parameters: the time horizon T, the covariate dimension d, and the parameters δ and α in our generalized margin condition.

When $\alpha = 1$, Theorem 2 yields a regret lower bound of $\Omega(d/\delta^2)$. In contrast, existing results of Goldenshluger and Zeevi (2013), He et al. (2022) established $\Omega(\log T)$ regret lower bound in this case, but did not track dependency on dimension d or margin gap parameter δ. Our lower bounds show that the minimax regret scales linearly in d and inversely quadratically in δ, which matches our regret upper bound in Theorem 1. Our results offer additional insights into the fundamental difficulty of this question, because covariate dimension and margin width (δ) are important and

could potentially be much larger than $\log T$ factors in applications with short or intermediate time horizons.

Furthermore, Theorem 2 provides the first result explicitly characterizing the lower bound dependency on d, T and δ for $\alpha \in (0,1)$ to the best of our knowledge. The lower bound of $\Omega(T^{(1-\alpha)/2} d^{(1+\alpha)/2} \delta^{-\alpha})$ shows how the margin parameters δ and α affect the problem difficulty.

In the remainder of this section we show how to obtain Theorem 2. We first introduce a general reduction, similar to Lemma 2 in He et al. (2022).

5.1. Reduction to Estimation Error Lower Bound

An estimator $\widehat{\Delta}_\beta \in \mathbb{R}^d$ is said to be T-admissible if it is $\mathcal{F}_T$ measuarable, where $\mathcal{F}_T$ is the σ-algebra we introduced in Secton 2 involving all (adaptively collected) samples over T periods. Our goal in this section is to establish a connection between the minimax regret of an admissible policy and the minimax estimation error of a T-admissible estimator, the latter of which is easier to lower bound subsequently. For lower bound purposes, noises are distributed as $\varepsilon_t \sim \mathcal{N}(0,1)$ for all $t \in [T]$. We define $\mathcal{P}(a, b_0, \underline{\lambda}, \delta, \alpha)$ as the collection of all triplets $(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)})$ satisfying Assumptions A1 to A4 with parameters $a, b_0, \underline{\lambda}, \delta$, and α. Because we only track dependency on T, d and δ, we assume that a, b_0 and $\underline{\lambda}$ are fixed constants, and henceforth write $\mathcal{P}(a, b_0, \underline{\lambda}, \delta, \alpha)$ as $\mathcal{P}(\delta, \alpha)$.

Given T, d and δ, the minimax regret and minimax estimation error over $\mathcal{P}(\delta, \alpha)$ are defined as

$$\begin{aligned}
\mathfrak{R}_{\delta,\alpha}(T) &:= \inf_{\pi} \sup_{(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)}) \in \mathcal{P}(\delta,\alpha)} \mathrm{Regret}_\pi(T), \\
\mathfrak{E}_{\delta,\alpha}(T) &:= \inf_{\widehat{\Delta}_\beta} \sup_{(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)}) \in \mathcal{P}(\delta,\alpha)} \mathbb{E}\left[\mathbb{P}_{x \sim \mathcal{P}_x}\left[\mathrm{sgn}\left(\langle x, \Delta_\beta \rangle\right) \neq \mathrm{sgn}\left(\langle x, \widehat{\Delta}_\beta \rangle\right)\right]\right].
\end{aligned} \tag{9}$$

The definition of minimax regret $\mathfrak{R}_{\delta,\alpha}(T)$ involves the infimum over all admissible policies $\pi = (\pi_1, \ldots, \pi_T)$, where each π_t is measurable to $\mathcal{F}_t$. In contrast, the definition of minimax estimation error $\mathfrak{E}_{\delta,\alpha}(T)$ involves the infimum over all T-admissible estimators $\widehat{\Delta}_\beta$ that are measurable to the largest σ-algebra $\mathcal{F}_T$. We note also that for $\mathfrak{E}_{\delta,\alpha}(T)$, the estimation error is defined as the probability that the signs of $\langle x, \beta^{(1)} - \beta^{(2)} \rangle$ and $\langle x, \widehat{\Delta}_\beta \rangle$ differ, instead of conventional ℓ_2 errors, which is more amenable to our lower bound analysis.

We then have the following lemma, reducing bandit optimization to statistical estimation by formally lower bounding the bandit minimax regret with minimax estimation errors.

Lemma 2 (Lower Bound Reduction) *For any $\delta \in (0,1), \alpha \in (0,1]$, we have*

$$\mathfrak{R}_{\delta,\alpha}(T) \geq 4^{-1-1/\alpha} \left[\mathfrak{E}_{\delta,\alpha}(T)\right]^{1+1/\alpha} \delta T.$$

This lemma demonstrates that the regret lower bound $\mathfrak{R}_{\delta,\alpha}(T)$ for all admissible policies $\pi = (\pi_1, \ldots, \pi_T)$ can be directly derived from the estimation error lower bound for all T-admissible estimators $\widehat{\Delta}_\beta$. Consequently, deriving a regret lower bound can be reduced to establishing an estimation error lower bound. The proof of this lemma is primarily based on Lemma 2 in He et al. (2022) and is deferred to the appendix. In the following sections, we focus primarily on deriving a lower bound of $\mathfrak{E}_{\delta,\alpha}(T)$.

5.2. Lower Bounds on Estimation Error of Admissible Estimators

The main purpose of this section is to discuss the following theorem lower bounding $\mathfrak{E}_{\delta,\alpha}(T)$:

Lemma 3 (Estimation Error Lower Bound) *Suppose $d \geq 9$. There exists a universal constant $\mathfrak{c}_1 > 0$ such that for $\alpha = 1$ and any $\delta \in (0,1)$, we have*

$$\mathfrak{E}_{\delta,1}(T) \geq \mathfrak{c}_1, \quad \textit{for all } T \leq 0.01 d\delta^{-2}. \tag{10}$$

Furthermore, for any $\alpha \in (0,1)$ and $\delta \in (0,1)$, there exists an α-dependent constant $\mathfrak{c}_\alpha > 0$ such that

$$\mathfrak{E}_{\delta,\alpha}(T) \geq \mathfrak{c}_\alpha \left(\frac{d}{T}\right)^{\alpha/2} \delta^{-\alpha}, \quad \textit{for all } T \geq 256^{-1}(d-1)\delta^{-2}\left[\Gamma\left(\frac{1-\alpha}{2}\right)\right]^{-2/\alpha}. \tag{11}$$

When $\alpha = 1$, (10) combined with Lemma 2 yield a regret lower bound of $\Omega(d/\delta^2)$ by considering the regret incurred in the first $\mathcal{O}(d/\delta^2)$ time periods. On the other hand, (11) shows that when $T \gtrsim d\delta^{-2}$, the minimax estimation error $\mathfrak{E}_{\delta,\alpha}(T)$ will be on the order of $\Omega\big((dT^{-1}\delta^{-2})^{\alpha/2}\big)$, which is consistent with (10) because when $T \lesssim d\delta^{-2}$ the minimax estimation error rate $\Omega\big((dT^{-1}\delta^{-2})^{\alpha/2}\big)$ is of constant order.

