Experimental Design for Semiparametric Bandits

Seok-Jin Kim SEOK-JIN.KIM@COLUMBIA.EDU
Columbia University

Gi-Soo Kim GISOOKIM@UNIST.AC.KR
UNIST

Min-hwan Oh MINOH@SNU.AC.KR
Seoul National University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study finite-armed semiparametric bandits, where each arm's reward combines a linear component with an unknown, potentially adversarial shift. This model strictly generalizes classical linear bandits and reflects complexities common in practice. We propose the first experimental-design approach that simultaneously offers a sharp regret bound, a PAC bound, and a best-arm identification guarantee. Our method attains the minimax regret $\widetilde{\mathcal{O}}(\sqrt{dT})$, matching the known lower bound for finite-armed linear bandits, and further achieves logarithmic regret under a positive suboptimality gap condition. These guarantees follow from our refined non-asymptotic analysis of orthogonalized regression that attains the optimal $\sqrt{d}$ rate, paving the way for robust and efficient learning across a broad class of semiparametric bandit problems.

Keywords: semiparametric bandits, experimental design, exploration-exploitation, G-optimal design

1. Introduction

Linear bandits, in which each arm is identified with a feature vector $x \in \mathbb{R}^d$ and the expected reward takes the form $x^\top \theta^\star$, have been extensively studied for both regret minimization and pure-exploration objectives (Abbasi-Yadkori et al., 2011; Soare et al., 2014; Jedra and Proutiere, 2020; Degenne et al., 2020; Xu et al., 2018; Fiez et al., 2019). Thanks to the linear structure, design-based methods such as the G-optimal design provide strong theoretical guarantees, enabling optimal $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$[1] regret bounds (Lattimore and Szepesvári, 2019) and tight sample complexities for *probably approximately correct* (PAC) learning (Li et al., 2022; Chaudhuri and Kalyanakrishnan, 2019; Sakhi et al., 2023) and *best arm identification* (BAI) (Soare et al., 2014; Fiez et al., 2019; Jedra and Proutiere, 2020; Komiyama et al., 2022).

However, many practical systems face adversarial or unpredictable effects that cannot be faithfully captured by a purely linear model. Baseline shifts or adversarial interventions often introduce reward variations beyond the reach of linear assumptions. To address these complexities, a semiparametric reward model (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim and Paik, 2019) incorporates an additional shift term. Specifically, if r_t denotes the reward obtained from taking action a_t whose feature vector is $x_{a_t} \in \mathbb{R}^d$ at time t, the model posits that

$$r_t = x_{a_t}^\top \theta^\star + \nu_t + \eta_t,$$

Corresponding authors: Gi-Soo Kim (gisookim@unist.ac.kr) and Min-hwan Oh (minoh@snu.ac.kr).

[1] $\widetilde{\mathcal{O}}$ suppresses logarithmic factors of d and T but *does not hide dependence on the number of arms, K.*

where $\theta^\star$ is an unknown parameter, ν_t is an arbitrary (potentially adversarial) shift that is not governed by a parametric form and is determined prior to action selection, and η_t is noise.

This formulation subsumes linear bandits as a special case (by taking $\nu_t = 0$) and captures extra complexities found in real-world applications, such as recommender systems (Mladenov et al., 2020). However, many real-world bandit problems involve fixed features with time-varying effects—for example, clinical trials with fixed treatments but varying subjects (but without subject information) (Kazerouni and Wein, 2021). A similar scenario appears in ad selection for a landing page (without access to personal information). The term ν_t accounts for such time-varying baselines, including changes in varying user propensity, external interference, or broader trends. Moreover, BAI is commonly studied under fixed features, where incorporating ν_t is still meaningful.

Existing analyses of semiparametric bandits currently provide only $\widetilde{\mathcal{O}}(d\sqrt{T})$ bounds (Greenewald et al., 2017; Krishnamurthy et al., 2018; Chowdhury et al., 2023) or $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ bounds (Kim and Paik, 2019). However, both rates fall short of the $\widetilde{\mathcal{O}}(\sqrt{dT\log K})$ bound known to be optimal in linear bandits with a finite action set (Li et al., 2019). This gap leaves open the question of whether more sophisticated design-based methods can attain a sharper $\sqrt{d}$-type regret bound in semiparametric settings, even under adversarial shifts. Moreover, no gap-dependent logarithmic regret result has been established, further contrasting with the well-studied linear bandit framework.

Another key limitation in the existing literature is the complete absence of work on experimental design in semiparametric bandits. Whereas G-optimal design is routinely employed in linear bandits to obtain PAC and BAI guarantees, no analogous design-based technique exists for semiparametric bandits. All prior approaches (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim and Paik, 2019; Chowdhury et al., 2023) have utilized *orthogonalized regression*, which introduces substantial new challenges to experimental design and prevents the direct extension of existing linear-bandit methods. Hence, the following research questions arise:

- *Can we develop an experimental design for semiparametric bandits?*
- *Can we develop an efficient algorithm that leverages this experimental design to simultaneously achieve sharp regret bounds, PAC, and BAI guarantees?*

In this paper, we address these questions by proposing new algorithmic and analysis frameworks for semiparametric bandits that incorporate optimal design methods into orthogonalized regression. We propose the first procedure that tackles the non-convex design problem of orthogonalized regression efficiently, enabling a sharp regret bound of $\widetilde{\mathcal{O}}(\sqrt{dT\log K})$. Moreover, by utilizing a dependency on the suboptimality gap, our approach also achieves logarithmic regret, bridging the gap between the theory of linear bandits and that of semiparametric bandits. Beyond regret minimization, we provide the first PAC and BAI results for semiparametric bandits by establishing sample complexities that match those of Soare et al. (2014) for linear bandits. Finally, we propose our main algorithm, which exhibits low regret as well as PAC and BAI guarantees. These new results are driven by a sharper non-asymptotic analysis of orthogonalized regression, which replaces the loose Cauchy–Schwarz arguments and delivers dimension-optimal statistical rates.

Our result can also be applied to a multi-armed bandit (MAB) problem with a semiparametric reward model. In this setting, each arm i has a base reward distribution with a mean of $\mu_i \subset \mathbb{R}$ for $i = 1, 2, \ldots, K$, but the reward observed at each time step t includes an additional, time-varying shift, ν_t. This setup can be framed within our model by defining the feature set $\mathcal{X}$ as the standard basis of $\mathbb{R}^K$. We discuss this application in greater detail in Section 5.4. For this bandit problem, our results achieve a regret bound of $\widetilde{O}(\sqrt{KT})$ as well as logarithmic regret when there is a suboptimality gap.

1.1. Our Contributions

We summarize our key contributions as follows:

1. **Experimental design for semiparametric bandits.** Prior to our work, experimental design for semiparametric reward models had not been investigated. We develop a novel design-based approach that, for the first time, enables both sharp regret bounds and exploration-based guarantees (PAC, BAI) in semiparametric bandits. In particular, our proposed algorithm, built around this new experimental design, attains a regret bound of $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$ while simultaneously ensuring PAC and BAI performance.

2. $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$ **regret bound for semiparametric bandits.** We propose a phase-elimination-based algorithm that achieves $\min(\widetilde{\mathcal{O}}(\sqrt{dT \log K}), \widetilde{\mathcal{O}}(d\sqrt{T}))$ cumulative regret (Theorem 6 and Theorem 18). To our best knowledge, this is the first result attaining the $\sqrt{d}$-rate regret in a semiparametric bandit setting. It matches the known lower bounds for linear bandits with finite arms (Lattimore and Szepesvári, 2019), and since the linear model is a special case of our framework, this bound is information-theoretically minimax optimal (up to logarithmic factors).

3. **Gap-dependent logarithmic regret bound for semiparametric bandits.** By allowing for suboptimality gap dependence, we also derive the regret bound logarithmic in T (Theorem 7), marking the first logarithmic regret result for semiparametric bandits. Hence, when the suboptimality gap is reasonably small, our result significantly improves over previous results (Krishnamurthy et al., 2018).

4. **PAC and BAI results for semiparametric bandits.** We provide the first exploration-based guarantees for semiparametric bandits (Corollary 5 and Theorem 8) including the PAC bound and BAI guarantee. It is important to note that our proposed algorithm simultaneously possesses the properties of low regret, PAC, and BAI. Our sample complexities match the results for linear bandits studied in Soare et al. (2014).

5. **Novel policy design for orthogonalized regression.** Orthogonalized regression is a key technique for semiparametric reward estimation (Krishnamurthy et al., 2018; Kim and Paik, 2019; Choi et al., 2023), requiring a G-optimal design step that is typically non-convex and challenging to solve. We propose an efficient algorithmic procedure (Theorem 3) that yields a suitable design solution while remaining computationally tractable, thereby enabling effective policy construction in practice.

6. **Sharper non-asymptotic analysis of orthogonalized regression.** Existing semiparametric bandit approaches often rely on the analysis based on the Cauchy–Schwarz inequality, which do not yield the optimal $\sqrt{d}$ dependence in estimation error. We develop a new analysis framework, achieving dimension-optimal statistical rates for orthogonalized regression (Theorem 4) and thereby improving upon the standard estimation error bounds, which can be of independent interest.

1.2. Related Work

Semiparametric Reward Model in Bandits. Several prior studies have explored the semiparametric reward model in various settings, including contextual and multi-agent bandit (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim and Paik, 2019; Chowdhury et al., 2023; Choi et al., 2023). Krishnamurthy et al. (2018) introduced an algorithm, achieving $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret, while Kim and Paik (2019) proposed a Thompson sampling-based method with regret $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$. Chowdhury et al. (2023) then developed a more computationally efficient algorithm with $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret, and Choi et al. (2023) studied multi-agent bandits under the same semiparametric framework. However, Krishnamurthy et al. (2018) left open the possibility of achieving $\widetilde{\mathcal{O}}(\sqrt{dT\log K})$ regret. All of the prior analyses rely on the Cauchy-Schwarz inequality-based approach, which cannot deliver the desired $\sqrt{d}$-rate in regret. Moreover, there has been the absence of exploration-based guarantees (such as PAC or BAI) in the semiparametric bandit literature.

Pure Exploration in Linear Bandits. PAC algorithms in bandits (Sakhi et al., 2023; Chaudhuri and Kalyanakrishnan, 2019; Wagenmaker and Jamieson, 2022) and RL (Strehl et al., 2009) are a key objective, alongside regret minimization, and have long been a prominent line of research. Particularly in linear bandits, the PAC and Best Arm Identification (BAI) objectives have been thoroughly investigated in works such as Soare et al. (2014); Degenne et al. (2020); Li et al. (2022); Jedra and Proutiere (2020); Fiez et al. (2019); Komiyama et al. (2022); Yang and Tan (2022), where G-optimal design plays a key role in achieving near-optimal sample complexity. Unlike regret minimization, these methods aim to identify the best action or to bound the performance relative to the optimum, leading to rich theoretical and practical implications. However, in the semiparametric reward model, orthogonalized regression is used in place of standard linear regression, and these existing design-based strategies for linear bandits do not extend to the semiparametric setting. As a result, there has been no prior work offering PAC or BAI guarantees for semiparametric bandits.

G-optimal Design for Linear Bandits. Let $\mathcal{X} = \{x_1, \ldots, x_K\} \subset \mathbb{R}^d$ be a set of features for actions in $\{1, ..., K\}$. The G-optimal design is defined as the solution to the following optimization problem (Lattimore and Szepesvári, 2019; Soare et al., 2014):

$$v^\star := \min_{(p_1,\ldots,p_K)\in\Delta^{(K)}} \max_{i\in[K]} \|x_i\|_{\left(\sum_i p_i x_i x_i^\top\right)^{-1}}, \tag{1}$$

and a known result gives $v^\star = \sqrt{d}$. The problem can be solved efficiently, and the support of the solution has bounded cardinality: $\left|\{i \mid p_i > 0\}\right| \leq \frac{d(d+1)}{2}$. The resulting policy minimizes the maximum prediction variance over $\mathcal{X}$, thereby efficiently solving the PAC and best arm identification problems for linear bandits. In semiparametric bandits, however, existing methods rely on *orthogonalized regression* (e.g., Krishnamurthy et al., 2018), leading to a different, non-convex design objective. We formulate a suitable G-optimal design for orthogonalized regression and propose an efficient algorithm to solve it, thus extending the benefits of experimental designs to semiparametric bandits.

1.3. Notations

We define $[n] := \{1, 2, \ldots, n\}$ for a positive integer n. We write $\mathcal{O}(\cdot)$ or $\lesssim$ to hide absolute constants, and $\widetilde{\mathcal{O}}(\cdot)$ to hide constants and logarithmic factors in d and T. We do *not* hide the factor $\log K$. We use $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$. We define $\Delta^{(n)}$ as the n-dimensional simplex. For a vector $x \in \mathbb{R}^d$, we let $\|x\|_p$ be the ℓ_p norm. For any positive semidefinite matrix $\mathbf{A}$, we define $\|x\|_{\mathbf{A}} = \sqrt{x^\top \mathbf{A} x}$.

Matrix-inverse-weighted Norm. In what follows, we generalize the definition of the matrix-inverse-weighted norm (e.g., $\|x\|_{\mathbf{A}^{-1}}$) to positive semidefinite matrices, regardless of invertibility. To define a matrix-inverse-weighted norm for non-invertible matrices, we introduce the following extended definition. Let $\mathbf{A}$ be a positive semidefinite matrix. We define

$$\|x\|_{\mathbf{A}^{-1}} := \lim_{\lambda \to 0} \|x\|_{(\mathbf{A}+\lambda \mathbf{I}_d)^{-1}}.$$

This reduces to the usual definition of $\|x\|_{\mathbf{A}^{-1}}$ when $\mathbf{A}$ is full rank. By allowing $\lambda \to 0$, we extend the definition to handle the non-invertible case in a well-defined way. Even though $\mathbf{A}$ is non-invertible, if x is contained in the subspace spanned by $\mathbf{A}$'s eigenvectors, then this value is finite and well-defined.

2. Preliminaries

In this section, we first present our problem setup of semiparametric bandits. Unlike the linear model, our framework includes an additional shift ν_t, determined prior to action selection, which may be adversarial. After that, we introduce orthogonalized regression—a common approach for obtaining the estimator used in all prior works (Krishnamurthy et al., 2018; Kim and Paik, 2019; Choi et al., 2023).

2.1. Problem Setup: Semiparametric Bandits

We consider a finite-armed bandit problem in which each arm $i \in [K]$ is represented by a feature vector $x_i \in \mathbb{R}^d$. Collectively, these vectors form the set $\mathcal{X} = \{x_1, \ldots, x_K\} \subset \mathbb{R}^d$. At each round $t = 1, 2, \ldots, T$, the learner selects an arm $a_t \in [K]$ and observes a reward

$$r_t \;=\; x_{a_t}^\top \theta^\star \;+\; \nu_t \;+\; \eta_t.$$

Here, $\theta^\star \in \mathbb{R}^d$ is an unknown parameter vector, $\nu_t \in \mathbb{R}$ is a bounded shift (which can be adversarially chosen), and η_t is a noise term. Specifically, we assume that η_t is independent and sub-Gaussian with variance proxy 1 (Wainwright, 2019). We set $\mathcal{H}_t$ as the sigma-algebra generated by $\{a_1, r_1, \ldots, a_t, r_t\}$. We allow the shift ν_t to be any $\mathcal{H}_{t-1}$-measurable random variable satisfying the boundedness condition in Assumption 1. This means ν_t may be adversarial, as long as it is determined before the choice of arm at time t.

Feature Span. Without loss of generality, we assume that $\{x_1, \ldots, x_K\}$ spans a d-dimensional subspace of $\mathbb{R}^d$. Should the actual rank be $d' < d$, we may re-parameterize the problem using a $d' \times d$ matrix A and a $d' \times 1$ vector $\theta^{\star\prime}$ so that for all $i \in [K]$, $x_i^\top \theta^\star = (Ax_i)^\top \theta^{\star\prime}$. Henceforth, we assume full column rank (i.e. rank = d) without loss of generality.

Optimal Arm and Suboptimality Gap. We define the optimal arm as $a^\star := \arg\max_{i \in [K]} x_i^\top \theta^\star$, and let $\Delta_\star := x_{a^\star}^\top \theta^\star - \max_{j \neq a^\star} x_j^\top \theta^\star$ denote the suboptimality gap. Throughout, we assume the optimal arm is unique, though our methods readily extend to settings with multiple optimal arms.

Cumulative Regret. Our primary performance criterion for the semiparametric bandit problem is the *cumulative regret*, defined by

$$\mathbf{Reg}(T) \;:=\; \sum_{t=1}^{T} \left(x_{a^\star}^\top \theta^\star \;-\; x_{a_t}^\top \theta^\star \right).$$

Since $a^\star$ maximizes the expected reward in the absence of the shift ν_t, the regret measures how much reward is lost by playing suboptimal arms over time.

Fixed Policies and Experimental Design. In addition to regret minimization, we consider exploration problems and, more generally, estimation tasks that require carefully chosen sampling policies. We use the term *design* to denote a (possibly randomized) policy that chooses arm $a_t \in [K]$ at each time t. Sections 3 and 4 focus on evaluating the estimation error when a fixed policy (design) is executed, i.e., when the policy is determined independently of the data. Our aim there is to identify designs that control the prediction variance and enable strong PAC or best arm identification guarantees. Section 5 then develops an adaptive, data-dependent policy that simultaneously achieves low cumulative regret, PAC, and BAI guarantees.

Assumption 1 (Boundedness of parameter and features) *We assume $\|\theta^\star\|_2 \le 1$, $\|x_i\|_2 \le 1$ for all $i \in [K]$, and $|\nu_t| \le 1$ for all $t \ge 1$.*

Using 1 as the bound is without loss of generality and maintains consistency with prior work (Krishnamurthy et al., 2018; Kim and Paik, 2019; Choi et al., 2023; Chowdhury et al., 2023). No additional assumption is placed on ν_t beyond boundedness; in particular, it may be adversarially chosen, state-dependent, or stochastic, provided it is fixed before each action is selected.

Next, we introduce the orthogonalized regression method for estimating $\theta^\star$, which is used in all the aforementioned prior works.

2.2. Estimation of $\theta^\star$ by Orthogonalized Regression

For action a_s sampled at time s according to a random policy, we define the *centered feature* $\widetilde{x}_{a_s} := x_{a_s} - \mathbb{E}[x_{a_s} \mid \mathcal{H}_{s-1}]$ so that $\mathbb{E}[\widetilde{x}_{a_s} \mid \mathcal{H}_{s-1}] = 0$. This expectation is taken with respect to the distribution of a_s, namely, $\mathbf{p}_s = \{p_{i,s}\}_{i=1}^K$. Since a_s is sampled from a multinomial distribution $\mathrm{Multinomial}(1, \mathbf{p}_s)$, we have $\mathbb{E}[x_{a_s}|\mathcal{H}_{s-1}] = \sum_{i=1}^K p_{i,s} x_i$. At time t, we define the *centered Gram matrix* $\widehat{\mathbf{V}}_t := \sum_{s=1}^t \widetilde{x}_{a_s}\widetilde{x}_{a_s}^\top$ and set the *empirical covariance* $\widehat{\boldsymbol{\Sigma}}_t := \frac{1}{t}\widehat{\mathbf{V}}_t$. We adopt the estimator proposed in Krishnamurthy et al. (2018), obtained by regressing the rewards on the centered features via ridge regression as follows,

$$\hat{\theta}_t = (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1} \sum_{s=1}^t \widetilde{x}_{a_s} r_s, \tag{2}$$

where β_t denotes the ridge regularizer. Roughly speaking, we perform ridge regression on the independent variables $\{\widetilde{x}_{a_s}\}_{s=1}^t$ and the responses $\{r_s\}_{s=1}^t$.

Consistency of the estimator can be inferred from the following decomposition of the estimation error:

$$\hat{\theta}_t - \theta^\star = (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1} \sum_{s=1}^t \widetilde{x}_{a_s}\left(\mathbb{E}[x_{a_s} \mid \mathcal{H}_{s-1}]^\top \theta^\star + \nu_s + \eta_s\right) - \beta_t(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\theta^\star.$$

We observe that $\sum_{s=1}^t \widetilde{x}_{a_s}\left(\mathbb{E}[x_{a_s} \mid \mathcal{H}_{s-1}]^\top \theta^\star + \nu_s + \eta_s\right)$ is a martingale adapted to the filtration $\{\mathcal{H}_{s-1}\}_{s=1}^t$. Therefore, the first term converges to 0 in probability under a suitably well-conditioned centered Gram matrix and covariance, $\widehat{\mathbf{V}}_t$ and $\widehat{\boldsymbol{\Sigma}}_t$. Compared to the first term, the second term is negligible and converges to 0 even faster. Later, we present a sharp analysis of the estimation error in Section 4, which is one of our main contributions.

3. Experimental Design for Orthogonalized Regression

In this section, we discuss efficient experimental design for orthogonalized regression. In our work, "design" refers to the policy $\mathbf{p} = (p_1, \ldots p_K) \in \Delta^{(K)}$, where p_i denotes the probability of selecting action a_i. Specifically, we consider a scenario in which a fixed policy $\mathbf{p}$ is employed for pure exploration over a predetermined period. Our goal is to identify an effective design for estimation. First, we establish the necessary notations and setup, and then we formulate the optimization problem for the design.

Throughout Sections 3 and 4, we consider the case where we pull arms up to time t according to a fixed policy $\mathbf{p} = (p_1, \ldots p_K)$ and obtain samples $\{x_{a_s}, r_s\}_{s=1}^t$. We then have $\mathbb{E}[x_{a_s} \mid \mathcal{H}_{s-1}] = \bar{x}_{\mathbf{p}}$ and $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_t] = \mathbb{E}[\widetilde{x}_{a_s}\widetilde{x}_{a_s}^\top] := \boldsymbol{\Sigma}_{\mathbf{p}}$, where $\bar{x}_{\mathbf{p}}$ and $\boldsymbol{\Sigma}_{\mathbf{p}}$ are defined as follows.

Definition 1 (Covariance and mean of policy $\mathbf{p}$) *We define the feature mean of a policy $\mathbf{p}$ as*

$$\bar{x}_{\mathbf{p}} := \sum_{i=1}^K p_i x_i.$$

We also define the covariance of a policy $\mathbf{p}$ as

$$\boldsymbol{\Sigma}_{\mathbf{p}} = \sum_{i=1}^K p_i \big(x_i - \bar{x}_{\mathbf{p}}\big)\big(x_i - \bar{x}_{\mathbf{p}}\big)^\top.$$

Then we have $\widetilde{x}_{a_s} := x_{a_s} - \bar{x}_{\mathbf{p}}$ and additionally define $\mathbf{V}_t := \mathbb{E}[\widehat{\mathbf{V}}_t] = t\boldsymbol{\Sigma}_{\mathbf{p}}$.

Ridge Regularizer Selection. In Lemma 24, we prove that when we choose $\beta_t = \log(t/\delta)$, we have

$$\frac{1}{c}\big(\mathbf{V}_t + \beta_t \mathbf{I}_d\big) \preceq \widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d \preceq c\big(\mathbf{V}_t + \beta_t \mathbf{I}_d\big)$$

with probability at least $1 - \frac{\delta}{10}$ for some absolute constant $c > 0$. From now on, we use the ridge regularizer $\beta_t = \log(t/\delta)$ when we perform regression with t samples. We also define the *normalized ridge regularizer* $\lambda_t = \frac{\beta_t}{t}$. Later, we prove in Appendix H that for our design (presented in Section 3), the constant c can be $c \in [1, 2]$ once $t \gtrsim d\log(dt/\delta)$.

3.1. Key Quantity to Bound Estimation Error

Consider the situation where we aim to estimate the value of $z^\top\theta^\star$ at some $z \in \mathbb{R}^d$. We observe below that $\|z\|_{\boldsymbol{\Sigma}_{\mathbf{p}}^{-1}}$ is the key quantity that must be controlled for minimizing the estimation error at $z \in \mathbb{R}^d$. For the extended definition of matrix-inverse-weighted norm (e.g. $\|\cdot\|_{\boldsymbol{\Sigma}_{\mathbf{p}}^{-1}}$), see Section 1.3.

To motivate this, we begin by presenting an error analysis based on the Cauchy–Schwarz inequality, as used in all prior works. Using Lemma 11 of Krishnamurthy et al. (2018) and Lemma 20 (modified version), we can derive, for any $z \in \mathbb{R}^d$:

$$|z^\top(\hat{\theta}_t - \theta^\star)| \le \|z\|_{(\mathbf{V}_t+\beta_t\mathbf{I}_d)^{-1}}\|(\hat{\theta}_t - \theta^\star)\|_{\mathbf{V}_t+\beta_t\mathbf{I}_d} \lesssim \frac{1}{\sqrt{t}}\sqrt{d\log\big(\frac{t}{\delta}\big)}\|z\|_{\boldsymbol{\Sigma}_{\mathbf{p}}^{-1}}, \tag{3}$$

where the first inequality holds via Cauchy–Schwarz. It is evident from this inequality that a low value of $\|z\|_{\boldsymbol{\Sigma}_{\mathbf{p}}^{-1}}$ guarantees a tight bound on the estimation error. This result is analogous to linear bandits, except that the covariance of the policy replaces the standard second moment.

Later in Theorem 4, we replace the Cauchy–Schwarz-based error analysis with a sharper inequality that removes the $\sqrt{d}$ factor from the bound above. (The suboptimality of the Cauchy–Schwarz-based analysis is discussed in Appendix G.) The resulting bound is still proportional to $\|z\|_{\mathbf{\Sigma}_{\mathbf{p}}^{-1}}$.

To identify the best arm or minimize the regret, it suffices to estimate the expected reward of each arm up to an additive constant, i.e., it suffices to estimate $x_i^\top \theta^\star + c$ for some constant c for all $i \in [K]$. Letting $c = -\mu^\top \theta^\star$ for some vector $\mu \in \mathbb{R}^d$, we have $x_i^\top \theta^\star + c = (x_i - \mu)^\top \theta^\star$. We therefore aim to find a design that minimizes the maximum value of $\|z\|_{\mathbf{\Sigma}_{\mathbf{p}}^{-1}}$ over $z \in \mathcal{X} - \mu := \{x_1 - \mu, \cdots, x_K - \mu\}$, where the value of $\mu \in \mathbb{R}^d$ is specified in the next section.

3.2. Main Challenges in Constructing the Experimental Design

We now formulate our optimization problem, aiming to find the optimal design for orthogonalized linear regression.

Parallel Shifting of Features. A natural choice of μ is $\mu = 0$. However, we find that it can sometimes be impossible to finitely bound $\max_{i \in [K]} \|x_i\|_{\mathbf{\Sigma}_{\mathbf{p}}^{-1}}$. This is because the subspace of $\mathbf{\Sigma}_{\mathbf{p}}$ is spanned by $\{x_i - x_1\}_{i \in [K]}$ (by Lemma 9), which can be a strict subspace of $\mathrm{span}(x_1, \ldots, x_K)$. This problem can be circumvented by shifting the contexts using an appropriate nonzero vector μ. Based on this observation, we present our G-optimal design problem for orthogonalized regression.

Definition 2 (G-optimal design of orthogonalized regression) *We aim to find a policy* $\mathbf{p}$ *that achieves*

$$\min_{\mu \in \mathbb{R}^d, \, \mathbf{p} \in \Delta^{(K)}} \max_{i \in [K]} \|x_i - \mu\|_{\mathbf{\Sigma}_{\mathbf{p}}^{-1}},$$

where

$$\mathbf{\Sigma}_{\mathbf{p}} = \sum_{i=1}^{K} p_i (x_i - \bar{x}_{\mathbf{p}})(x_i - \bar{x}_{\mathbf{p}})^\top \quad \text{and} \quad \bar{x}_{\mathbf{p}} = \sum_{i=1}^{K} p_i x_i.$$

Challenges and Differences from Linear Bandits. Our optimization problem includes the covariance term $\sum_{i=1}^K p_i (x_i - \bar{x}_{\mathbf{p}})(x_i - \bar{x}_{\mathbf{p}})^\top$, which is a **third-order** polynomial of p_i and hence is **non-convex**. In contrast, the optimal design for linear bandits is formulated with the second moment $\sum_{i=1}^K p_i x_i x_i^\top$, which is linear in p_i. The non-convexity renders our problem much more challenging, calling for non-convex optimization methods.

Our Goal. For linear bandits, the G-optimal design Eq. (1) always has optimal cost $v^\star = \sqrt{d}$. Since our reward model includes the linear model as a special case ($\nu_t = 0$), if we solve our optimization problem and obtain $\mathcal{O}(\sqrt{d})$, we can view it as a sufficiently "nice" solution. We propose a design that achieves $\mathcal{O}(\sqrt{d})$, leading to an optimal regret bound and tight PAC and BAI results.

3.3. Proposed Experimental Design and Performance Guarantees

Due to non-convexity, the optimization problem in Definition 2 is hard to solve exactly. In this Section, instead of solving the problem exactly, we propose an algorithm which achieves $\mathcal{O}(\sqrt{d})$, which is sufficiently good for the estimation, which we will discuss more in next Section 4.

There are two main challenges. The first is handling the covariance of the policy, which involves third-order terms of $\mathbf{p}$. The second is choosing μ. We need to pick an appropriate μ that effectively minimizes Problem 2.

Our algorithm is surprisingly simple, yet it handles the non-convex optimization problem efficiently. We define $b_i = x_i - x_1$ for $i = 2, \dots, K$. First, we find the G-optimal design over $b_2, \dots, b_K$ for standard linear regression (as defined in Eq. (1)) and denote it by $(\widetilde{p}_2, \dots, \widetilde{p}_K)$. Then we return our final policy $\mathbf{p} = \left(\frac{1}{2}, \frac{\widetilde{p}_2}{2}, \dots, \frac{\widetilde{p}_K}{2}\right)$. Our optimal design algorithm, named `DEO`, is presented in Algorithm 1.

Algorithm 1 `DEO`: **D**esign of **E**xperiment for **O**rthogonalized Regression

Require: Feature set $\mathcal{X}$ and $\delta > 0$.

Calculate $b_2 = x_2 - x_1, \dots, b_K = x_K - x_1$.

Find the G-optimal design Eq. (1) for $b_2, \dots, b_K$, and set the obtained policy as $\widetilde{p}_2, \dots, \widetilde{p}_K$.

Set $p_1 = \frac{1}{2}$ and $p_i = \frac{\widetilde{p}_i}{2}$ for $i \geq 2$.

return $\mathbf{p}^{\text{deo}} = (p_1, \dots, p_K)$.

We denote the design of `DEO` as $\mathbf{p}^{\text{deo}} = (p_1^{\text{deo}}, \dots, p_K^{\text{deo}})$ and redefine the covariance of $\mathbf{p}^{\text{deo}}$ as $\mathbf{\Sigma}_{\text{deo}}$. We also redefine the feature mean of $\mathbf{p}^{\text{deo}}$ as $\bar{x}_{\mathbf{p}^{\text{deo}}} := \bar{x}_{\text{deo}}$.