On the other hand, the inequality (11) aims to show the specific minimax estimation error for different $\alpha \in (0,1)$ when $T \gtrsim d\delta^{-2}$, as the rate $\Omega\big((dT^{-1}\delta^{-2})^{\alpha/2}\big)$ will decay faster for larger α and slower for smaller α. Combining (11) with Lemma 2 directly gives us the regret lower bound in Theorem 2 for $\alpha \in (0,1)$.

The proof of Lemma Theorem is carried out by constructing a suitable distribution $\mathcal{P}_x$ and a set of parameter pairs $(\beta^{(1)}, \beta^{(2)})$, denoted by $\mathcal{B}^{(1)} \times \mathcal{B}^{(2)}$, where $\{\mathcal{P}_x\} \times \mathcal{B}^{(1)} \times \mathcal{B}^{(2)} \subseteq \mathcal{P}_{\delta,\alpha}$. Using Fano's method, we then show that with a constant probability that we cannot identify the true parameter pair $(\beta^{(1)}, \beta^{(2)})$ if it is chosen uniformly at random from $\mathcal{B}^{(1)} \times \mathcal{B}^{(2)}$. This leads to the conclusion that the minimax estimation error exceeds the minimum distance between any two tuples in $\mathcal{B}^{(1)} \times \mathcal{B}^{(2)}$.

The key challenge in the proof lies in constructing the suitable distribution $\mathcal{P}_x$ and set $\mathcal{B}^{(1)} \times \mathcal{B}^{(2)}$ such that the generalized margin condition holds for all $(\beta^{(1)}, \beta^{(2)}) \in \mathcal{B}^{(1)} \times \mathcal{B}^{(2)}$ and $\mathcal{P}_x$. For the special case of $\alpha = 1$, He et al. (2022) shows that $\mathcal{P}_x$ can be chosen as a normal distribution or a truncated normal distribution for the class $\mathcal{P}(\delta, 1)$. However, existing literature provides no clear

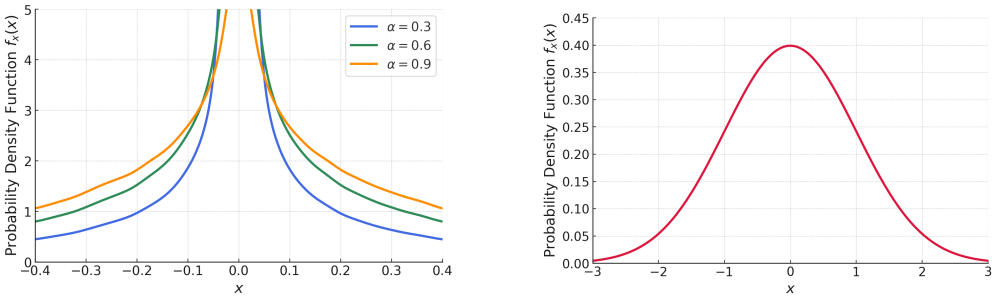

Figure 1 Probability density functions of constructed $\mathcal{P}_x$ (left panel) and $\mathcal{N}(0,1)$ (right panel). Note the constructed density diverges at zero which is consistent with margin conditions of parameters $\alpha < 1$.

guidance on the choice of $\mathcal{P}_x$ for the class $\mathcal{P}(\delta,\alpha)$ when $\alpha \in (0,1)$. When $\alpha \in (0,1)$, a tight lower bound requires a distribution $\mathcal{P}_x$ such that $\mathbb{P}_{x\sim\mathcal{P}_x}[|\Delta(x)| \le h] \propto h^\alpha$ for all $h > 0$, which yields an unbounded density of x near the decision boundary. In this case, many commonly used distributions such as the normal distribution and the Bernoulli distribution, are not suitable for the choice of $\mathcal{P}_x$ to obtain a tight lower bound for $\alpha \in (0,1)$. In our proof, we address this by bringing in the Beta distribution, which yields an unbounded density near $x = 0$. Furthermore, we mix the Beta distribution with components of Normal distributions. We adjust the shape parameters of both distributions, so that the generalized margin condition for different $\alpha \in (0,1)$ is satisfied. We illustrate the probability density function for our constructed $\mathcal{P}_x$ and $\mathcal{N}(0,1)$ for the case of $d = 1$ in Figure 1.

5.3. Proof of Lemma 3

We first introduce an "extended" filtration which incorporates the information from both arms up to time t. For any $t \in [T]$, let $\widetilde{\mathcal{F}}_t$ be the $\sigma-$algebra generated by the distribution of $\mathcal{S}_t := \{(x_1, y_1^{(1)}, y_2^{(2)}), (x_2, y_2^{(1)}, y_2^{(2)}), \ldots, (x_t, y_t^{(1)}, y_t^{(2)})\}$. Then we let $d_x^r(\cdot,\cdot)$ be a distance metric on the sphere $\mathbb{S}^r = \{z \in \mathbb{R}^d : \|z\|_2 = r\}$, defined as $d_x^r(z_1, z_2) := \mathbb{P}_{x\sim\mathcal{P}_x}\left[\operatorname{sgn}(\langle x, z_1\rangle) \ne \operatorname{sgn}(\langle x, z_2\rangle)\right]$. Since this distance metric is invariant to the radius r, we can prove Lemma 3 by constructing a set $\widetilde{\mathcal{P}} \subseteq \mathcal{P}_{\delta,\alpha}$ and lower bounding

$$\inf_{\widehat{\Delta}_\beta} \sup_{(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)}) \in \widetilde{\mathcal{P}}} \mathbb{E}\left[d_x^r\left((r/\|\Delta_\beta\|_2)\Delta_\beta, (r/\|\widehat{\Delta}_\beta\|_2)\widehat{\Delta}_\beta, \right)\right], \tag{12}$$

for some $r > 0$, where the infimum is taken over all $\widehat{\Delta}_\beta$ that are $\widetilde{\mathcal{F}}_T$ measurable.