Theorem 3 (Performance of Algorithm 1) *Our policy obtained by Algorithm 1 satisfies*

$$\|x_i - x_1\|_{\mathbf{\Sigma}_{\text{deo}}^{-1}} \leq 2\sqrt{d},$$

for all $i \in [K]$. Also, for all $i \in [K]$,

$$\|x_i - \bar{x}_{\text{deo}}\|_{\mathbf{\Sigma}_{\text{deo}}^{-1}} \leq 4\sqrt{d}.$$

Additionally, the support of the policy satisfies $\left|\{i \in [K] \colon p_i^{\text{deo}} > 0\}\right| \leq \frac{d(d+1)}{2}$.

Discussion of Theorem 3. This theorem shows that our policy obtained by Algorithm 1 effectively solves the main optimization problem of Problem 2 and achieves a result up to a constant factor of the golden value, $\sqrt{d}$. By simply selecting $\mu = x_1$ in Problem 2, it achieves a 2-approximation with respect to our benchmark quantity $\sqrt{d}$. Implementing this design, we can efficiently estimate $\left(x_i - x_1\right)^\top \theta^\star$ for all $i \in [K]$, which is sufficient for obtaining optimal regret bound. Moreover, even if we choose $\mu = \bar{x}_{\text{deo}}$, we still attain performance on the order of $\mathcal{O}(\sqrt{d})$. Its proof is deferred to Appendix A.

4. Analysis of Estimation Error for Orthogonalized Regression

Previously, we obtained a policy design that successfully bound $\max_{x\in\mathcal{X}} \|x - \mu\|_{\mathbf{\Sigma}_{\mathbf{p}}^{-1}} \lesssim \sqrt{d}$ for $\mu = x_1$. The next goal is to develop a sharp estimation error bound for orthogonalized regression under the fixed policy. All of the previous studies Krishnamurthy et al. (2018); Kim and Paik (2019); Choi et al. (2023) obtained error bounds using a Cauchy–Schwarz-based analysis. However from Eq. (3), we see that this approach leads to a estimation error of order $\mathcal{O}(\frac{d}{\sqrt{t}})$ which is suboptimal in d. We discuss this further in Appendix G. Even in linear bandits, sharp results in dimension d are never obtained via a Cauchy–Schwarz-based analysis (Lattimore and Szepesvári, 2019). In this Section, we provide a novel non-asymptotic estimation error analysis for orthogonalized regression, which yields an error bound of order $\mathcal{O}(\frac{\sqrt{d}}{\sqrt{t}})$. This rate is optimal for linear reward models. Since the linear model is a special case of our setup, the rate cannot be further improved.

4.1. Novel and Sharp Estimation Error Analysis

We present our novel error analysis, which does not use a Cauchy–Schwarz-based approach. Because of orthogonalized regression, we devise entirely new techniques for bounding the estimation error.

Theorem 4 (Estimation error upper bound) *Suppose we obtain t samples $\{x_{a_s}, r_s\}_{s=1}^{t}$ from pure exploration with a fixed policy $\mathbf{p}$. For some $z \in \mathbb{R}^d$, set $\|z\|^2_{\mathbf{\Sigma}_\mathbf{p}^{-1}} = L$ and $\max_{i\in[K]} \|x_i - \bar{x}_\mathbf{p}\|^2_{\mathbf{\Sigma}_\mathbf{p}^{-1}} = M$. Then the estimator $\hat{\theta}_t$ obtained by orthogonalized regression with ridge regularizer $\beta_t = \log(t/\delta)$ satisfies*

$$|z^\top(\hat{\theta}_t - \theta^\star)| \le C_1\Big(\frac{\sqrt{L\log\big(\frac{t}{\delta}\big)}}{\sqrt{t}} + \frac{\sqrt{L}M\log\big(\frac{d}{\delta}\big)}{t}\Big)$$

with probability at least $1-\delta$ for a universal constant $C_1 > 0$. Also, the leading term $\frac{\sqrt{L}}{\sqrt{t}}$ matches the lower bound up to some constant and cannot be improved.

Discussion of Theorem 4. This theorem provides a novel and sharp non-asymptotic error bound for orthogonalized regression with an improved rate. When we use our experimental design $\mathbf{p}$ from `DEO` and choose $z = x_i - x_1$, we have $L, M \lesssim d$ by Theorem 3, leading to an estimation error bound of $\widetilde{\mathcal{O}}\big(\frac{\sqrt{d\log K}}{\sqrt{t}}\big)$, which is on the $\sqrt{d}$ scale. It matches the minimax rate of linear regression with dimension d, and since our model is broader and more challenging than the standard linear model, it also matches the minimax rate in this context. To the best of our knowledge, this is the first dimension-optimal result ($\sqrt{d}$-rate upper bound) for the non-asymptotic analysis of orthogonalized regression. A Cauchy–Schwarz-based analysis, which is used in all previous literature on orthogonalized regression, cannot meet this rate, making this a significant improvement. Its proof is deferred to Appendix B.

4.2. Warm-up: Pure Exploration with `DEO`

We now state our pure exploration strategy using the policy design `DEO`. The procedure is simple:

1. Find a policy $\mathbf{p}^{\text{deo}}$ by running `DEO`.
2. Sample actions with $\mathbf{p}^{\text{deo}}$ for t rounds and then stop.
3. The output is a greedy policy $a_t = \arg\max_i x_i^\top \hat{\theta}_t$, where $\hat{\theta}_t$ is our estimator (Eq. (2)) with regularizer $\beta_t > 0$.

This is a warm-up version of pure exploration; in the next section, we propose an algorithm that achieves low regret, PAC guarantees, and also performs BAI.

For any $\varepsilon, \delta > 0$, an (ε, δ)-PAC algorithm aims to ensure that the value function is close to the optimal value. It aims to find a policy π satisfying $V^\pi \ge V^\star - \varepsilon$ with probability at least $1-\delta$, where $V^\star$ is the value of the optimal policy and V^π is the value of policy π. We define the sample complexity required to achieve this as $\boldsymbol{\tau}(\varepsilon, \delta)$, which we refer to as the (ε, δ)-PAC bound. We next present the PAC bound of our pure exploration strategy.

Corollary 5 (PAC bound) *Suppose we are conducting pure exploration with policy $\mathbf{p}^{\text{deo}}$ and $\beta_t = \log(t/\delta)$ as described above. Its (ε, δ)-PAC bound satisfies*

$$\boldsymbol{\tau}(\varepsilon, \delta) \ge C_2\Big(\frac{d\log\big(\frac{dK}{\varepsilon\delta}\big)}{\varepsilon^2} + \frac{d^{\frac{3}{2}}\log\big(\frac{dK}{\delta}\big)}{\varepsilon}\Big) = \widetilde{\mathcal{O}}\Big(\frac{d}{\varepsilon^2}\log K\Big)$$

for some absolute constant $C_2 > 0$.

Discussion of Corollary 5. This is the first result on pure exploration and PAC bound for the class of bandit problems with a semiparametric reward model. The proof is deferred to Appendix C. By using an arm elimination technique, with a simple modification, we can obtain a BAI strategy with sample complexity $\widetilde{\mathcal{O}}\big(d \log K/\Delta_\star^2\big)$. Later, in Section 5, we show that our low-regret algorithm (Algorithm 2) also achieves BAI while enjoying PAC guarantees.

5. Main Algorithm with Low Regret, PAC and BAI Properties

We now present our main algorithm. This algorithm exhibits an instance-independent regret of order $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$, which is the first optimal result in bandit problems with a semiparametric reward model. We also derive a problem-dependent regret bound of order $\widetilde{\mathcal{O}}(d \log K/\Delta_\star)$. Furthermore, our algorithm simultaneously achieves BAI with sample complexity $\widetilde{\mathcal{O}}(d \log K/\Delta_\star^2)$, and enjoys an (ϵ, δ)-PAC guarantee with sample complexity $\widetilde{\mathcal{O}}(d \log K/\varepsilon^2)$.

5.1. Proposed Algorithm

Our algorithm, SBE, is shown in Algorithm 2. SBE adopts the phase-elimination scheme and incorporates our experimental design DEO. Given a G-optimal design, Lattimore and Szepesvári (2019) studied a phase elimination scheme that achieves $\widetilde{\mathcal{O}}(\sqrt{dT})$ regret for linear reward model. By combining this scheme with the results of Theorems 3 and 4, we propose the following low-regret algorithm. At the beginning of the ℓ-th phase, we denote the set of arms that were not eliminated up to the previous phase as $\mathcal{A}_\ell$. We compute the policy $\mathbf{p}_\ell^{\mathrm{deo}}$ over $\mathcal{A}_\ell$ via Algorithm 1, wherein the role of x_1 is replaced by $\mathcal{A}_\ell(1)$, the arm in $\mathcal{A}_\ell$ with the smallest index. Then sample actions according to $\mathbf{p}_\ell^{\mathrm{deo}}$ for

$$n_\ell := 4C_2 \left[\frac{d}{\varepsilon_\ell^2} \log\Big(\frac{dK\ell(\ell+1)}{\delta \varepsilon_\ell}\Big) + \frac{d^{\frac{3}{2}}}{\varepsilon_\ell} \log\Big(\frac{dK\ell(\ell+1)}{\delta}\Big) \right] \tag{4}$$

times. At the end of the phase, we calculate $\hat{\theta}_{(\ell)}$ using orthogonalized regression with ridge regularizer $\beta_{(\ell)} = \log\big(\frac{n_\ell \ell(\ell+1)}{\delta}\big)$ on the samples obtained during phase ℓ. We then eliminate arms from $\mathcal{A}_\ell$ whose estimated rewards are less than the maximum by more than ε_ℓ. The aforementioned procedure is repeated for $\ell = 1, 2, \cdots$ until only one arm survives. If there is only one arm left, declare it as the best arm and select that arm until the end.

5.2. Regret Analysis

First, we present the regret bound of SBE without any dependency on the suboptimality gap. Our result is a $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$ cumulative regret, which is the first result for bandits with a semiparametric reward model.

Theorem 6 (Regret bound of SBE) *The SBE algorithm has the following cumulative regret bound with probability at least $1 - \delta$:*

$$\mathbf{Reg}(T) \lesssim \sqrt{dT \log\big(K/\delta\big)} + \sqrt{dT} \log T + d^{3/2} \log\left(\frac{dKT}{\delta}\right) \log\Big(\frac{T}{d}\Big) = \widetilde{\mathcal{O}}\big(\sqrt{dT \log K}\big).$$

Algorithm 2 SBE: **S**emiparametric **B**andits with **E**limination

Input: Features $\mathcal{X}$, $\delta > 0$ and $\mathbf{B}_0 = 0\mathbf{I}_d, \mathbf{b}_0 = \mathbf{0}$.
Initialize $\mathcal{A}_1 = [K]$, $t = 1$.
for $\ell = 1, 2, \ldots$ **do**
  if $|\mathcal{A}_\ell| = 1$ **then**
    declare the arm in $\mathcal{A}_\ell$ as the best arm and select that arm until the end.
  end
  else
    Set $\varepsilon_\ell = \frac{1}{2^\ell}$ and calculate a policy $\mathbf{p}_\ell^{\mathrm{deo}}$ using DEO for the remaining arms $\mathcal{A}_\ell$. The role of x_1 in DEO is replaced by $\mathcal{A}_\ell(1)$.
    Set n_ℓ from Eq. (4).
    for $j = 1, 2, \ldots, n_\ell$ **do**
      Pull an arm according to $\mathbf{p}_\ell^{\mathrm{deo}}$. Let the sampled arm be a_t, and receive reward r_t.
      Set $\widetilde{x}_{a_t} = x_{a_t} - \mathbb{E}[x_{a_t} \mid \mathcal{H}_{t-1}]$. Update $\mathbf{B}_j = \mathbf{B}_{j-1} + \widetilde{x}_{a_t}\widetilde{x}_{a_t}^\top$, $\mathbf{b}_j = \mathbf{b}_{j-1} + \widetilde{x}_{a_t} r_t$.
      Update $t \leftarrow t + 1$. If $t + 1 \geq T$, exit.
    end
    Calculate $\hat{\theta}_{(\ell)} = (\mathbf{B}_{n_\ell} + \beta_{(\ell)}\mathbf{I}_d)^{-1}\mathbf{b}_{n_\ell}$ for $\beta_{(\ell)} = \log\big(n_\ell \ell(\ell+1)/\delta\big)$.
    Eliminate arms $\big\{x \in \mathcal{A}_\ell \big| \max_{x' \in \mathcal{A}_\ell}(x')^\top \hat{\theta}_{(\ell)} - x^\top \hat{\theta}_{(\ell)} > \varepsilon_\ell\big\}$ and update the remaining arm set to $\mathcal{A}_{\ell+1}$.
    Reset $\mathbf{B}_0 = 0 \cdot \mathbf{I}_d$, $\mathbf{b}_0 = 0_d$.
  end
end

Discussion of Theorem 6. The bound matches known optimal results for linear bandits (up to logarithmic factors), even though our problem is more challenging. Compared to known results, this is the first $\sqrt{d}$-rate regret bound for bandits with a semiparametric reward model. With small modifications, one obtains $\min\big(\widetilde{\mathcal{O}}(\sqrt{dT \log K}), \widetilde{\mathcal{O}}(d\sqrt{T})\big)$ as shown in Appendix E and F. Hence, with a reasonably finite number of arms, our result provides the sharpest regret bound among bandits with a semiparametric reward model. Its proof is deferred to Appendix D.

Next, we present a gap-dependent regret bound that achieves a logarithmic scale in the time horizon T.

Theorem 7 (Gap-dependent regret bound of SBE) *The SBE algorithm has the following cumulative regret bound with probability at least $1 - \delta$:*

$$\mathbf{Reg}(T) \lesssim \Big(\frac{d}{\Delta_\star} + d^{3/2}\Big) \log\Big(\frac{dK}{\delta \Delta_\star}\Big) \log(\frac{1}{\Delta_\star}) = \widetilde{\mathcal{O}}\Big(\frac{d}{\Delta_\star} \log K\Big).$$

Discussion of Theorem 7. This result shows logarithmic regret, and we highlight that it is the first such result in bandits with a semiparametric reward model. Similarly to the high-probability regret, one can derive the expected regret of order $\mathcal{O}\big(\frac{d}{\Delta_\star} \log(\frac{TdK}{\Delta_\star})\big)$ by setting $\delta = \frac{1}{T}$. Hence, when the suboptimality gap $\Delta_\star$ is reasonably small, our result improves upon the existing algorithms Greenewald et al. (2017); Krishnamurthy et al. (2018); Kim and Paik (2019). The proof is in Appendix D.

5.3. PAC and BAI Properties

Even though our algorithm has sublinear regret, it possesses exploration-based properties such as PAC and BAI. Algorithm 2 declares the remaining action as the best arm when only one action is left. We prove that the declared arm is indeed the best arm with high probability and provide its sample complexity.

Theorem 8 (PAC and BAI properties of SBE) *Our SBE enjoys both PAC and BAI properties, as follows:*
(BAI): *Our SBE algorithm outputs the best arm with probability at least $1-\delta$ upon selecting $\boldsymbol{\tau}_{BAI}$ samples, where*

$$\boldsymbol{\tau}_{BAI} = \widetilde{\mathcal{O}}\Big(\frac{d}{\Delta_\star^2}\log K\Big).$$

(PAC): *At time t, let the policy of SBE be π_t. with probability at least $1-\delta$, we have $V^{\pi_t} \geq V^\star - \varepsilon$ whenever $t \geq \boldsymbol{\tau}(\varepsilon,\delta)$ for*

$$\boldsymbol{\tau}(\varepsilon,\delta) = \widetilde{\mathcal{O}}\Big(\frac{d}{\varepsilon^2}\log K\Big).$$

Discussion of Theorem 8. These corollaries show that SBE achieves low regret and also performs BAI while satisfying the PAC property. Previously proposed low-regret algorithms in Kim and Paik (2019); Krishnamurthy et al. (2018) do not have BAI and PAC guarantees as stated in Theorem 8. Hence, the proposed algorithm SBE is the first algorithm that comes with such guarantees while also achieving near-optimal regret, which we believe is a major contribution to the literature.

However, the BAI sample complexity presented above is not minimax optimal when the instance is fixed. In the case of linear bandits, the instance-dependent minimax sample complexity can be smaller than $\Omega(\frac{d}{\Delta_\star^2})$, as studied in many literature (Jedra and Proutiere, 2020; Fiez et al., 2019). Therefore, this result can be practically and theoretically suboptimal, and improving it is considered a promising future direction.

5.4. Application to MAB with Semiparametric Rewards

Furthermore, we would like to mention that our results are also applicable to the MAB with a semiparametric reward model. The semiparametric MAB assumes the following model: there are K arms and each arm $i \in [K]$ has reward distribution of mean $\mu_i \in \mathbb{R}$. When arm i is pulled at time t, a shifted reward $r_t = \mu_i + \varepsilon_t + \nu_t$ is received. Here, ε_t is a mean-zero sub-Gaussian noise, and ν_t is a shift determined before the arm selection. This can be incorporated into our setup by setting the features as standard basis of $\mathbb{R}^K$. In this case, $d = K$. Directly applying the results from previous work (Kim and Paik, 2019; Krishnamurthy et al., 2018) yields a regret bound of $\widetilde{\mathcal{O}}(K\sqrt{T})$, which is suboptimal. Our algorithm and results attain a regret bound of $\widetilde{\mathcal{O}}(\sqrt{KT})$ (also hiding $\log K$ terms), which matches the known lower bound. It is also important to note that an instance-dependent logarithmic regret is achieved. Our results guarantee a regret bound of $\widetilde{\mathcal{O}}(K/\Delta_\star)$, which, while not matching the known MAB lower bound, is still favorable.

6. High-level Proof Sketch

The core of our results lies in Theorem 3 and Theorem 4. Once these two theorems are established, other results can be obtained by applying standard techniques used in the linear bandit literature.

Theorem 3 is obtained through Lemma 9. Let $\widetilde{p}_2, \dots \widetilde{p}_K$ be the G-optimal design computed for the features $\{x_2 - x_1, \dots x_K - x_1\}$, and let $\mathbf{\Sigma}_{\text{opt},1}$ be the second moment of $\{x_2 - x_1, \dots x_K - x_1\}$ under that policy (G-optimal design). By Lemma 9, our second moment satisfies $\mathbf{\Sigma}_{\text{deo}} \succeq \frac{1}{4}\mathbf{\Sigma}_{\text{opt},1}$, which completes the proof. A detailed proof is provided in Appendix A.

The key to the proof of Theorem 4 is performing decorrelation. The estimation error $\hat{\theta}_t - \theta^\star$ is decomposed as follows; see Appendix B for detailed notation:

$$\hat{\theta}_t - \theta^\star = \underbrace{(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1} \sum_{s=1}^{t} \widetilde{x}_{a_s} \underbrace{\left(\bar{x}^\top \theta^\star + \nu_s\right)}_{:=q_s}}_{:=\mathscr{A}} + \underbrace{(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1} \sum_{s=1}^{t} \widetilde{x}_{a_s} \eta_s}_{:=\mathscr{B}} - \underbrace{\beta_t (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1} \theta^\star}_{:=\mathscr{C}}.$$

Among these, $\mathscr{B}$ and $\mathscr{C}$ are terms that also appear in the error decomposition of standard linear regression and are easily controlled. The problematic term is $\mathscr{A}$. The difficulty in the analysis arises because the randomness of the mean-zero vector $\widetilde{x}_{a_s}$ is correlated with $(\hat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}$. First, we define $e_s := (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-1} \widetilde{x}_{a_s}$, which is a mean-zero vector. By manipulating the expression, for any $z \in \mathbb{R}^d$, we can decompose $z^\top \mathscr{A}$ as follows:

$$z^\top \mathscr{A} = \frac{1}{t} z^\top \left(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d\right)^{-\frac{1}{2}} \sum_{s=1}^{t} e_s q_s + \frac{1}{t} z^\top \left(\widehat{\mathbf{\Sigma}}_t + \lambda_t \mathbf{I}_d\right)^{-1} \left(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_t\right) \left(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d\right)^{-\frac{1}{2}} \sum_{s=1}^{t} e_s q_s$$

The first term on the right-hand side can be analyzed because the matrix inverse term $(\left(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d\right)^{-\frac{1}{2}})$ and the martingale term $(\sum_{s=1}^{t} e_s q_s)$ are decorrelated, and the second term becomes a higher-order term of order $\widetilde{\mathcal{O}}(\frac{1}{t})$. A detailed proof can be found in Appendix B.

7. Conclusion

We introduced the first experimental design framework for semiparametric bandits, yielding both pure exploration guarantees (PAC and BAI) and tight regret bounds, including the $\widetilde{\mathcal{O}}(\sqrt{dT \log K})$ rate and its gap-dependent logarithmic counterpart. Our work opens several avenues for further study: (1) refining pure exploration algorithms to achieve optimal sample complexities (which may not simultaneously achieve optimal regret), and (2) developing methods for the fixed-budget setting, a natural counterpart to fixed-confidence pure exploration in linear bandits.

Acknowledgements

We are very grateful for the constructive comments provided by the reviewers during our COLT submission. Gi-Soo Kim was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Program (UNIST); No. 2022-0-00469, Development of Core Technologies for Task-oriented Reinforcement Learning for Commercialization of Autonomous Drones). Min-hwan Oh 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) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-02263754)

References

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

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Pac identification of many good arms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 991–1000. PMLR, 2019.

Young-Geun Choi, Gi-Soo Kim, Seunghoon Paik, and Myunghee Cho Paik. Semi-parametric contextual bandits with graph-laplacian regularization. *Information Sciences*, 645:119367, 2023.

Mubarrat Chowdhury, Elkhan Ismayilzada, Khalequzzaman Sayem, and Gi-Soo Kim. Gbose: Generalized bandit orthogonalized semiparametric estimation. *arXiv preprint arXiv:2301.08781*, 2023.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Xiequan Fan and Shen Wang. Bernstein type inequalities for self-normalized martingales with applications. *Statistics*, 53(2):245–260, 2019.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, and Predag Klasnja. Action centered contextual bandits. *Advances in neural information processing systems*, 30, 2017.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

Gi-Soo Kim and Myunghee Cho Paik. Contextual multi-armed bandit algorithm for semiparametric reward model. In *International Conference on Machine Learning*, pages 3389–3397. PMLR, 2019.

Wonyoung Kim, Gi-soo Kim, and Myunghee Cho Paik. Doubly robust thompson sampling for linear payoffs. In *Advances in neural information processing systems*, 2021.

Junpei Komiyama, Taira Tsuchiya, and Junya Honda. Minimax optimal algorithms for fixed-budget best arm identification. *Advances in Neural Information Processing Systems*, 35:10393–10404, 2022.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785. PMLR, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press (preprint), 2019.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.

Zhaoqi Li, Lillian Ratliff, Kevin G Jamieson, Lalit Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

Martin Mladenov, Chih-wei Hsu, Vihan Jain, Eugene Ie, Christopher Colby, Nicolas Mayoraz, Hubert Pham, Dustin Tran, Ivan Vendrov, and Craig Boutilier. Demonstrating principled uncertainty modeling for recommender ecosystems with recsim ng. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 591–593, 2020.

Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. Pac-bayesian offline contextual bandits with guarantees. In *International Conference on Machine Learning*, pages 29777–29799. PMLR, 2023.

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

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(11), 2009.

Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

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

Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.

Appendix

Contents

Appendix A. Proof of Theorem 3

We define $\widetilde{p}_2, \ldots, \widetilde{p}_K$ as the G-optimal design for $\{x_2 - x_1, \ldots, x_K - x_1\}$, which solves Eq. (1). Also, we define

$$\mathbf{\Sigma}_{\text{opt},1} := \sum_{i=2}^{K} \widetilde{p}_i (x_i - x_1)(x_i - x_1)^\top.$$

By the definition of the G-optimal design, we have

$$\max_{i \in [K]} \|x_i - x_1\|_{\mathbf{\Sigma}_{\text{opt},1}^{-1}} \leq \sqrt{d}.$$

Recall that we set $\mathbf{p}^{\text{deo}} = \left(\frac{1}{2}, \frac{1}{2}\widetilde{p}_2, \ldots, \frac{1}{2}\widetilde{p}_K\right)$ and define $\mathbf{\Sigma}_{\text{deo}} := \sum_{i=1}^{K} p_i^{\text{deo}} (x_i - \bar{x})(x_i - \bar{x})^\top$, which is the covariance of the policy $\mathbf{p}^{\text{deo}}$. Applying Lemma 9, we obtain

$$\mathbf{\Sigma}_{\text{deo}} \succeq p_1^{\text{deo}} \sum_{j \neq 1} p_j^{\text{deo}} (x_j - x_1)(x_j - x_1)^\top = \frac{1}{2} \sum_{j \neq 1} \frac{1}{2} \widetilde{p}_j (x_j - x_1)(x_j - x_1)^\top = \frac{1}{4} \mathbf{\Sigma}_{\text{opt},1}.$$

Hence, the covariance $\mathbf{\Sigma}_{\text{deo}}$ of our policy satisfies

$$\mathbf{\Sigma}_{\text{deo}} \succeq \frac{1}{4} \mathbf{\Sigma}_{\text{opt},1}.$$

We are now ready to prove the theorem. By the definition of the G-optimal design, $\mathbf{\Sigma}_{\text{opt},1}$, we get for all $i \geq 2$:

$$\|x_i - x_1\|_{\mathbf{\Sigma}_{\text{deo}}^{-1}} \leq 2 \|x_i - x_1\|_{\mathbf{\Sigma}_{\text{opt},1}^{-1}} \leq 2\sqrt{d},$$

and this is the desired result.

Next, we prove the second inequality. Recall that we define $\widetilde{x}_i := x_i - \bar{x}_{\text{deo}}$ and observe

$$x_i - \bar{x}_{\text{deo}} = (x_i - x_1) - \sum_{j=1}^{K} p_j^{\text{deo}} (x_j - x_1),$$

and hence,

$$\begin{aligned} \|\widetilde{x}_i\|_{\mathbf{\Sigma}_{\text{deo}}^{-1}} \leq 2 \|\widetilde{x}_i\|_{\mathbf{\Sigma}_{\text{opt},1}^{-1}} &\leq 2 \Big(\sum_{j=1}^{K} p_j^{\text{deo}} \|x_j - x_1\|_{\mathbf{\Sigma}_{\text{opt},1}^{-1}} + \|x_i - x_1\|_{\mathbf{\Sigma}_{\text{opt},1}^{-1}} \Big) \\ &\leq 2 \big(\sum_{j=1}^{K} p_j^{\text{deo}} \sqrt{d} + \sqrt{d} \big) \leq 4\sqrt{d}. \end{aligned}$$

■

Lemma 9 *For any policy $\mathbf{p} = (p_1, \ldots, p_K) \in \Delta^{(K)}$ and $\bar{x}_{\mathbf{p}} = \sum_{i=1}^{K} p_i x_i$, the following holds:*

$$\mathbf{\Sigma}_{\mathbf{p}} := \sum_{i=1}^{K} p_i (x_i - \bar{x}_{\mathbf{p}})(x_i - \bar{x}_{\mathbf{p}})^\top = \sum_{i<j} p_i p_j \big(x_i - x_j\big)\big(x_i - x_j\big)^\top.$$

Proof We define $\text{out}(a, b) := ab^\top$, which is a bilinear function. Observe that

$$\begin{aligned}
\mathbf{\Sigma_p} &:= \sum_{i=1}^{K} p_i \big(x_i - \bar{x}_\mathbf{p}\big)\big(x_i - \bar{x}_\mathbf{p}\big)^\top \\
&= \sum_{i=1}^{K} p_i \text{out}(x_i, x_i) - \sum_{i=1}^{K} p_i \text{out}(x_i, \bar{x}_\mathbf{p}) - \sum_{i=1}^{K} p_i \text{out}(\bar{x}_\mathbf{p}, x_i) + \sum_{i=1}^{K} p_i \text{out}(\bar{x}_\mathbf{p}, \bar{x}_\mathbf{p}) \\
&= \sum_{i=1}^{K} p_i \text{out}(x_i, x_i) - \text{out}(\bar{x}_\mathbf{p}, \bar{x}_\mathbf{p}) \quad (\text{since } \sum_{i=1}^{K} p_i = 1) \\
&= \sum_{i=1}^{K} p_i \text{out}(x_i, x_i) - \sum_{i=1}^{K} p_i^2 \text{out}(x_i, x_i) - \sum_{i<j} p_i p_j (\text{out}(x_i, x_j) + \text{out}(x_j, x_i)) \\
&= \sum_{i=1}^{K} (p_i - p_i^2) \text{out}(x_i, x_i) - \sum_{i<j} p_i p_j (\text{out}(x_i, x_j) + \text{out}(x_j, x_i)) \\
&= \sum_{i=1}^{K} \sum_{j:j\neq i} p_i p_j \text{out}(x_i, x_i) - \sum_{i<j} p_i p_j (\text{out}(x_i, x_j) + \text{out}(x_j, x_i)) \\
&= \sum_{i\neq j} p_i p_j \text{out}(x_i, x_i) - \sum_{i<j} p_i p_j (\text{out}(x_i, x_j) + \text{out}(x_j, x_i)) \\
&= \sum_{i<j} p_i p_j \text{out}(x_i, x_i) + \sum_{i<j} p_i p_j \text{out}(x_j, x_j) - \sum_{i<j} p_i p_j (\text{out}(x_i, x_j) + \text{out}(x_j, x_i)) \\
&= \sum_{i<j} p_i p_j \text{out}(x_i - x_j, x_i - x_j),
\end{aligned}$$

and this concludes the proof.

Second proof. [2] Let X and Y be independent and identically distributed (i.i.d.) random vectors. The expected outer product of their difference is twice the covariance of X:

$$\mathbb{E}\left[(X - Y)(X - Y)^\top\right] = 2\mathbb{E}\left[XX^\top\right] - 2\mathbb{E}[X]\mathbb{E}[X]^\top = 2\,\text{Cov}(X).$$

Now, consider a discrete random vector X with the distribution $\mathbb{P}(X = x_i) = p_i$ for $i = 1, \ldots, K$. The left-hand side can be expressed as:

$$\mathbb{E}\left[(X - Y)(X - Y)^\top\right] = 2 \sum_{i<j} p_i p_j \left(x_i - x_j\right)\left(x_i - x_j\right)^\top.$$

The covariance of X is given by:

$$\text{Cov}(X) = \sum_{i=1}^{K} p_i (x_i - \bar{x}_\mathbf{p})(x_i - \bar{x}_\mathbf{p})^\top,$$

where $\bar{x}_\mathbf{p} = \mathbb{E}[X]$. Equating the two expressions concludes the proof. ■

[2] We thank reviewer #2 for valuable comments about correcting our proofs and for providing this simple and elegant second proof.

Appendix B. Proof of Theorem 4

We define $\boldsymbol{\Sigma}_{\mathbf{p}} := \mathbb{E}[\widetilde{x}_{a_s}\widetilde{x}_{a_s}^\top] = \sum_{i=1}^K p_i(x_i - \bar{x}_{\mathbf{p}})(x_i - \bar{x}_{\mathbf{p}})^\top$. Abusing notation slightly, we simply write $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\mathbf{p}}$ and $\bar{x} := \bar{x}_{\mathbf{p}}$ for this proof. We first prove the upper bound.