Proof of (10). We aim to use the Fano's method in Proposition 15.12 in Wainwright (2019) to lower bound the minimax risk in (12). Let $\mathcal{P}_x = N(0, \mathbf{I}_d)$ and $\beta^{(2)} = 0$. Now we will construct a discrete set $\mathcal{B}$ and show that $\{\mathcal{P}_x\} \times \mathcal{B} \times \{0\} \subseteq \mathcal{P}_{\delta,1}$. By the Gilbert-Varshamov bound, there exists a set $\mathcal{B}$ such that the following holds: 1) for any $\beta \in \mathcal{B}$, $\gamma^{-1}\beta \in \{-1,1\}^d, \gamma = \delta/\sqrt{d}$; 2) for any $\beta, \beta' \in \mathcal{B}$, $\|\gamma^{-1}(\beta - \beta')\|_1 \ge d/2$; 3) the size of set $\mathcal{B}$, $|\mathcal{B}| \ge \exp(d/8)$. It is easy to show that for any $\beta^{(1)} \in \mathcal{B}$,

Assumption A4 is satisfied with δ and $\alpha = 1$ and Assumptions A1 to A3 hold with $a^2 = 1$, $\underline{\lambda} = 1/4$, $b_0 = 1$, so $\{\mathcal{P}_x\} \times \mathcal{B} \times \{0\} \subseteq \mathcal{P}_{\delta,1}$.

Now we lower bound (12) with $\widetilde{\mathcal{P}} = \{\mathcal{P}_x\} \times \mathcal{B} \times \{0\}$ and $r = \delta$. For brevity, let $M = |\mathcal{B}|$ and index elements in $\mathcal{B}$ as $\beta^1, \beta^2, \ldots, \beta^M$. We first show that the distance between any $\beta^k, \beta^{k'} \in \mathcal{B}$ is sufficiently large. Note that for any $k \in [M]$, $\|\beta^k\|_2 = \gamma\sqrt{d} = \delta$, and thus $\mathcal{B}$ is a subset of points on a sphere in $\mathbb{R}^d$ with δ radius. Then by Lemma 9 we have for any $k, k' \in [M], k \neq k'$,

$$d_x^\delta\left(\beta^k, \beta^{k'}\right) = \mathbb{P}\left[\operatorname{sgn}\left(\langle x, \gamma^{-1}\beta^k\rangle\right) \neq \operatorname{sgn}\left(\langle x, \gamma^{-1}\beta^{k'}\rangle\right)\right] \geq \mathbb{P}\left[|\zeta| \leq \sqrt{\|\gamma^{-1}(\beta^k - \beta^{k'})\|_1/(2d)}\right],$$

where ζ follows a standard Cauchy distribution. By the construction of the set $\mathcal{B}$, there exists a global constant $\mathfrak{c}_1' > 0$ such that

$$d_x^\delta\left(\beta^k, \beta^{k'}\right) \geq \mathbb{P}\left[|\zeta| \leq \sqrt{\|\gamma^{-1}(\beta^k - \beta^{k'})\|_1/(2d)}\right] \geq \mathbb{P}\left[|\zeta| \leq 1/2\right] \geq 2\mathfrak{c}_1', \forall k, k' \in [M], k \neq k'. \tag{13}$$

This yields that $\mathcal{B}$ is a $2\mathfrak{c}_1'$-seperated set under the distance metric $d_x^\delta(\cdot,\cdot)$ by the definition of seperated set in Proposition 15.12 in Wainwright (2019). Then Proposition 15.12 in Wainwright (2019) and the upper bound of mutual information in Eq. (27) in Scarlett and Cevher (2019) yields

$$\inf_{\widehat{\Delta}_\beta} \sup_{\beta^{(1)} \in \mathcal{B}} \mathbb{E}\left[d_x^\delta(\beta^{(1)}, (\delta/\|\widehat{\Delta}_\beta\|_2)\widehat{\Delta}_\beta)\right] \geq \mathfrak{c}_1'\left(1 - \left(\max_{k,k' \in [M]} \mathrm{KL}(\mathcal{P}^k || \mathcal{P}^{k'}) + \log 2\right)/\log M\right), \tag{14}$$

where the infimum is taken over all $\widetilde{\mathcal{F}}_T$ measurable estimator $\widehat{\Delta}_\beta \neq 0$, $M = |\mathcal{B}|$, and $\mathcal{P}^k$ is the distribution of $\mathcal{S}_T$ induced by $\beta^{(1)} = \beta^k, \beta^{(2)} = 0$ as well as our choice of $\mathcal{P}_x$ for any $k \in [M]$. By the construction of set $\mathcal{B}$ we can derive

$$\mathrm{KL}(\mathcal{P}^k || \mathcal{P}^{k'}) = T\mathbb{E}\left[\langle x, \beta^k - \beta^{k'}\rangle^2\right]/2 \leq T\left\|\beta^k - \beta^{k'}\right\|_2^2/2 \leq 2T\delta^2, \forall k, k' \in [M].$$

Plugging this along with $M \geq \exp(d/8)$ into (14) gives us $\inf_{\widehat{\Delta}_\beta} \sup_{\beta^{(1)} \in \mathcal{B}} \mathbb{E}\left[d_x^\delta(\beta^{(1)}, (\delta/\|\widehat{\Delta}_\beta\|_2)\widehat{\Delta}_\beta)\right] \geq 0.08\mathfrak{c}_1'$ for any $\widetilde{\mathcal{F}}_T$-measurable estimator $\widehat{\Delta}_\beta$. Since any T-admissible estimator with respect to $\mathcal{F}_T$ is also $\widetilde{\mathcal{F}}_T$-measurable, (10) holds with $\mathfrak{c}_1 = 0.08\mathfrak{c}_1'$ by our choice of $\widetilde{\mathcal{P}} = \{\mathcal{P}_x\} \times \mathcal{B} \times \{0\}$.

Proof of (11). Fix an $\alpha \in (0,1)$. For $x \sim \mathcal{P}_x$, let $x = wz$, where w is independent of z and $w \sim \mathrm{Beta}(\alpha, 1)$. Define z as follows: $z = bz_0 + (1-b)z_1, z_0 \sim \mathcal{N}(0, \mathbf{I}_d), z_1 \sim \mathcal{N}\left(0, \mathrm{diag}(0, \mathbf{I}_{d-1})\right)$, where b is a Bernoulli random variable whose mean will be defined later, and b, z_0, z_1 are independent of each other. Let $\Gamma_\alpha = \Gamma\left((1-\alpha)/2\right)$. Again by the Gilbert-Varshamov bound, there exists a set $\widetilde{\mathcal{B}}$ where the following holds: 1) for any $\beta \in \widetilde{\mathcal{B}}$, $\beta(1) = (2\Gamma_\alpha)^{1/\alpha}\delta$, $(1/\widetilde{\gamma})\beta(2:d) \in \{-1,1\}^{d-1}$ where $\widetilde{\gamma} = (256T)^{-1/2}$; 2) for any $\beta, \beta' \in \widetilde{\mathcal{B}}$, we have $(1/\widetilde{\gamma})\|\beta(2:d) - \beta'(2:d)\|_1 \geq (d-1)/2; 3)|\widetilde{\mathcal{B}}| = \widetilde{M} \geq \exp\left((d-1)/8\right)$.