Step 0: Preparations. We set the ridge regularizer $\beta_t = \log(t/\delta)$ and the normalized ridge parameter $\lambda_t = \frac{\beta_t}{t}$. We set $\widehat{\mathbf{V}}_t := \sum_{i=1}^t \widetilde{x}_{a_s}\widetilde{x}_{a_s}^\top$ and $\mathbf{V}_t = t\boldsymbol{\Sigma} = \mathbb{E}[\widehat{\mathbf{V}}_t]$. Also, set $\widehat{\boldsymbol{\Sigma}}_t = \frac{\widehat{\mathbf{V}}_t}{t}$. For our choice of $\beta_t = \log(t/\delta)$, by Lemma 24, the following holds with probability at least $1 - \frac{\delta}{10}$ for some absolute constant $c > 0$:

$$\frac{1}{c}\big(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I}_d\big) \preceq \boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d \preceq c\big(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I}_d\big), \quad \frac{1}{c}\big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big) \preceq \mathbf{V}_t + \beta_t \mathbf{I}_d \preceq c\big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big). \tag{5}$$

Recall that we define $\widetilde{x}_{a_s} = x_{a_s} - \bar{x}$, and then $\mathbb{E}[\widetilde{x}_{a_s} \mid \mathcal{H}_{s-1}] = 0$. Also, we have $\mathbb{E}[\widetilde{x}_{a_s}\widetilde{x}_{a_s}^\top] = \boldsymbol{\Sigma}$.

We prove the upper bound of the theorem in four steps; afterward, we show its optimality.

Step 1: Error Decomposition. Using the definition of $\hat{\theta}_t$ from Eq. (2), decompose the estimation error $\hat{\theta}_t - \theta^\star$ as

$$\begin{aligned}
\hat{\theta}_t - \theta^\star &= (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\sum_{s=1}^t \widetilde{x}_{a_s}(x_{a_s}^\top\theta^\star + \nu_s + \eta_s) - (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)\theta^\star \\
&= (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\sum_{s=1}^t \widetilde{x}_{a_s}\big(\widetilde{x}_{a_s}^\top\theta^\star + \bar{x}^\top\theta^\star + \nu_s + \eta_s\big) \\
&\quad - (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\Big(\sum_{s=1}^t \widetilde{x}_{a_s}\widetilde{x}_{a_s}^\top\theta^\star\Big) - (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\beta_t\theta^\star \\
&= (\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\sum_{s=1}^t \widetilde{x}_{a_s}\big(\bar{x}^\top\theta^\star + \nu_s + \eta_s\big) - \beta_t(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\theta^\star \\
&= \underbrace{(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\sum_{s=1}^t \widetilde{x}_{a_s}\underbrace{\big(\bar{x}^\top\theta^\star + \nu_s\big)}_{:=q_s}}_{:=\mathscr{A}} + \underbrace{(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\sum_{s=1}^t \widetilde{x}_{a_s}\eta_s}_{:=\mathscr{B}} - \underbrace{\beta_t(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-1}\theta^\star}_{:=\mathscr{C}} \\
&:= \mathscr{A} + \mathscr{B} + \mathscr{C}.
\end{aligned}$$

Step 2: Bounding $\mathscr{C}$. For any $z \in \mathbb{R}^d$, we can bound $z^\top\mathscr{C}$ with probability at least $1 - \frac{\delta}{10}$ using Eq. (5):

$$\begin{aligned}
|z^\top\mathscr{C}| &\le \Big|\sqrt{\beta_t}z^\top(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-\frac{1}{2}}(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-\frac{1}{2}}\sqrt{\beta_t}\theta^\star\Big| \\
&\le \sqrt{\beta_t}\big\|z^\top(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-\frac{1}{2}}\big\|_2\big\|\sqrt{\beta_t}(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-\frac{1}{2}}\theta^\star\big\|_2 \\
&\le \sqrt{\beta_t}\frac{1}{\sqrt{t}}\sqrt{z^\top\big(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1}z} \times \sqrt{\beta_t}\big\|(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-\frac{1}{2}}\theta^\star\big\|_2 \\
&\lesssim \sqrt{\beta_t}\frac{1}{\sqrt{t}}\sqrt{z^\top\big(\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d\big)^{-1}z} \times \sqrt{\beta_t}\big\|(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d)^{-\frac{1}{2}}\theta^\star\big\|_2 \quad \text{(by Eq. (5))} \\
&\lesssim \frac{\sqrt{\beta_t L}}{\sqrt{t}}\|\theta^\star\|_2 \lesssim \frac{\sqrt{\beta_t L}}{\sqrt{t}} = \frac{\sqrt{\log(t/\delta)L}}{\sqrt{t}}.
\end{aligned}$$

Step 3: Bounding $\mathscr{B}$. Next, we bound $z^\top \mathscr{B}$. Recall

$$z^\top \mathscr{B} = \sum_{s=1}^{t} z^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} \widetilde{x}_{a_s} \eta_s.$$

Given $\{\widetilde{x}_{a_1}, \ldots \widetilde{x}_{a_t}\}$, the random variable $z^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} \widetilde{x}_{a_s} \eta_s$ is mean-zero and sub-Gaussian with proxy $\alpha_s := |z^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} \widetilde{x}_{a_s}|$ (since noise η_s is sampled independently). See that

$$\begin{aligned}
\sum_{s=1}^{t} \alpha_s^2 &= \sum_{s=1}^{t} z^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} \widetilde{x}_{a_s} \widetilde{x}_{a_s}^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} z \\
&= z^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} \widehat{\mathbf{V}}_t \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} z \\
&\leq z^\top \big(\widehat{\mathbf{V}}_t + \beta_t \mathbf{I}_d\big)^{-1} z.
\end{aligned}$$

Applying Bernstein's inequality for sub-Gaussian random variables (Boucheron et al., 2013), with probability at least $1 - \frac{\delta}{10}$,

$$|z^\top \mathscr{B}| \lesssim \sqrt{\frac{z^\top \big(\widehat{\mathbf{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} z}{t} \log\big(\frac{1}{\delta}\big)}.$$

Since Eq. (5) holds with probability at least $1 - \frac{\delta}{10}$, we have with probability at least $1 - \frac{2\delta}{10}$:

$$|z^\top \mathscr{B}| \lesssim \sqrt{\frac{z^\top \big(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d\big)^{-1} z}{t} \log\big(\frac{1}{\delta}\big)} \lesssim \frac{\sqrt{L \log\big(\frac{t}{\delta}\big)}}{\sqrt{t}}.$$

Step 3: Bounding $\mathscr{A}$. Lastly, we bound the most challenging term, $\mathscr{A}$. We define $q_s := \bar{x}^\top \theta^\star + \nu_s$. Under the boundedness Assumption 1, $|q_s| \leq 2$. Observe that

$$\begin{aligned}
\mathscr{A} &:= \frac{1}{t} \big(\widehat{\mathbf{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} \sum_{s=1}^{t} (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}} (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \widetilde{x}_{a_s} q_s \\
&= \frac{1}{t} \big(\widehat{\mathbf{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}} \sum_{s=1}^{t} \underbrace{(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \widetilde{x}_{a_s}}_{:= e_s} q_s \\
&:= \frac{1}{t} \big(\widehat{\mathbf{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}} \sum_{s=1}^{t} e_s q_s.
\end{aligned}$$

Here, $e_s := (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \widetilde{x}_{a_s} \in \mathbb{R}^d$ is mean zero given $\mathcal{H}_{s-1}$, and its variance satisfies

$$\begin{aligned}
\mathrm{var}(e_s \mid \mathcal{H}_{s-1}) &= (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \mathbb{E}\big[\widetilde{x}_{a_s} \widetilde{x}_{a_s}^\top \mid \mathcal{H}_{s-1}\big] (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \\
&= (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \mathbf{\Sigma} (\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \\
&\preceq \mathbf{I}_d.
\end{aligned}$$

Also, $\|e_s\|_2^2 \le M$ by the given condition. Then, for any $z \in \mathcal{X} - x_1$, we have

$$
\begin{aligned}
z^\top \mathscr{A} &= \frac{1}{t} z^\top \big(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}} \sum_{s=1}^t e_s q_s \\
&= \frac{1}{t} z^\top \big(\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d\big)^{-1} (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}} \sum_{s=1}^t e_s q_s \\
&\quad + \frac{1}{t} z^\top \big(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} \big(\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d - \widehat{\boldsymbol{\Sigma}}_t - \lambda_t \mathbf{I}_d\big) \big(\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d\big)^{-1} (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}} \sum_{s=1}^t e_s q_s \\
&= \frac{1}{t} z^\top \big(\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d\big)^{-\frac{1}{2}} \sum_{s=1}^t e_s q_s + \frac{1}{t} z^\top \big(\widehat{\boldsymbol{\Sigma}}_t + \lambda_t \mathbf{I}_d\big)^{-1} \big(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_t\big) \big(\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d\big)^{-\frac{1}{2}} \sum_{s=1}^t e_s q_s \\
&:= I + II,
\end{aligned}
$$

where we used the identity $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$ in the second equality.

<u>[3-1] Bounding Term I.</u> We first bound term I. See that

$$
|I| \le \frac{1}{t} \Big| z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \sum_{s=1}^t e_s q_s \Big| = \frac{1}{t} \Big| \sum_{s=1}^t q_s z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} e_s \Big|.
$$

Since $|q_s| \le 2$ and $\mathbb{E}[e_s e_s^\top \mid \mathcal{H}_{s-1}] \preceq \mathbf{I}_d$, we get

$$
\begin{aligned}
\operatorname{var}\big(q_s z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} e_s \big| \mathcal{H}_{s-1}\big) &= q_s^2 \mathbb{E}\big[z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} e_s e_s^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} z \big| \mathcal{H}_{s-1}\big] \\
&\le 4 z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-1} z \lesssim L.
\end{aligned}
$$

Also,

$$
\big| q_s z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} e_s \big| \lesssim \big\| z^\top (\boldsymbol{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}} \big\|_2 \|e_s\|_2 \lesssim \sqrt{L}\sqrt{M} \lesssim \sqrt{LM}.
$$

Since $\mathbb{E}[e_s \mid \mathcal{H}_{s-1}] = 0$ and q_s is measurable w.r.t. $\mathcal{H}_{s-1}$, by applying Bernstein's inequality for sum of martingale differences (see Fan and Wang 2019), we get with probability at least $1 - \frac{\delta}{10}$:

$$
|I| \lesssim \frac{1}{t} \sqrt{Lt \log\big(\frac{10}{\delta}\big)} + \frac{1}{t} \sqrt{LM} \log\big(\frac{10}{\delta}\big) \lesssim \sqrt{\frac{L}{t} \log\big(\frac{1}{\delta}\big)} + \frac{\sqrt{LM}}{t} \log\big(\frac{1}{\delta}\big).
$$

[3-2] Bounding Term II. This term is a higher order term, such as $\widetilde{\mathcal{O}}(\frac{1}{t})$. Observe that

$$\begin{aligned}
|II| &= \left|\frac{1}{t}z^\top\big(\widehat{\boldsymbol{\Sigma}}_t+\lambda_t\mathbf{I}_d\big)^{-1}\big(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_t\big)(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\sum_{s=1}^{t}e_sq_s\right| \\
&= \left|\frac{1}{t}z^\top\underbrace{\big(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d\big)^{-\frac{1}{2}}(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}}\big(\widehat{\boldsymbol{\Sigma}}_t+\lambda_t\mathbf{I}_d\big)^{-1}\right. \\
&\quad \left.\times\underbrace{(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}}\big(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_t\big)(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\sum_{s=1}^{t}e_sq_s\right| \\
&\le \frac{1}{t}\big\|z^\top(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big\|_2\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}\big(\widehat{\boldsymbol{\Sigma}}_t+\lambda_t\mathbf{I}_d\big)^{-1}(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}\big\|_{\mathrm{op}} \\
&\quad \times\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_t\big)(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big\|_{\mathrm{op}}\Big\|\sum_{s=1}^{t}e_sq_s\Big\|_2 \\
&\le \frac{1}{t}\sqrt{L}\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_t+\lambda_t\mathbf{I}_d)^{-1}(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}\big\|_{\mathrm{op}} \\
&\quad \times\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_t\big)(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big\|_{\mathrm{op}}\Big\|\sum_{s=1}^{t}e_sq_s\Big\|_2.
\end{aligned}$$

First, with probability at least $1-\frac{\delta}{10}$, Eq. (5) gives

$$\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_t+\lambda_t\mathbf{I}_d)^{-1}(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{\frac{1}{2}}\big\|_{\mathrm{op}}\le 2.$$

Next, by applying the dimension-free martingale difference bound (Lemma 10 of Kim et al. (2021)), we aim to bound $\|\sum_{s=1}^{t}e_sq_s\|_2$. Since $\|q_se_s\|_2\le 2\sqrt{M}$ and $\mathbb{E}[e_s\mid\mathcal{H}_{s-1}]=0$, we can bound it with probability at least $1-\frac{\delta}{10}$:

$$\Big\|\sum_{s=1}^{t}e_sq_s\Big\|_2\lesssim\sqrt{M}\sqrt{t\log\big(\frac{10}{\delta}\big)}\lesssim\sqrt{tM\log\big(\frac{1}{\delta}\big)}.$$

Lastly, to bound $\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_t\big)(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big\|_{\mathrm{op}}$, we use the result of Lemma 10 and obtain with probability at least $1-\frac{\delta}{10}$:

$$\big\|(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big(\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}_t\big)(\boldsymbol{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\big\|_{\mathrm{op}}\lesssim\sqrt{\frac{M\log\big(\frac{10d}{\delta}\big)}{t}}\lesssim\sqrt{\frac{M\log\big(\frac{d}{\delta}\big)}{t}}.$$

Combining the above, with probability at least $1 - \frac{3\delta}{10}$:

$$\begin{aligned}
|II| &\lesssim \frac{\sqrt{L}}{t}\left\|(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}}(\widehat{\mathbf{\Sigma}}_t + \lambda_t \mathbf{I}_d)^{-1}(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{\frac{1}{2}}\right\|_{\mathrm{op}} \\
&\quad \times \left\|(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}}\big(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_t\big)(\mathbf{\Sigma} + \lambda_t \mathbf{I}_d)^{-\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\sum_{i=1}^{t} e_i a_i\right\|_2 \\
&\lesssim \frac{\sqrt{L}}{t}\frac{\sqrt{M\log\left(\frac{d}{\delta}\right)}}{\sqrt{t}}\sqrt{t}\sqrt{M\log\left(\frac{1}{\delta}\right)} \\
&\lesssim \frac{\sqrt{L}M\log\left(\frac{d}{\delta}\right)}{t}.
\end{aligned}$$

Step 4: Final Upper Bound. Combining the established bounds, with probability at least $1 - \delta$, we have

$$|z^\top(\hat{\theta}_t - \theta^\star)| \lesssim \frac{\sqrt{\log\left(\frac{t}{\delta}\right)L}}{\sqrt{t}} + \frac{\sqrt{L}M}{t}\log\left(\frac{d}{\delta}\right).$$

Step 5: Proof of Optimality. Consider the case when the shift is given by $\nu_s = -\bar{x}_{\mathbf{p}}\theta^\star$ for all $1 \leq s \leq t$. Then, our problem has the reward structure

$$r_s = (x_{a_s} - \bar{x}_{\mathbf{p}})^\top \theta^\star + \eta_s,$$

which corresponds to a linear model with independent variables (covariates) $\{x_{a_s} - \bar{x}_{\mathbf{p}}\}_{s=1}^t$ and responses $\{r_s\}_{s=1}^t$. However, our problem is more challenging since we do not know the exact shift (so we do not know this is actually a linear model), nor do we know $\{x_{a_s} - \bar{x}_{\mathbf{p}}\}$ are the covariates of the linear model.

In this case, the second-moment matrix of the linear model's covariates is

$$\mathbb{E}\left[(x_{a_s} - \bar{x}_{\mathbf{p}})(x_{a_s} - \bar{x}_{\mathbf{p}})^\top\right] = \mathbf{\Sigma}_{\mathbf{p}}.$$

Thus, the known lower bound for the estimation error at z for this linear model is $\frac{\sqrt{L}}{\sqrt{t}}$. This result is widely known, but we include a proof reference for completeness. We can interpret this as a transfer learning problem in a linear model, where the source distribution has a second-moment matrix $\mathbf{\Sigma}_{\mathbf{p}}$, while the target distribution is a Dirac distribution at z. By applying Theorem 3.2 of Ge et al. (2023), we obtain the desired result directly.

■

We now present the lemma and its proof used in the proof of Theorem 4.

Lemma 10 *For any $\lambda > 0$,*

$$\left\|(\mathbf{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}}\big(\widehat{\mathbf{\Sigma}}_t - \mathbf{\Sigma}\big)(\mathbf{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}}\right\|_{\mathrm{op}} \lesssim \sqrt{\frac{M\log\left(\frac{d}{\delta}\right)}{t}}$$

holds with probability at least $1 - \delta$.

Proof We apply Lemma 27 to $\boldsymbol{x}_s = (\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}} \widetilde{x}_{a_s}$. Then $\|\boldsymbol{x}_s\|_2 \leq \sqrt{M}$, so we set $S \leftarrow \sqrt{M}$. Next, we set $v = 1$ in the lemma because $\|\mathbb{E}[\boldsymbol{x}_s \boldsymbol{x}_s^\top]\|_{\text{op}} = \|(\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}}\|_{\text{op}} \leq 1$. Finally, we set $r = d$ because $\operatorname{tr}(\mathbb{E}[\boldsymbol{x}_s \boldsymbol{x}_s^\top]) = \operatorname{tr}\big((\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}}\big) \leq \operatorname{tr}\big((\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}} (\boldsymbol{\Sigma} + \lambda \mathbf{I}) (\boldsymbol{\Sigma} + \lambda \mathbf{I})^{-\frac{1}{2}}\big) = d$. ■

Appendix C. Proof of Corollary 5

Assume that we use the greedy policy with the estimator $\hat{\theta}_t$, i.e., $a_t = \arg\max_{i \in [K]} x_i^\top \hat{\theta}_t$ and it is equivalent to $a_t = \arg\max_{i \in [K]} \big(x_i - x_1\big)^\top \hat{\theta}_t$. Suppose $\max_{i \in [K]} |(x_i - x_1)^\top (\hat{\theta}_t - \theta^\star)| \leq \frac{\varepsilon}{2}$ holds. Then

$$\big(x_{a_t} - x_1\big)^\top \hat{\theta}_t \geq \big(x_{a^\star} - x_1\big)^\top \hat{\theta}_t,$$
$$\big(x_{a_t} - x_1\big)^\top \theta^\star \leq \big(x_{a^\star} - x_1\big)^\top \theta^\star,$$

holds and it leads

$$\big(x_{a^\star} - x_{a_t}\big)^\top \theta^\star \leq \varepsilon.$$

This implies that our policy is ε-PAC.

Hence, to achieve the (ε, δ)-PAC property, by Theorem 4, it suffices to ensure

$$C_1 \Big(\frac{\sqrt{d \log\big(\frac{Kt}{\delta}\big)}}{\sqrt{t}} + \frac{d^{\frac{3}{2}} \log\big(\frac{dK}{\delta}\big)}{t} \Big) < \frac{\varepsilon}{2}.$$

The term K appears to guarantee concentrations over all $\mathcal{X} - x_1$. If

$$C_1 \frac{\sqrt{d \log\big(\frac{Kt}{\delta}\big)}}{\sqrt{t}} < \frac{\varepsilon}{4}, \tag{6}$$

$$C_1 \frac{d \log\big(\frac{dK}{\delta}\big)}{t} < \frac{\varepsilon}{4}, \tag{7}$$

then the above relation also holds. The first inequality (6) suffices if

$$C_1 \frac{\sqrt{d \log t}}{\sqrt{t}} < \frac{\varepsilon}{8} \quad \text{and} \quad C_1 \frac{\sqrt{d \log\big(\frac{K}{\delta}\big)}}{\sqrt{t}} < \frac{\varepsilon}{8}.$$

A sufficient condition is

$$\begin{aligned} & \frac{t}{\log t} > c \frac{d}{\varepsilon^2} \quad \text{and} \quad t > c \frac{d \log\big(\frac{K}{\delta}\big)}{\varepsilon^2}, \\ \Longleftarrow \quad & t \gtrsim \frac{d \log\big(\frac{d}{\varepsilon}\big)}{\varepsilon^2} + \frac{d \log\big(\frac{K}{\delta}\big)}{\varepsilon^2} \\ \Longleftarrow \quad & t \gtrsim \frac{d \log\big(\frac{dK}{\varepsilon \delta}\big)}{\varepsilon^2}. \end{aligned}$$

The second inequality (7) is equivalent to

$$t \gtrsim \frac{d^{\frac{3}{2}} \log\left(\frac{dK}{\delta}\right)}{\varepsilon}.$$

Hence, if

$$t \gtrsim \frac{d \log\left(\frac{dK}{\varepsilon\delta}\right)}{\varepsilon^2} + \frac{d^{\frac{3}{2}} \log\left(\frac{dK}{\delta}\right)}{\varepsilon},$$

we obtain the desired ε-PAC property with probability at least $1 - \delta$.

Appendix D. Proofs for Section 5

We define the end-time of phase ℓ as t_ℓ. Recall that we denote the estimator computed at the end of phase ℓ by $\hat{\theta}_{(\ell)}$. Define $L_\star := \lceil \log_2(1/\Delta_\star) \rceil + 1$. We set L_T to be the last phase until time T. Note that $L_T, L_\star$ are deterministic values. Recall that $\mathcal{A}_\ell(1)$ is defined as the arm of $\mathcal{A}_\ell$ with smallest index.

D.1. Framework for Regret Analysis

Lemma 11 (Good event) *With probability at least* $1 - \delta$*,*

$$\left| x^\top \left(\hat{\theta}_{(\ell)} - \theta^\star\right) \right| \le \frac{\varepsilon_\ell}{2}$$

for any $x \in \mathcal{A}_\ell - \mathcal{A}_\ell(1)$ *and any phase* $1 \le \ell \le L_T$*. We call this event* $\mathscr{E}_T$*.*

Proof By Corollary 5, for any phase ℓ,

$$\mathbb{P}\left[\exists\, x \in \mathcal{A}_\ell - \mathcal{A}_\ell(1) \text{ s.t. } \left| x^\top \left(\hat{\theta}_{(\ell)} - \theta^\star\right) \right| > \frac{\varepsilon_\ell}{2}\right] \le \frac{\delta}{\ell(\ell+1)}.$$

By taking the union bound over all ℓ, we obtain the desired result. ■

Lemma 12 *Under the event* $\mathscr{E}_T$*, for each phase* ℓ*, the action set* $\mathcal{A}_\ell$ *contains* $a^\star$*.*

Proof We prove this by contradiction. Suppose $a^\star$ is eliminated after some phase ℓ. Then there exists some arm a' such that

$$x_{a'}^\top \hat{\theta}_{(\ell)} > x_{a^\star}^\top \hat{\theta}_{(\ell)} + \varepsilon_\ell \;\Leftrightarrow\; \left(x_{a'} - x_{\mathcal{A}_\ell(1)}\right)^\top \hat{\theta}_{(\ell)} > \left(x_{a^\star} - x_{\mathcal{A}_\ell(1)}\right)^\top \hat{\theta}_{(\ell)} + \varepsilon_\ell.$$

However, under the event $\mathscr{E}_T$, by Lemma 11, we have

$$\left| \left(x_{a'} - x_{\mathcal{A}_\ell(1)}\right)^\top \hat{\theta}_{(\ell)} - \left(x_{a'} - x_{\mathcal{A}_\ell(1)}\right)^\top \theta^\star \right| \le \frac{\varepsilon_\ell}{2}, \; \left| \left(x_{a^\star} - x_{\mathcal{A}_\ell(1)}\right)^\top \hat{\theta}_{(\ell)} - \left(x_{a^\star} - x_{\mathcal{A}_\ell(1)}\right)^\top \theta^\star \right| \le \frac{\varepsilon_\ell}{2}$$

and

$$\left(x_{a^\star} - x_{\mathcal{A}_\ell(1)}\right)^\top \theta^\star > \left(x_{a'} - x_{\mathcal{A}_\ell(1)}\right)^\top \theta^\star$$

which leads a contradiction. ■

Lemma 13 *Under the event $\mathscr{E}_T$, any arm a in $\mathcal{A}_\ell$ satisfies*

$$x_a^\top \theta^\star > x_{a^\star}^\top \theta^\star - 4\varepsilon_\ell.$$

Proof Since $a^\star \in \mathcal{A}_{\ell-1}$ and arm a is not eliminated, we have

$$\begin{aligned} & x_a^\top \hat{\theta}_{(\ell-1)} > x_{a^\star}^\top \hat{\theta}_{(\ell-1)} - \varepsilon_{\ell-1} \\ \Leftrightarrow\ & (x_a - x_{\mathcal{A}_{\ell-1}(1)})^\top \hat{\theta}_{(\ell-1)} > (x_{a^\star} - x_{\mathcal{A}_{\ell-1}(1)})^\top \hat{\theta}_{(\ell-1)} - \varepsilon_{\ell-1}. \end{aligned}$$

Under the event $\mathscr{E}_T$, by Lemma 11, we get

$$\begin{aligned} \left|(x_{a'} - x_{\mathcal{A}_{\ell-1}(1)})^\top \hat{\theta}_{(\ell-1)} - (x_{a'} - x_{\mathcal{A}_{\ell-1}(1)})^\top \theta^\star\right| &\leq \frac{\varepsilon_{\ell-1}}{2} \\ \left|(x_{a^\star} - x_{\mathcal{A}_{\ell-1}(1)})^\top \hat{\theta}_{(\ell-1)} - (x_{a^\star} - x_{\mathcal{A}_{\ell-1}(1)})^\top \theta^\star\right| &\leq \frac{\varepsilon_{\ell-1}}{2}, \end{aligned}$$

and it leads

$$(x_a - x_{\mathcal{A}_{\ell-1}(1)})^\top \theta^\star > (x_{a^\star} - x_{\mathcal{A}_{\ell-1}(1)})^\top \theta^\star - 2\varepsilon_{\ell-1}.$$

which is equivalent to

$$x_a^\top \theta^\star > x_{a^\star}^\top \theta^\star - 2\varepsilon_{\ell-1}.$$

■

Lemma 14 (Regret formula without gap) *Under the event $\mathscr{E}_T$, the regret is bounded by*

$$\mathbf{Reg}(T) \lesssim d2^{L_T}\Big(\log\big(\frac{dKL_T}{\delta}\big) + L_T\Big) + d^{\frac{3}{2}} \log\big(\frac{dKL_T}{\delta}\big) L_T.$$

Proof Under the event $\mathscr{E}_T$, by Lemma 13, we have

$$\mathbf{Reg}(T) \lesssim \sum_{\ell=1}^{L_T} n_\ell \varepsilon_\ell \lesssim \sum_{\ell=1}^{L_T} \Big(\frac{d\log\big(\frac{dK\ell}{\delta\varepsilon_\ell}\big)}{\varepsilon_\ell^2} + \frac{d^{\frac{3}{2}}}{\varepsilon_\ell} \log\big(\frac{dK\ell}{\delta}\big)\Big)\varepsilon_\ell.$$

Hence,

$$\begin{aligned} \mathbf{Reg}(T) &\lesssim \sum_{\ell=1}^{L_T} \frac{d\log\big(\frac{dKL_T}{\delta\varepsilon_\ell}\big)}{\varepsilon_\ell} + d^{\frac{3}{2}} \log\big(\frac{dK\ell}{\delta}\big) \\ &\lesssim \sum_{\ell=1}^{L_T} \big(d\log\big(\frac{dKL_T}{\delta}\big) + d\log\frac{1}{\varepsilon_\ell}\big)\frac{1}{\varepsilon_\ell} + d^{\frac{3}{2}} \log\big(\frac{dKL_T}{\delta}\big). \end{aligned}$$

Noting that $\log(\frac{1}{\varepsilon_\ell}) \leq \ell$, the above is

$$\begin{aligned} \mathbf{Reg}(T) &\lesssim \sum_{\ell=1}^{L_T} \frac{d\log\big(\frac{dKL_T}{\delta}\big) + d\ell}{\varepsilon_\ell} + L_T d^{\frac{3}{2}} \log\big(\frac{dKL_T}{\delta}\big) \\ &\lesssim d2^{L_T}\Big(\log\big(\frac{dKL_T}{\delta}\big) + L_T\Big) + d^{\frac{3}{2}} \log\big(\frac{dKL_T}{\delta}\big) L_T. \end{aligned}$$

■

Lemma 15 *The total number of phases L_T is bounded by*

$$2^{L_T} \le c\sqrt{\frac{T}{d\log(\frac{dK}{\delta})}}.$$

Proof Observe that

$$T \gtrsim c\sum_{\ell=1}^{L_T-1} d\log\Big(\frac{dK}{\delta}\Big)4^{\ell} \gtrsim cd\log\Big(\frac{dK}{\delta}\Big)4^{L_T}.$$

Hence

$$4^{L_T} \lesssim \frac{T}{d\log(\frac{dK}{\delta})} \Longrightarrow 2^{L_T} \lesssim \sqrt{\frac{T}{d\log(\frac{dK}{\delta})}}.$$

■

D.2. Proof of Theorem 6

Proof Under the event $\mathscr{E}_T$, by combining Lemma 13 and 14 and Lemma 15, we get

$$\begin{aligned}
&\mathbf{Reg}(T)\\
&\lesssim d2^{L_T}(\log(\frac{dKL_T}{\delta}) + L_T) + d^{\frac{3}{2}}\log(\frac{dKL_T}{\delta})L_T\\
&\lesssim d\sqrt{\frac{T}{d\log(\frac{dK}{\delta})}}\left(\log(\frac{dKL_T}{\delta}) + \log(\frac{T}{d\log(\frac{dK}{\delta})})\right) + d^{\frac{3}{2}}\log(\frac{dKL_T}{\delta})L_T \quad \text{(by Lemma 15)}\\
&\lesssim \sqrt{\frac{dT}{\log(\frac{dK}{\delta})}}(\log(\frac{dK}{\delta}) + \log L_T + \log(\frac{T}{d})) + d^{\frac{3}{2}}\log(\frac{dKT}{\delta})L_T\\
&\lesssim \sqrt{\frac{dT}{\log(\frac{dK}{\delta})}}(\log(\frac{K}{\delta}) + \log T + \log\log T) + d^{\frac{3}{2}}\log(\frac{dKT}{\delta})\log(T/d)\\
&\lesssim \sqrt{dT\log(\frac{K}{\delta})} + \sqrt{dT}\log T + d^{\frac{3}{2}}\log(\frac{dKT}{\delta})\log(T/d).
\end{aligned}$$

Since $\mathbb{P}\big[\mathscr{E}_T\big] \ge 1-\delta$ by Lemma 11, the result follows. ■