Now set $\mathbb{E}[b] = 1 - 2^{-1}\delta^{-\alpha}\Gamma_\alpha^{-1}\left[\widetilde{\gamma}^2(d-1)\right]^{\alpha/2}$. Notice that $\mathbb{E}[b] \in [1/2, 1)$ since $\delta^{-\alpha}\Gamma_\alpha^{-1}\left[\widetilde{\gamma}^2(d-1)\right]^{\alpha/2} \leq \delta^{-\alpha}\Gamma_\alpha^{-1}\Gamma_\alpha\delta^\alpha = 1$, where the inequality is given by our condition $T \geq 256^{-1}(d-1)\delta^{-2}\Gamma_\alpha^{-2/\alpha}$. Thus the definition of the bernoulli variable b in $\mathcal{P}_x$ is valid. Also note

that $\|\beta\|_2 = \sqrt{4^{1/\alpha}+1}\Gamma_\alpha{}^{1/\alpha}\delta := \widetilde{r}, \forall \beta \in \widetilde{\mathcal{B}}$. Then we have the following two lemmas, which show that $\{\mathcal{P}_x\} \times \widetilde{\mathcal{B}} \times \{0\}$ is a subset of $\mathcal{P}_{\delta,\alpha}$ and the distance between any $\beta, \beta' \in \widetilde{\mathcal{B}}$ is large enough, respectively. These lemmas are proved in the appendix.

Lemma 4 *Let $\beta^{(2)} = 0$. For any $\beta^{(1)} \in \widetilde{\mathcal{B}}$, Assumption A4 is satisfied with δ and α and Assumptions A1 to A3 are satisfied with $a^2 = 1$, $\underline{\lambda} = 1/4$, $b_0 = \widetilde{r}$.*

Lemma 5 *There exists a positive constant $\mathfrak{c}_1'(\alpha)$ which only depends on α such that*

$$d_x^{\widetilde{r}}(\beta, \beta') \geq 2\mathfrak{c}_1'(\alpha) d^{\alpha/2} T^{-\alpha/2} \delta^{-\alpha}, \forall \beta, \beta' \in \widetilde{\mathcal{B}}, \beta \neq \beta'. \tag{15}$$

Lemma 4 shows that we can then lower bound (12) by choosing $\widetilde{\mathcal{P}}$ as $\{\mathcal{P}_x\} \times \widetilde{\mathcal{B}} \times \{0\}$. Lemma 5 futher yields that $\widetilde{\mathcal{B}}$ is a $2\mathfrak{c}_1'(\alpha)\left(\frac{d}{T}\right)^{\alpha/2}\delta^{-\alpha}$-seperated set under the distance metric $d_x^{\widetilde{r}}(\cdot,\cdot)$. Then similar to the proof of (10), we can use Proposition 15.12 in Wainwright (2019) and Eq. (27) in Scarlett and Cevher (2019) to show that

$$\inf_{\widehat{\Delta}_\beta} \sup_{\beta^{(1)} \in \widetilde{\mathcal{B}}} \mathbb{E}\left[d_x^{\widetilde{r}}(\beta^{(1)}, (\widetilde{r}/\|\widehat{\Delta}_\beta\|_2)\widehat{\Delta}_\beta)\right] \geq 0.08\mathfrak{c}_1'(\alpha)\left(\frac{d}{T}\right)^{\alpha/2}\delta^{-\alpha}$$

by $\mathrm{KL}(\mathcal{P}^k||\mathcal{P}^{k'}) \leq 4T\widetilde{\gamma}^2(d-1) = (d-1)/64, \forall k, k' \in [\widetilde{M}]$ and $\widetilde{M} \geq \exp((d-1)/8)$, where the infimum is taken over all $\widetilde{\mathcal{F}}_T$ measurable estimator $\widehat{\Delta}_\beta \neq 0$. Since any T-admissible estimator with respect to $\mathcal{F}_T$ is also $\widetilde{\mathcal{F}}_T$-measurable, this completes the proof of (12) by our choice of $\mathcal{P} = \{\mathcal{P}_x\} \times \widetilde{\mathcal{B}} \times \{0\}$. □

6. Conclusion

In this paper we study linear contextual bandits under a generalized margin condition. Instead of relying on standard explore-then-commit algorithms or greedy algorithms, we propose successive elimination-based methods that achieve optimal margin parameters and contextual dimensionality dependency without knowing margin condition parameters in advance. It is interesting to see whether or how our algorithm and analysis could be further extended to more complex models, such as generalized linear models, high-dimensional sparse models, or problem settings with missing data or errors in variables.

Appendix A: Proof of Lemma 1.

For any $t \in [T]$ and any $i \in \{1,2\}$ let $\mathcal{B}_t^{(i)} = \{\beta : \|\beta - \widehat{\beta}_{t-1}^{(i)}\| \le \rho_{t-1}^{(i)}\omega_0/(a\sqrt{2\log(2T^4)})\}$. We then define an event $\widetilde{\mathcal{G}}_t \supseteq \mathcal{G}_t$: For any $t \in [T]$, let $\widetilde{\mathcal{G}}_t := \left\{\left|\langle x_t, \widehat{\beta}_{t-1}^{(i)} - \beta^{(1)}\rangle\right| \le \rho_{t-1}^{(i)}\omega_0 \text{ for } i = 1,2\right\}$. We first show that event $\widetilde{\mathcal{G}}_t$ holds with high probability for any $t > t_0$. By tail probability properties of sub-Gaussian random vectors (e.g., Jin et al. 2019, Lemma 2), the following claim holds.

Claim 1 *For any $t \in [T]$, $\mathbb{P}(\widetilde{\mathcal{G}}_t) \ge 1 - 2T^{-4} - \mathbb{P}\left(\beta^{(i)} \notin \mathcal{B}_t^{(i)} \text{ for } i = 1 \text{ or } i = 2\right)$.*

The following lemma then bounds the probability of the event in the claim above.

Lemma 6 *Under the conditions in Theorem 1, $\mathbb{P}\left(\beta^{(i)} \notin \mathcal{B}_t^{(i)} \text{ for } i = 1 \text{ or } i = 2\right) \le 4T^{-3}$, $\forall t \in [T]$.*

The lemma below further bounds the width of confidence intervals.

Lemma 7 *For any $t > t_0$, $\mathbb{P}\left(\rho_{t-1}^{(i)} \le \sqrt{8\underline{\lambda} + 16a^2}/(\underline{\lambda}\sqrt{t-1}), i = 1,2\right) \ge 1 - 4T^{-3} - \mathbb{P}\left(\cup_{\tau=1}^{t}\widetilde{\mathcal{G}}_\tau^c\right)$.*

Lemma 1 is then a direct consequence of Claim 1 and Lemmas 6, 7. □

Proof of Lemma 6. Without loss of generality we focus only on the event about $\beta^{(1)}$. Let $\mathbf{X} = (x_1\mathbb{1}\{a_1 = 1\}, x_2\mathbb{1}\{a_2 = 1\}\ldots, x_{t-1}\mathbb{1}\{a_{t-1} = 1\})$ be the design matrix for arm 1 and $Y = (y_1\mathbb{1}\{a_1 = 1\}, y_2\mathbb{1}\{a_2 = 1\}\}, \ldots, y_{t-1}\mathbb{1}\{a_{t-1} = 1\})$ be the response vector for arm 1 at the beginning of time period t. We then have $Y = \mathbf{X}\beta^{(1)} + \varepsilon$, where $\varepsilon = (\varepsilon_1\mathbb{1}\{a_1 = 1\}, \varepsilon_2\mathbb{1}\{a_1 = 1\}, \ldots, \varepsilon_{t-1}\mathbb{1}\{a_{t-1} = 1\})$ is the noise vector. When $\lambda_{\min}(\mathbf{X}^\top\mathbf{X}) = 0$, our argument naturally holds since $\rho_t = +\infty$. Otherwise, standard linear algebra of ordinary least squares yields $\widehat{\beta}_{t-1}^{(1)} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta^{(1)} + \varepsilon) = \beta^{(1)} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\epsilon$. Subsequently, noting that $\mathbf{X}^\top\mathbf{X} = \mathbf{V}_{t-1}^{(1)}$, we have

$$\|\widehat{\beta}_{t-1}^{(1)} - \beta^{(1)}\|_2 = \|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon\|_2 \le \lambda_{\max}((\mathbf{X}^\top\mathbf{X})^{-1})\|\mathbf{X}^\top\varepsilon\|_2 = \|\mathbf{X}^\top\varepsilon\|_2/\lambda_{\min}(\mathbf{V}_{t-1}^{(1)}). \tag{16}$$

We next proceed to upper bound $\|\mathbf{X}^\top\varepsilon\|_2$ with high probability. Let $\mathcal{N}_\epsilon$ be an ϵ-covering of the unit ball in $\mathbb{R}^d$ with respect to $\|\cdot\|_2$, so that $\sup_{u\in\mathbb{R}^d, \|u\|_2=1}\min_{z\in\mathcal{N}_\epsilon}\|u - z\|_2 \le \epsilon$. Setting $\epsilon = 1/4$, it is easy to verify that there exists an $\mathcal{N}_{1/4}$ with $|\mathcal{N}_{1/4}| \le 9^d$. Subsequently,

$$\left\|\mathbf{X}^\top\varepsilon\right\|_2^2 = \sup_{u\in\mathbb{R}^d:\|u\|_2=1}|\langle\mathbf{X}^\top\varepsilon, u\rangle|^2 \le 2\max_{u\in\mathcal{N}_{1/4}}|\langle\mathbf{X}^\top\varepsilon, u\rangle|^2. \tag{17}$$

Additionally, for any $u \in \mathcal{N}_{1/4}$, applying the Azuma-Hoeffding's inequality we obtain with probability $1 - 2/(9^dT^3)$ that

$$\left|\langle\mathbf{X}^\top\varepsilon, u\rangle\right| \le \left|\sum_{\tau=1}^{t-1}\varepsilon_\tau\mathbb{1}\{a_\tau = 1\}\langle x_\tau, u\rangle\right| \le \sigma\sqrt{6(d + \log T)\sum_{\tau=1}^{t-1}\mathbb{1}\{a_\tau = 1\}(\langle x_\tau, u\rangle)^2}. \tag{18}$$

Note also that $\sum_{\tau=1}^{t-1}\mathbb{1}\{a_\tau = 1\}(\langle x_\tau, u\rangle)^2 \le \lambda_{\max}(\sum_{\tau=1}^{t-1}\mathbb{1}\{a_\tau = 1\}x_\tau x_\tau^\top) = \lambda_{\max}(\mathbf{V}_{t-1}^{(1)})$, where the first inequality holds because $\|u\|_2 = 1$. Consequently, incorporating the definitions of $\rho_{t-1}^{(1)}$, ω_0 and applying the union bound over all $u \in \mathcal{N}_{1/4}$, Eqs. (16,17,18) yield with probability $1 - 2T^{-3}$ that $\|\widehat{\beta}_{t-1}^{(1)} - \beta^{(1)}\|_2 \le \rho_{t-1}^{(1)}\sigma\sqrt{6(d + \log T)}$, which is to be proved. □

Proof of Lemma 7. We consider arm 1 and upper bound $\lambda_{\max}(\mathbf{V}_{t-1}^{(1)})$ and lower bound $\lambda_{\min}(\mathbf{V}_{t-1}^{(1)})$ for any $t > t_0, t \in [T]$. Let $b_1, b_2, \ldots, b_T$ be a sequence of independently and identically distributed Bernoulli random variables with mean $1/2$, such that the algorithm takes action $a_t = 1$ if $b_t = 1$ and $a_t = 2$ otherwise when randomly choosing actions. Fix arbitraty $t > t_0$ and we first consider the lower bound of $\lambda_{\min}(\mathbf{V}_{t-1}^{(1)})$.

For brevity, $\forall \tau \in [t-1]$, let $g_\tau^1 := \mathbb{1}\{L_\tau^{(1)}(x_\tau) > U_\tau^{(2)}(x_\tau)\}$ and $g_\tau^2 := \mathbb{1}\{L_\tau^{(1)}(x_\tau) \le U_\tau^{(2)}(x_\tau), L_\tau^{(2)}(x_\tau) \le U_\tau^{(1)}(x_\tau)\}$. Subsequently, $\lambda_{\min}(\sum_{\tau=1}^{t-1} x_\tau x_\tau^\top \mathbb{1}\{a_\tau = 1\}) \ge \lambda_{\min}(\sum_{\tau=1}^{t-1} x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\}(g_\tau^1 + g_\tau^2)) := h_t^0$ because $\mathbb{1}\{a_\tau = 1\} = g_\tau^1 + g_\tau^2 b_\tau \ge b_\tau(g_\tau^1 + g_\tau^2)$ and $b_\tau \le 1$. To further lower bound h_t^0, note that for any $\tau \in [t-1]$, if $\widetilde{\mathcal{G}}_\tau$ holds, then $a_\tau^* = 1$ and $L_\tau^{(2)}(x_\tau) > U_\tau^{(1)}(x_\tau)$ cannot happen simultaneously, because $\widetilde{\mathcal{G}}_\tau$ with $a_\tau^* = 1$ implies $L_\tau^{(2)}(x_\tau) \le \langle x_\tau, \beta^{(2)} \rangle \le \langle x_\tau, \beta^{(1)} \rangle \le U_\tau^{(1)}(x_\tau)$, and therefore $\mathbb{1}\{a_\tau^* = 1\}\mathbb{1}\{\widetilde{\mathcal{G}}_\tau\}\mathbb{1}\{L_\tau^{(2)}(x_\tau) > U_\tau^{(1)}(x_\tau)\} = 0$. Consequently, if $\widetilde{\mathcal{G}}_\tau$ holds for $\tau \le t-1$, then