D.3. Proof of Theorem 7

Proof Under the event $\mathscr{E}_T$, by Lemma 12, the optimal arm $a^\star$ is contained in every phase ℓ. Consider the phase ℓ with $\ell > L_\star$. Under the event $\mathscr{E}_T$, by Lemma 13, the optimal arm $a^\star$ is the unique arm in the phase ℓ, i.e. $\mathcal{A}_\ell = \{a^\star\}$. Recall that

$$2^{L_\star} \asymp \frac{1}{\Delta_\star} \asymp \frac{1}{\varepsilon_{L_\star}}.$$

Using the Lemma 13, under the event $\mathscr{E}_T$, the regret is bounded by

$$\begin{aligned}
\mathbf{Reg}(T) &\lesssim \sum_{\ell=1}^{L_\star} n_\ell \varepsilon_{\ell-1} \lesssim \sum_{\ell=1}^{L_\star} n_\ell \varepsilon_\ell \\
&\lesssim \sum_{\ell=1}^{L_\star} \Big(\frac{d \log\big(\frac{dK\ell}{\delta\varepsilon_\ell}\big)}{\varepsilon_\ell^2} + \frac{d^{\frac{3}{2}}}{\varepsilon_\ell} \log\big(\frac{dK\ell}{\delta}\big) \Big) \varepsilon_\ell \\
&\lesssim \sum_{\ell=1}^{L_\star} \frac{d \log\big(\frac{dKL_\star}{\delta\varepsilon_\ell}\big)}{\varepsilon_\ell} + d^{\frac{3}{2}} \log\big(\frac{dK\ell}{\delta}\big) \\
&\lesssim d 2^{L_\star} \log\Big(\frac{dK}{\delta\varepsilon_{L_\star}}\Big) + d 2^{L_\star} \log L_\star + d^{\frac{3}{2}} \log\Big(\frac{dKL_\star}{\delta}\Big) L_\star \\
&\lesssim d 2^{L_\star} \log\Big(\frac{dK}{\delta\Delta_\star}\Big) + d 2^{L_\star} \log\log\big(\frac{1}{\Delta_\star}\big) + d^{\frac{3}{2}} \log\Big(\frac{dKL_\star}{\delta}\Big) L_\star \\
&\lesssim d 2^{L_\star} \log\Big(\frac{dK}{\delta\Delta_\star}\Big) + d^{\frac{3}{2}} \log\Big(\frac{dK}{\delta\Delta_\star}\Big) L_\star.
\end{aligned}$$

Hence, finally we get

$$\mathbf{Reg}(T) \lesssim \Big(\frac{d}{\Delta_\star} + d^{\frac{3}{2}}\Big) \log\Big(\frac{dK}{\delta\Delta_\star}\Big) \log\big(\frac{1}{\Delta_\star}\big).$$

Since $\mathbb{P}\big[\mathscr{E}_T\big] \geq 1 - \delta$ by Lemma 11, we get the desired result. ■

D.4. Proof of Theorem 8

Proof of BAI. As shown in the proof of Theorem 7, under the event $\mathscr{E}_T$, the arm $a_\star$ is the unique remaining arm after phase $\ell > L_\star$. Since $\sum_{\ell=1}^{L_\star} n_\ell \asymp n_{L_\star} = \widetilde{\mathcal{O}}(\frac{d}{\Delta_\star^2})$, we get the wanted result.

Proof of PAC. Under the event $\mathscr{E}_T$, by Lemma 13, in phase ℓ, each remaining arm has suboptimality gap smaller than $4\varepsilon_\ell = \frac{1}{2^{\ell-1}}$. Then we obtain the result directly.

■

Appendix E. K-Independent Results

Next, we present a result without K (action cardinality) dependence. When $\log K \gtrsim d$, i.e., $K \gg d$, the previously established results may be suboptimal. First, using Lemma 20, we can easily get an estimation error bound and a sample complexity without explicit K-dependence.

Corollary 16 (K-independent estimation error bound) *When we run pure exploration with the policy $\mathbf{p}^{\text{deo}}$, the estimator at time t with ridge regularizer $\beta_t = \log(t/\delta)$, named $\hat{\theta}_t$, satisfies*

$$\max_{i\in[K]} \big| (x_i - x_1)^\top \big(\hat{\theta}_t - \theta^\star\big) \big| \lesssim \frac{d\sqrt{\log\big(\frac{t}{\delta}\big)}}{\sqrt{t}}.$$

Corollary 17 (K-independent PAC bound) *For a fixed $\varepsilon > 0$, using the policy $\mathbf{p}^{\mathrm{deo}}$, we sample from that policy for t rounds and then exit, as same as Section 4.2. Then if*

$$t \geq C_3 \frac{d^2}{\varepsilon^2} \log\Big(\frac{d}{\varepsilon\delta}\Big) = \widetilde{\mathcal{O}}\Big(\frac{d}{\varepsilon^2}\Big),$$

for some absolute constant C_3, our greedy policy with the estimator $\hat{\theta}_t$ is (ε, δ)-PAC.

Corollary 16 and Corollary 17 give the sample complexity to achieve an ε-estimation error of the estimator. This coincides with the known optimal result for linear bandits, without any dependence on K.

Adaptive Low-Regret Algorithm. For SBE, there is a sampling part with each arm pulled n_ℓ times. With a slight modification of the sampling number n_ℓ in our low-regret algorithm SBE, we can obtain an adaptive regret that enjoys

$$\min\Big(\widetilde{\mathcal{O}}\big(\sqrt{dT\log K}\big), \widetilde{\mathcal{O}}\big(d\sqrt{T}\big)\Big)$$

while maintaining PAC and BAI properties. This algorithm is exactly the same as SBE, except setting

$$n_\ell = 4\min\Big(C_2\Big\lceil\frac{d}{\varepsilon_\ell^2}\log\big(\frac{dK\ell(\ell+1)}{\delta\varepsilon_\ell}\big) + \frac{d^{\frac{3}{2}}}{\varepsilon_\ell}\log\big(\frac{dK\ell(\ell+1)}{\delta}\big)\Big\rceil, C_3\Big\lceil\frac{d^2}{\varepsilon_\ell^2}\log\big(\frac{d\ell(\ell+1)}{\delta\varepsilon_\ell}\big)\Big\rceil\Big).$$

We then have the following regret bound for this modified algorithm.

Theorem 18 (Adaptive regret bound) *The adaptive version of SBE has cumulative regret bound*

$$\mathbf{Reg}(T) \lesssim \min\Big(\widetilde{\mathcal{O}}\big(\sqrt{dT\log K}\big), \widetilde{\mathcal{O}}\big(d\sqrt{T}\big)\Big)$$

with probability at least $1 - \frac{1}{T^2}$.

Discussion. Our regret bound is robust for the case $d \ll K$. It is strictly sharper than the previous works Kim and Paik (2019); Krishnamurthy et al. (2018), which exhibit $\widetilde{\mathcal{O}}(d\sqrt{T})$ or $\widetilde{\mathcal{O}}(d^{\frac{3}{2}}\sqrt{T})$-type bounds.

Appendix F. Proofs for K-independent Results

F.1. Cauchy-Schwarz Based Error Analysis

In this section, we present the concentration of the estimation error bound using the self-normalized inequality from Krishnamurthy et al. (2018), specifically Lemma 11 of that work. We consider the situation in which arms are sampled from a fixed policy.

Lemma 19 (Modified Lemma 11 from Krishnamurthy et al. 2018) *Under Assumption 1 (boundedness), with probability at least $1 - \delta$, the following holds for t:*

$$\|\hat{\theta}_t - \theta^\star\|_{\widehat{\mathbf{V}}_t + \beta\mathbf{I}_d} \leq \sqrt{\beta} + \sqrt{27d\log\big(1 + \frac{t}{d\delta}\big) + 54\log\big(\frac{t}{\delta}\big)},$$

for $\beta = \log(t/\delta)$ under the fixed policy.

The original selection of β is $\beta \asymp d\log(t/\delta)$; however, using Lemma 26, we can relax it for the fixed policy case. Even though we choose the larger $\beta \asymp d\log(t/\delta)$, it yields the same bound when we use that lemma.

Proof Following the proof of Lemma 11 of Krishnamurthy et al. (2018), the only role of the ridge regularizer β is to guarantee

$$\frac{1}{c}(\widehat{\mathbf{V}}_t + \beta\mathbf{I}_d) \preceq \mathbf{V}_t + \beta\mathbf{I}_d \preceq c(\widehat{\mathbf{V}}_t + \beta\mathbf{I}_d).$$

Using Lemma 26, we can prove that $\beta \asymp \log(t/\delta)$ is enough to satisfy this relation. The remaining proof is exactly the same as in Krishnamurthy et al. (2018). ■

Lemma 20 (Cauchy-Schwarz based estimation error bound) *Under the boundedness assumption, with probability at least $1-\delta$, the following holds for t:*

$$\sqrt{t}\|\hat{\theta}_t - \theta^\star\|_{\mathbf{\Sigma}+\lambda\mathbf{I}_d} = \|\hat{\theta}_t - \theta^\star\|_{\mathbf{V}_t+\beta\mathbf{I}_d} \lesssim \sqrt{d\log(t/\delta)}.$$

when we choose $\beta = \log(t/\delta)$ under the fixed policy. Here, c is an absolute constant.

Proof Using Lemma 26, our choice of β makes the following with probability at least $1-\frac{\delta}{10}$:

$$\frac{1}{c}(\widehat{\mathbf{V}}_t + \beta\mathbf{I}_d) \preceq \mathbf{V}_t + \beta\mathbf{I}_d \preceq c(\widehat{\mathbf{V}}_t + \beta\mathbf{I}_d),$$

and it follows directly by combining Lemma 19. ■

F.2. Proof of Corollary 16

For any $z \in \mathcal{X} - x_1$, using Lemma 20, we get

$$\begin{aligned}
|z^\top(\hat{\theta} - \theta^\star)| &\leq \|z\|_{(\mathbf{\Sigma}_{\mathrm{deo}}+\lambda\mathbf{I}_d)^{-1}}\|\hat{\theta}_t - \theta^\star\|_{\mathbf{\Sigma}_{\mathrm{deo}}+\lambda\mathbf{I}_d} \\
&\lesssim \|z\|_{(\mathbf{\Sigma}_{\mathrm{deo}}+\lambda\mathbf{I}_d)^{-1}}\frac{\sqrt{d\log(t/\delta)}}{\sqrt{t}} \\
&\lesssim \frac{d\sqrt{\log(t/\delta)}}{\sqrt{t}}
\end{aligned}$$

with probability $1-\delta$.

F.3. Proofs of Corollary 17.

We can use the same argument from the proof of Corollary 5, combined with the result of Corollary 16. ■

F.4. Proof of Theorem 18

We now prove our theorem.

Lemma 21 (Good events) *With probability at least* $1 - 2\delta$,

$$|x^T(\hat{\theta}_{(\ell)} - \theta^\star)| \leq \frac{\varepsilon_\ell}{2}$$

for any arm $x \in \mathcal{A} - \mathcal{A}_\ell(1)$ *and any phase* $1 \leq \ell \leq L_T$. *We call this event* $\mathscr{F}_T$.

Proof We use the same argument from the proof of Lemma 11. Combining Corollary 5 and Corollary 17, we can prove it in the same way. ■

Lemma 22 *Under the event* $\mathscr{F}_T$, *for each phase* ℓ, *the action set* A_ℓ *contains* $a^\star$. *Any arm* a *in* $\mathcal{A}_\ell$ *satisfies*

$$x_a^\top \theta^\star > x_{a^\star}^\top \theta^\star - 4\varepsilon_\ell.$$

Proof The same as the proof of Lemma 12 and 13. ■

Main Proof of Theorem 18 Define

$$\begin{aligned} n_\ell &= \min\left(4\lceil C_4 \frac{d^2}{\varepsilon_\ell^2}\log(\frac{d\ell(\ell+1)}{\varepsilon\delta})\rceil, \lceil C_2 \frac{d\log(\frac{dK\ell(\ell+1)}{\delta\varepsilon_\ell})}{\varepsilon_\ell^2} + C_3 \frac{d^{\frac{3}{2}}\log^2(\frac{dK\ell(\ell+1)}{\delta})}{\varepsilon_\ell}\rceil\right) \\ &:= \min(n_\ell^{(1)}, n_\ell^{(2)}). \end{aligned}$$

For each time t, define the phase $\ell(t)$ that t belongs to and define $e(t) := \frac{1}{2^{\ell(t)}}$. Under the event $\mathscr{F}_T$, by Lemma 22, we have

$$\begin{aligned} \mathbf{Reg}(T) &\lesssim \sum_{t=1}^T e(t) \\ &\lesssim \sum_{t=1}^T \min(e_1(t), e_2(t)) \\ &\lesssim \min\Big(\sum_{t=1}^T e_1(t), \sum_{t=1}^T e_2(t)\Big). \end{aligned}$$

First, using the same argument from the proof of Theorem 6, we have

$$\sum_{t=1}^T e_1(t) \lesssim \sqrt{dT\log(\frac{K}{\delta})} + \sqrt{dT}\log T + d^{\frac{3}{2}}\log(\frac{dKT}{\delta}).$$

Next, we aim to bound $\sum_{t=1}^T e_2(t)$. Let $L_T^{(k)}$ for $k \in \{0, 1\}$ be the smallest value of L satisfying

$$\sum_{\ell=1}^L n_\ell^{(k)} \geq T.$$

Also, observe that

$$
\begin{aligned}
\sum_{t=1}^{T} e_2(t) &\lesssim \sum_{\ell=1}^{L_T^{(2)}} n_\ell^{(2)} \varepsilon_{\ell-1} \\
&\lesssim \sum_{\ell=1}^{L_T^{(2)}} n_\ell^{(2)} \varepsilon_\ell \\
&\lesssim \sum_{\ell=1}^{L_T^{(2)}} \frac{d^2 \log\left(\frac{d\ell(\ell+1)}{\varepsilon\delta}\right)}{\varepsilon_\ell^2} \varepsilon_\ell \\
&\lesssim d^2 \log\left(\frac{dL_T^{(2)}(L_T^{(2)}+1)}{\varepsilon_{L_T^{(2)}}\delta}\right) \sum_{\ell=1}^{L_T^{(2)}} \frac{1}{\varepsilon_\ell} \\
&\lesssim d^2 \left(\log\left(\frac{dL_T^{(2)}}{\delta}\right) + L_T^{(2)} \right) L_T^{(2)} 2^{L_T^{(2)}}.
\end{aligned}
$$

Since

$$
\begin{aligned}
T &\gtrsim \sum_{\ell=1}^{L_T^{(2)}-1} d^2 \log(1/\delta) 4^\ell \\
&\gtrsim d^2 \log(1/\delta) 4^{L_T^{(2)}},
\end{aligned}
$$

we get the upper bound of $L_T^{(2)}$ as

$$
2^{L_T^{(2)}} \lesssim \sqrt{\frac{T}{d^2 \log(1/\delta)}}.
$$

Using the above result, we finally get

$$
\sum_{t=1}^{T} e_2(t) \lesssim d\sqrt{T} \log\left(\frac{T}{d}\right) \log\left(\frac{dT}{\delta}\right) = \widetilde{\mathcal{O}}(d\sqrt{T}).
$$

By setting $\delta = \frac{1}{T^2}$, it completes the proof.