$$\begin{aligned} h_t^0 &= \lambda_{\min}\left(\sum\nolimits_{\tau=1}^{t-1} x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\}\mathbb{1}\{\widetilde{\mathcal{G}}_\tau\}\left(g_\tau^1 + g_\tau^2 + \mathbb{1}\{L_\tau^{(2)}(x_\tau) > U_\tau^{(1)}(x_\tau)\}\right)\right) \\ &\ge \lambda_{\min}\left(\sum\nolimits_{\tau=1}^{t-1} x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\}\right), \end{aligned} \tag{19}$$

where the second inequality holds from the observations that $g_\tau^1 + g_\tau^2 + \mathbb{1}\{L_\tau^{(2)}(x_\tau) > U_\tau^{(1)}(x_\tau)\} = 1$. To lower bound the first term in (19), first note by concavity of the minimum eigenvalues of PSD matrices that

$$\lambda_{\min}\left(\sum_{\tau=1}^{t-1} \mathbb{E}[x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\}]\right) \ge \sum_{\tau=1}^{t-1} \mathbb{E}[b_\tau] \lambda_{\min}\left(\mathbb{E}\left[x_\tau x_\tau^\top \mathbb{1}\{a_\tau^* = 1\}\right]\right) \ge \frac{(t-1)\underline{\lambda}}{2}. \tag{20}$$

Next, to upper bound the difference between the minimum eigenvalues of $\sum_{\tau=1}^{t-1} x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\}$ and its expectation, we apply standard matrix concentration inequalities (Vershynin 2018, Theorem 4.7.1, Exercise 4.7.3) to obtain that, with probability greater than $1 - T^{-3}$

$$\lambda_{\max}\left(\sum\nolimits_{\tau \in [t-1]} \left(x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\} - \mathbb{E}[x_\tau x_\tau^\top b_\tau \mathbb{1}\{a_\tau^* = 1\}]\right)\right) \le (t-1)\underline{\lambda}/4 \tag{21}$$

provided that $t > t_0 \ge \lceil (d + 3\log T) \max\{256a^4/\underline{\lambda}^2, 16a^2/\underline{\lambda}\} \rceil$. If $\widetilde{\mathcal{G}}_\tau$ holds for $\tau \le t-1$, then combining Eqs. (19,20,21) gives

$$\lambda_{\min}(\mathbf{V}_{t-1}^{(1)}) = \lambda_{\min}\left(\sum_{\tau=1}^{t-1} x_\tau x_\tau^\top \mathbb{1}\{a_\tau = 1\}\right) \ge \underline{\lambda}(t-1)/4. \tag{22}$$

By matrix concentration inequality and $\lambda_{\max}(\mathbb{E}[x_\tau x_\tau^\top]) \le a^2$ as x_τ is a^2 subgaussian, we can show that

$$\lambda_{\max}(\mathbf{V}_{t-1}^{(1)} + \mathbf{V}_{t-1}^{(2)}) \le \underline{\lambda}(t-1)/2 + a^2(t-1). \tag{23}$$

with probability higher than $1 - 2T^{-3}$. Since $\lambda_{\max}(\mathbf{V}_{t-1}^{(1)}) \le \lambda_{\max}(\mathbf{V}_{t-1}^{(1)} + \mathbf{V}_{t-1}^{(2)})$, by plugging (22) and (23) into the definition $\rho_{t-1}^{(1)} = \sqrt{\lambda_{\max}(\mathbf{V}_{t-1}^{(1)})}/\lambda_{\min}(\mathbf{V}_{t-1}^{(1)})$ in Algorithm 1 we complete our proof. □

Appendix B: Proofs Omitted in Section 5

Proof of Lemma 2. For any $t \in [T]$, recall that a_t^π is the arm π selects at time step t for any policy π. Now let $r_t^\pi = |\Delta(x_t)| \mathbb{1}\{a_t^\pi \neq a_t^*\}$ denote the instant regret of π at time step t and let Δ_t^π to be a π-induced estimator at time step t defined by

$$\Delta_t^\pi \in \arg\max_\Delta \left(\mathbb{P}[a_t^\pi = 1, \langle x_t, \Delta\rangle \geq 0] + \mathbb{P}[a_t^\pi = 2, \langle x_t, \Delta\rangle < 0]\right). \tag{24}$$

By Lemma 3.1 in He et al. (2022) and the margin condition, we have for any policy π, $r_t^\pi \geq \frac{h}{2}\mathbb{P}\left[\operatorname{sgn}\left(\Delta(x_t)\right) \neq \operatorname{sgn}(\langle x_t, \Delta_t^\pi\rangle)\right] - h^{\alpha+1}\delta^{-\alpha}, \forall h > 0$. Therefore we have $\forall h > 0$,

$$\mathfrak{R}_{\delta,\alpha}(T) = \inf_\pi \sup \mathbb{E}\left[\sum_{t=1}^T r_t^\pi\right] \geq \frac{h}{2} \inf_\pi \sup \mathbb{E}\left[\sum_{t=1}^T \left(\mathbb{P}\left[\operatorname{sgn}\left(\Delta(x_t)\right) \neq \operatorname{sgn}(\langle x_t, \Delta_t^\pi\rangle)\right]\right)\right] - T\frac{h^{\alpha+1}}{\delta^\alpha}, \tag{25}$$

where the supremums is taker over all $(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)}) \in \mathcal{P}(\delta, \alpha)$. Note that for all admissible policies π and $\forall t \in [T]$, the π-induced estimator at time step t, Δ_t^π, is also admissible. This yields

$$\begin{aligned}
&\inf_\pi \sup \mathbb{E}\left[\sum_{t=1}^T \left(\mathbb{P}\left[\operatorname{sgn}\left(\Delta(x_t)\right) \neq \operatorname{sgn}(\langle x_t, \Delta_t^\pi\rangle)\right]\right)\right] \text{ (the infinum is taken over all admissible policy } \pi) \\
&\geq \inf_{\widetilde{\Delta}_t, t\in[T]} \sup \sum_{t=1}^T \mathbb{E}\left[\left(\mathbb{P}\left[\operatorname{sgn}\left(\Delta(x_t)\right) \neq \operatorname{sgn}(\langle x_t, \widetilde{\Delta}_t\rangle)\right]\right)\right] \text{ (the infinum is taken over all } \widetilde{\mathcal{F}}_T\text{-measurable } \widetilde{\Delta}_t) \\
&= T \inf_{\widehat{\Delta}_\beta} \sup_{(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)}) \in \mathcal{P}(\delta,\alpha)} \mathbb{E}\left[\left(\mathbb{P}\left[\operatorname{sgn}\left(\Delta(x_t)\right) \neq \operatorname{sgn}(\langle x_t, \widehat{\Delta}_\beta\rangle)\right]\right)\right] = \mathfrak{E}_{\delta,\alpha}(T)T,
\end{aligned}$$

where all supremums is taker over all $(\mathcal{P}_x, \beta^{(1)}, \beta^{(2)}) \in \mathcal{P}(\delta, \alpha)$, the thrid infimum is taken over all T-admissible estimators $\widehat{\Delta}_\beta$, and the second last equality holds because $x_1, \ldots, x_T$ are identically distributed. Plugging the above into (25) and taking $h = \delta\left[\mathfrak{E}_{\delta,\alpha}(T)\right]^{1/\alpha} 4^{-1/\alpha}$ yields the final result. □

Proof of Lemma 4. We only verify the margin condition since the choice of other parameters is trivial. By our condition, $\beta^{(2)} = 0$. We have for any $\beta^{(1)} \in \widetilde{\mathcal{B}}$,

$$\mathbb{P}\left[|\Delta(x)| \leq h\right] = \mathbb{P}\left[\left|\langle x, \beta^{(1)}\rangle\right| \leq h\right] = \mathbb{P}\left[\left|\langle wz_0, \beta^{(1)}\rangle\right| \leq h\right]\mathbb{E}[b] + \mathbb{P}\left[\left|\langle wz_1, \beta^{(1)}\rangle\right| \leq h\right]\mathbb{E}[1-b]. \tag{26}$$

For the first term on the right-hand side, first notice that $\langle z_0, \beta^{(1)}\rangle$ follows a normal distribution with variance $\beta^{(1)}(1)^2 + \widetilde{\gamma}^2(d-1) \geq \beta^{(1)}(1)^2$. Thus we have

$$\begin{aligned}
\mathbb{P}\left[\left|\langle wz_0, \beta^{(1)}\rangle\right| \leq h\right] &= \mathbb{P}\left[\left|w\left(\beta^{(1)}(1)^2 + \widetilde{\gamma}^2(d-1)\right)^{-1/2}\langle z_0, \beta^{(1)}\rangle\right| \leq h\left(\beta^{(1)}(1)^2 + \widetilde{\gamma}^2(d-1)\right)^{-1/2}\right] \\
&\leq \left(h/\beta^{(1)}(1)\right)^\alpha \Gamma_\alpha = h^\alpha \left[2\Gamma_\alpha\right]^{-1}\delta^{-\alpha}\Gamma_\alpha = (h/\delta)^\alpha/2, \forall h > 0,
\end{aligned}$$

where the inequality is given by Lemma 8. For the second term on the right-hand side of (26), similarly we have $\langle z_1, \beta^{(1)}\rangle$ follows a Gaussian distribution with variance $\widetilde{\gamma}^2(d-1)$, and thus

$$\mathbb{P}\left[\left|\langle wz_1, \beta^{(1)}\rangle\right| \leq h\right]\mathbb{E}[1-b] \leq \Gamma_\alpha h^\alpha \left(\widetilde{\gamma}^2(d-1)\right)^{-\alpha/2} \cdot \frac{1}{2}\delta^{-\alpha}\Gamma_\alpha^{-1}\left[\widetilde{\gamma}^2(d-1)\right]^{\alpha/2} = (h/\delta)^\alpha/2, \forall h > 0.$$

By combining the above with (26), we have for any $\beta^{(1)} \in \widetilde{\mathcal{B}}$,

$$\mathbb{P}\left[|\Delta(x)| \le h\right] = \mathbb{P}\left[\left|\langle wz_0, \beta^{(1)}\rangle\right| \le h\right]\mathbb{E}[b] + \mathbb{P}\left[\left|\langle wz_1, \beta^{(1)}\rangle\right| \le h\right]\mathbb{E}[1-b] \le (h/\delta)^{\alpha}, \forall h > 0,$$

which means the margin condition holds for all $\beta^{(1)} \in \mathcal{B}$. □

Proof of Lemma 5. For any $\beta, \beta' \in \widetilde{\mathcal{B}}$, by the independence between z_1 and b, we have

$$\begin{aligned}
d_x^{\widetilde{r}}(\beta, \beta') &\ge \mathbb{P}\left[\operatorname{sgn}\left(\langle bz_0 + (1-b)z_1, \beta\rangle\right) \ne \operatorname{sgn}\left(\langle bz_0 + (1-b)z_1, \beta'\rangle\right) \text{ and } b = 0\right] \\
&= \mathbb{P}\left[\operatorname{sgn}\left(\langle z_1, \beta\rangle\right) \ne \operatorname{sgn}\left(\langle z_1, \beta'\rangle\right)\right]\mathbb{P}\left[b = 0\right] \\
&= \mathbb{P}\left[\operatorname{sgn}\left(\left\langle z_1(2:d), \frac{1}{\widetilde{\gamma}}\beta(2:d)\right\rangle\right) \ne \operatorname{sgn}\left(\left\langle z_1(2:d), \frac{1}{\widetilde{\gamma}}\beta'(2:d)\right\rangle\right)\right]\mathbb{E}[1-b] \\
&\ge \mathbb{P}\left[|\zeta| \le \sqrt{\frac{\|\beta(2:d) - \beta'(2:d)\|_1}{2\widetilde{\gamma}(d-1)}}\right]\mathbb{E}[1-b] \ge \mathbb{P}\left[|\zeta| \le \frac{1}{2}\right]\mathbb{E}[1-b] \ge 0.29 \cdot \frac{\delta^{-\alpha}}{2}\Gamma_{\alpha}^{-1}\left[\widetilde{\gamma}^2(d-1)\right]^{\alpha/2},
\end{aligned}$$

where ζ obeys a standard Cauchy distribution and the second inequality is given by Lemma 9. By $\widetilde{\gamma}^2 = (256T)^{-1}$, this yields that (15) holds for $\mathfrak{c}_1'(\alpha) = 0.29 \times 4^{-\alpha-1}\Gamma_{\alpha}^{-1}$. □

Appendix C: Additional Technical Lemmas

Lemma 8 *Let w, z be independent random variables with $w \sim \mathsf{Beta}(\alpha, 1)$, $z \sim N(0,1)$ for some $\alpha \in (0,1)$. Then for any $h > 0$, $\mathbb{P}[|wz| \le h] \le h^{\alpha}\Gamma((1-\alpha)/2)$.*

Proof. Using basic probability calculus we have the following derivation, which proves the lemma.

$$\mathbb{P}\left[|wz| \le h\right] = 2\int_0^{\infty} (2\pi)^{-1/2}\exp(-x^2/2)\mathbb{P}\left[|wx| \le h\right]dx \le h^{\alpha}\int_0^{\infty}\exp(-x^2/2)x^{-\alpha}dx \le h^{\alpha}\Gamma\left(\frac{1-\alpha}{2}\right).$$

Lemma 9 *Let x, ζ be independent random variables such that ζ follows a standard Cauchy distribution and $x \sim N(0, \mathbf{I}_d)$. For any $s, s' \in \{-1, 1\}^d$, $\mathbb{P}[\operatorname{sgn}(\langle x, s\rangle) \ne \operatorname{sgn}(\langle x, s'\rangle)] \ge \mathbb{P}[|\zeta| \le \sqrt{\|s - s'\|_1/(2d)}]$.*

Proof. The probability of $\langle x, s\rangle, \langle x, s'\rangle$ having different signs can be written as

$$\mathbb{P}\left[\frac{\left|\langle x, (s+s')/2\rangle\right|}{\left|\langle x, (s'-s)/2\rangle\right|} \le 1\right] = \mathbb{P}\left[\frac{\left|\langle x, s+s'\rangle\right|/\|s+s'\|_2}{\left|\langle x, s'-s\rangle\right|/\|s-s'\|_2} \le \frac{\|s-s'\|_2}{\|s+s'\|_2}\right] = \mathbb{P}\left[|\zeta| \le \frac{\|s-s'\|_2}{\|s+s'\|_2}\right],$$

where the second equality is given by that $\langle x, s+s'\rangle$ is independent with $\langle x, s-s'\rangle$ and the distribution of the fraction of two independent stadard normal random variables is standard Cauchy distribution. Finally, note that because $s, s' \in \{-1, 1\}^d$, $\|s+s'\|_2 \le 2\sqrt{d}$ and $\|s-s'\|_1 = 2\sqrt{\|s-s'\|_1/2}$ because each coordinate of $s - s'$ is either 0 or ± 2. This yields $\mathbb{P}[|\zeta| \le \|s-s'\|_2/\|s+s'\|_2] \ge \mathbb{P}[|\zeta| \le \sqrt{\|s-s'\|_1/(2d)}]$, which completes the proof of Lemma 9. □

References

Abbasi-Yadkori, Yasin, Dávid Pál, Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

Abbasi-Yadkori, Yasin, David Pal, Csaba Szepesvari. 2012. Online-to-confidence-set conversions and application to sparse stochastic bandits. *Artificial Intelligence and Statistics*. PMLR.

Agarwal, Alekh, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, Robert Schapire. 2014. Taming the monster: A fast and simple algorithm for contextual bandits. *International conference on machine learning*. PMLR, 1638–1646.

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

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

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

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

Besbes, Omar, Will Ma, Omar Mouchtaki. 2023. From contextual data to newsvendor decisions: On the actual performance of data-driven algorithms. *arXiv preprint arXiv:2302.08424* .

Bubeck, Sébastien, Nicolo Cesa-Bianchi, Sham M Kakade. 2012. Towards minimax policies for online linear optimization with bandit feedback. *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 41–1.

Chen, Yi, Yining Wang, Ethan X Fang, Zhaoran Wang, Runze Li. 2024. Nearly dimension-independent sparse linear bandit over small action spaces via best subset selection. *Journal of the American Statistical Association* **119**(545) 246–258.

Cohen, Maxime C, Ilan Lobel, Renato Paes Leme. 2020. Feature-based dynamic pricing. *Management Science* **66**(11) 4921–4943.

Ding, Jingying, Woonghee Tim Huh, Ying Rong. 2024. Feature-based inventory control with censored demand. *Manufacturing & Service Operations Management* **26**(3) 1157–1172.

Duan, Congyuan, Wanteng Ma, Jiashuo Jiang, Dong Xia. 2024. Regret minimization and statistical inference in online decision making with high-dimensional covariates. *arXiv preprint arXiv:2411.06329* .

Foster, Dylan, Alexander Rakhlin. 2020. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. *International conference on machine learning*. PMLR, 3199–3210.

Goldenshluger, Alexander, Assaf Zeevi. 2009. Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability* **19**(4) 1603–1633.

Goldenshluger, Alexander, Assaf Zeevi. 2013. A linear response bandit problem. *Stochastic Systems* **3**(1) 230–261. doi:10.1287/11-SSY032.

Gur, Yonatan, Ahmadreza Momeni, Stefan Wager. 2022. Smoothness-adaptive contextual bandits. *Operations Research* **70**(6) 3198–3216.

He, Jiahao, Jiheng Zhang, Rachel Q Zhang. 2022. A reduction from linear contextual bandits lower bounds to estimations lower bounds. *International Conference on Machine Learning*. PMLR, 8660–8677.

Hu, Yichun, Nathan Kallus, Xiaojie Mao. 2022. Smooth contextual bandits: Bridging the parametric and nondifferentiable regret regimes. *Operations Research* **70**(6) 3261–3281. doi:10.1287/opre.2021.2237.

Javanmard, Adel, Hamid Nazerzadeh. 2019. Dynamic pricing in high-dimensions. *Journal of Machine Learning Research* **20**(9) 1–49.

Jin, Chi, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, Michael I. Jordan. 2019. A short note on concentration inequalities for random vectors with subgaussian norm. URL `https://arxiv.org/abs/1902.03736`.

Kallus, Nathan, Madeleine Udell. 2020. Dynamic assortment personalization in high dimensions. *Operations Research* **68**(4) 1020–1037.

Kim, Seok-Jin, Min-hwan Oh. 2024. Local anti-concentration class: Logarithmic regret for greedy linear contextual bandit. *Advances in Neural Information Processing Systems* **37** 77525–77592.

Lee, Harin, Taehyun Hwang, Min hwan Oh. 2025. Lasso bandit with compatibility condition on optimal arm. *The Thirteenth International Conference on Learning Representations*. URL `https://openreview.net/forum?id=f3jySJpEFT`.

Oh, Min-hwan, Garud Iyengar, Assaf Zeevi. 2021. Sparsity-agnostic lasso bandit. *International Conference on Machine Learning*. PMLR, 8271–8280.

Perchet, Vianney, Philippe Rigollet. 2013. The multi-armed bandit problem with covariates. *The Annals of Statistics* **41**(2) 693–721.

Ren, Zhimei, Zhengyuan Zhou. 2024. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science* **70**(2) 1315–1342.

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

Scarlett, Jonathan, Volkan Cevher. 2019. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555* .

Simchi-Levi, David, Yunzong Xu. 2022. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research* **47**(3) 1904–1931.

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

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge university press.

Wang, Xue, Mike Mingcheng Wei, Tao Yao. 2024. Online learning and decision making under generalized linear model with high-dimensional data. *Management Science* .

Xu, Kan, Hamsa Bastani. 2025. Multitask learning and bandits via robust statistics. *Management Science* .