Appendix G. Suboptimality of Cauchy-Schwarz Based Analysis

Using the self-normalized bound of Krishnamurthy et al. (2018); Kim and Paik (2019), one obtains the following bound with probability at least $1-\delta$ for all t:

$$
\|\hat{\theta}_t - \theta^\star\|_{\widehat{\mathbf{V}}_t + \beta \mathbf{I}} \le \sqrt{\beta} + \sqrt{27 d \log\big(1 + 1/d\delta\big) + 54 \log\left(\frac{t}{\delta}\right)}
$$

for $\beta \asymp d\log\left(\frac{t}{\delta}\right)$. All studies Krishnamurthy et al. (2018); Kim and Paik (2019); Choi et al. (2023) used this method. It gives

$$\|\hat{\theta}_t - \theta^\star\|_{\widehat{\mathbf{V}}_t+\beta\mathbf{I}_d} \lesssim \sqrt{d\log(\frac{t}{\delta})}.$$

Using this, they bound the estimation error as

$$|z^\top(\widehat{\theta}_t - \theta^\star)| \lesssim \|z\|_{(\widehat{\mathbf{\Sigma}}_t+\lambda_t\mathbf{I}_d)^{-1}}\frac{1}{\sqrt{t}} \times \sqrt{d\log(\frac{t}{\delta})}.$$

If $z \notin \mathrm{span}(\mathcal{X} - x_1)$, the lower bound result of Theorem 4 implies we may not get finite estimation error bound. For $z \in \mathrm{span}(\mathcal{X} - x_1)$, for sufficiently large t, we have

$$\|z\|_{(\widehat{\mathbf{\Sigma}}_t+\lambda_t\mathbf{I}_d)^{-1}} \asymp 2\|z\|_{\mathbf{\Sigma}_{\mathbf{P}}^{-1}}$$

and get

$$|z^\top(\hat{\theta}_t - \theta^\star)| \lesssim \|z\|_{\mathbf{\Sigma}_{\mathbf{P}}^{-1}}\frac{\sqrt{d\log(\frac{t}{\delta})}}{\sqrt{t}},$$

which has the additional factor $\sqrt{d}$ compared to Theorem 4.

Appendix H. Discussion for Computational Efficiency and Size of Absolute Constants

Compuational Effeiciency of SBE BOSE algorithm (Krishnamurthy et al., 2018) requires to solve minimax quadratic optimization problem every time and it requires heavy $\Omega(T)$ computations. Semiparametric-TS algorithm (Kim and Paik, 2019) requires to calculate **exact** probability of Thompson sampling policy and it requires large sampling every time. Both makes heavy computations (quadratic optimization, gram matrix inverse, sampling) with $\Omega(T)$ times. Our algorithm requires $\mathcal{O}(\log T)$ times of policy updates, so we need to calculate poicy and matrix inverse at most $\mathcal{O}(\log T)$ times. Also, if there is a gap, then $\mathcal{O}(\log(1/\Delta_{\min}))$ policy update is required, which is significantly computational efficient.

Size of Absolute Constants. The absolute constants in our analysis depend critically on the constants from Theorem 3 and Theorem 4. The constant in Theorem 3 is specified as value 2. The constant in Theorem 4, as detailed in its proof, is a product of the constant for second-moment comparability (from Eq. (8)) and constants from the standard Bernstein inequality. The proof is presented in Appendix B. The constants from standard concentration inequalities are well-known and are typically mild.

We now investigate the magnitude of the constant c related to second-moment comparability, which satisfies:

$$\frac{1}{c}(\widehat{\mathbf{V}}_t + \log(t/\delta)\mathbf{I}) \preceq \mathbf{V}_t + \log(t/\delta)\mathbf{I} \preceq c(\widehat{\mathbf{V}}_t + \log(t/\delta)\mathbf{I}) \tag{8}$$

The following lemma shows that this constant c is also mild. In our algorithm, the length of each phase is on the order of $n_\ell \gtrsim \frac{d}{\varepsilon^2} + \frac{d^{3/2}}{\varepsilon}$, so the condition in Lemma 23 is easily satisfied.

Lemma 23 *Under the same setup as Theorem 4, when* $t \gtrsim M\log(d/\delta)$*, the following holds with a probability of at least* $1-\delta/10$*:*

$$\frac{1}{2}\left(\widehat{\mathbf{\Sigma}}_t + \frac{\log(t/\delta)}{t}\mathbf{I}\right) \preceq \mathbf{\Sigma} + \frac{\log(t/\delta)}{t}\mathbf{I} \preceq \frac{3}{2}\left(\widehat{\mathbf{\Sigma}}_t + \frac{\log(t/\delta)}{t}\mathbf{I}\right)$$

Recall that when we use our design `DEO`*, we have* $M = d$*.*

Proof In the proof of Theorem 4, for $\lambda_t = \log(t/\delta)/t$, we showed that with a probability of at least $1-\delta/10$:

$$\left\|(\mathbf{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\left(\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_t\right)(\mathbf{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\right\|_{\text{op}} \lesssim \sqrt{\frac{M\log\left(\frac{10d}{\delta}\right)}{t}} \lesssim \sqrt{\frac{M\log\left(\frac{d}{\delta}\right)}{t}}.$$

Hence, when $t \gtrsim d\log(d/\delta)$, we have:

$$\left\|(\mathbf{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\left(\mathbf{\Sigma}-\widehat{\mathbf{\Sigma}}_t\right)(\mathbf{\Sigma}+\lambda_t\mathbf{I}_d)^{-\frac{1}{2}}\right\|_{\text{op}} \leq \frac{1}{2},$$

which leads to the desired result. ∎

Appendix I. Experiments

We conducted experiments with two shifts $\nu_t^1 = 1+\sin(2t)$ and $\nu_t^2 = \max(\frac{\log(t+1)}{5}, 2)\times(-1)^{q_t}$ where q_t is remainder of t dividing by 3. We first simulate the cumulative regret of `SBE` and compare them with two existing semiparametric bandit algorithms: BOSE from Krishnamurthy et al. (2018) and semiparametric-TS of Kim and Paik (2019). Second, to demonstrate our algorithm is PAC, we examine the maximum estimation error, $\mathbf{e}_t := \max_{i\in[K]} |(x_i - x_1)^\top(\hat{\theta}_t - \theta^\star)|$ over time t and we plot the value of $\sqrt{t}\cdot\mathbf{e}_t$ over time. If $\sqrt{t}\cdot\mathbf{e}_t$ is bounded by some constant, it means that our algorithm has $\sqrt{t}$-rate PAC property.

I.1. Cumulative Regret

For the comparison of cumulative regret, we conducted to cases: $d=5, K=10$ and $d=20, K=30$ with noise level $\sigma=1$. We constructed feature set with suboptimality gap $\Delta_\star = 1/2$. All previous literature Kim and Paik (2019); Krishnamurthy et al. (2018) set one parameter as tuning parameter, so we also regard the exploration paramaters as hyperparameters.

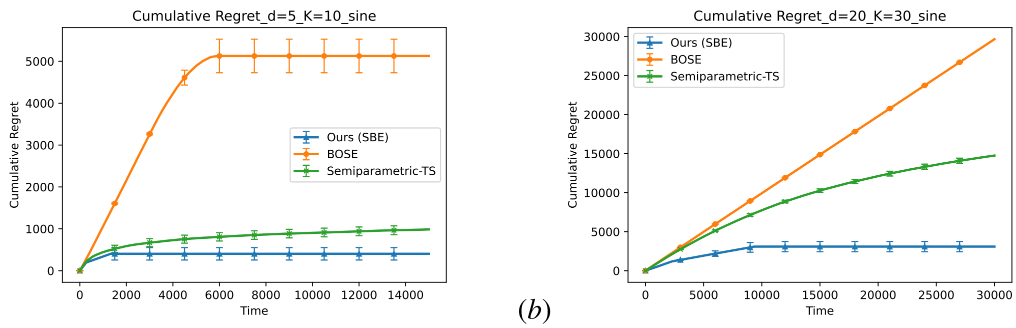
Hyperparameters. For each algorithm, we tuned a single exploration parameter to its theoretical value, up to a constant factor.

For semiparametric-TS algorithm (Kim and Paik, 2019), we set v, which is an exploration parameter for Thompson sampling, as $v = c\sqrt{d\log T}$ for $c = 1,2,4,8$ and we chose the optimal hyperparameter.

For BOSE algorithm, we set ridge regularizer $\lambda = 4d\log(9T) + 8\log(4T/\delta)$ as suggested in their algorithm and we only tune γ, which is a parameter related to exploration level of the algorithm. We set $\gamma = c(\sqrt{\lambda} + \sqrt{d\log(\frac{T}{d}) + \log(T/\delta)})$ for $c = 1,2,4,8$, up to constant with their theoretical value.

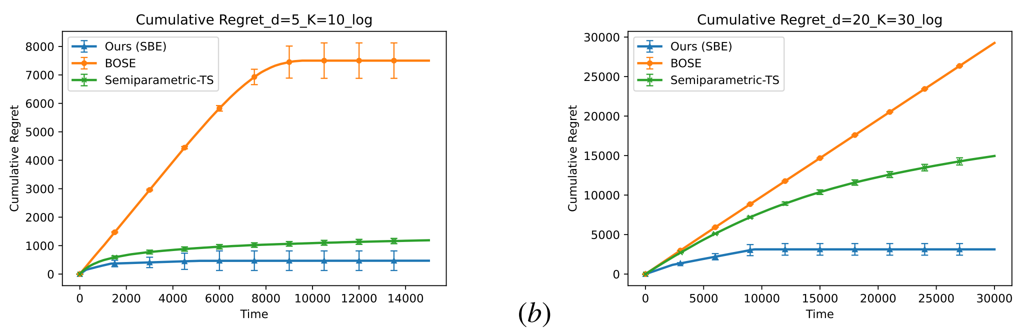
For our algorithm, we set $n_\ell = c \min\left(\lceil \frac{d \log(\frac{dK\ell}{\delta \varepsilon_\ell})}{\varepsilon_\ell^2} + \frac{d^{\frac{3}{2}} \log(\frac{dK\ell}{\delta})}{\varepsilon_\ell} \rceil, \lceil \frac{d^2}{\varepsilon_\ell^2} \log(\frac{d\ell}{\varepsilon_\ell \delta}) \rceil \right)$ for hyperparameter $c \in \{1, 2, 4, 8\}$ and used the best one.

We can see that our algorithm SBE's plot became flat after certain time. It means that our algorithm did a best arm identification. Although the graph shapes for the sine shift and cosine shift appear similar in the case of $d = 20, K = 30$, the actual regret values differ: the cumulative regret of our algorithm for ν_t^1 is 2677.66 but for ν_t^2 is 2703.10.



(a) (b)

Figure 1: Plots (a) and (b) are the cumulative regret plots of SBE with sine shift ν_t^1. We conducted two cases: $d = 5, K = 10$ and $d = 20, K = 30$.



(a) (b)

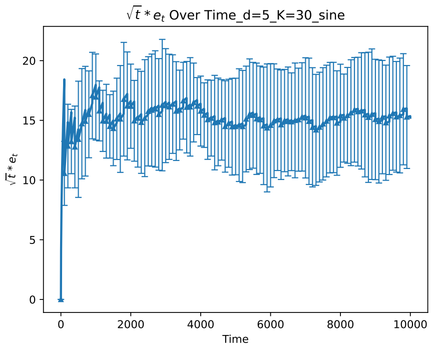
Figure 2: Plots (a) and (b) are the cumulative regret plots of SBE with log shift, ν_t^2. We conducted two cases: $d = 5, K = 10$ and $d = 20, K = 30$.

I.2. Experiments for PAC Property

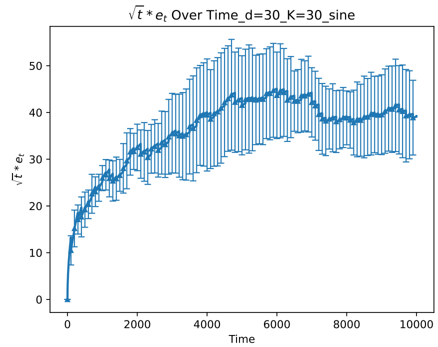
Next we conducted experiment to investigate the maximum estimation error over arms,

$$\mathbf{e}_t := \max_{i \in [K]} |(x_i - x_1)^\top (\hat{\theta}_t - \theta^\star)|$$

for fixed policy of DEO. We plot the $\sqrt{t}\mathbf{e}_t$ over time $t \geq 1$. We run 10 times and plot the error bar. We set same $K = 30$, but change $d = 5$ and $d = 30$ to see the dimension dependency. The theoretical value is $\sqrt{t}\mathbf{e}_t \lesssim \widetilde{\mathcal{O}}(\sqrt{d \log K})$ and we can see the value of $\sqrt{t}\mathbf{e}_t \leq 10\sqrt{d \log K}$ for both cases. Since we calculate every arm $x_i, i \in [K]$'s estimation error, this means that our design DEO performs good pure exploration and matches our theory.

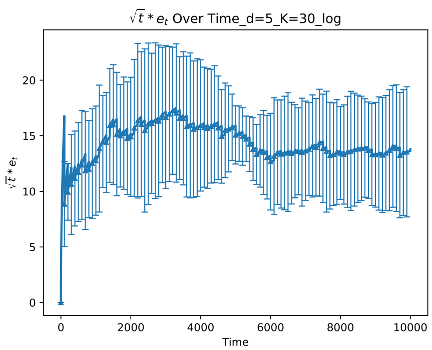


(a)

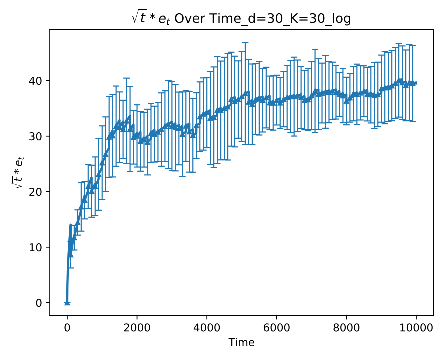


(b)

Figure 3: (a), (b) are plots of $\sqrt{t}\mathbf{e}_t$ for $d = 5, K = 30$ and $d = 30, K = 30$.



(a)



(b)

Figure 4: (a), (b) is plot of $\sqrt{t}\mathbf{e}_t$ for $d = 5, K = 30$ and $d = 30, K = 30$.

Appendix J. Technical Lemmas

Lemma 24 *Assume $z_i \sim \mathcal{P}$ i.i.d. for $1 \leq i \leq t$ and define $\widehat{\mathbf{V}}_t := \sum_{i=1}^t z_i z_i^\top$ and $\mathbf{V}_t = t\mathbb{E}[z_1 z_1^\top]$. If $\|z_i\|_2 \leq R$ for some absolute constant $R > 0$ almost surely, then for $\beta = \log(t/\delta)$, we have*

$$\frac{1}{c}(\widehat{\mathbf{V}}_t + \beta \mathbf{I}_d) \preceq \mathbf{V}_t + \beta \mathbf{I}_d \preceq c(\widehat{\mathbf{V}}_t + \beta \mathbf{I}_d) \tag{9}$$

holds with probability at least $1 - \frac{\delta}{10}$ for some absolute constant $c > 0$.

Proof By applying Lemma 26 directly, for some absolute constant $c_0 > 1$ and $\xi = c_0 \log(\frac{t}{\delta})$, with probability at least $1 - \frac{\delta}{10}$,

$$\frac{1}{2}(\widehat{\mathbf{V}}_t + \xi \mathbf{I}_d) \preceq \mathbf{V}_t + \xi \mathbf{I}_d \preceq 2(\widehat{\mathbf{V}}_t + \xi \mathbf{I}_d)$$

holds. Thus for $\beta = \log(\frac{t}{\delta})$, we have

$$\widehat{\mathbf{V}}_t + \beta \mathbf{I}_d \succeq (\widehat{\mathbf{V}}_t + \xi \mathbf{I}_d)\frac{\beta}{\xi} \succeq \frac{1}{c_0}(\widehat{\mathbf{V}}_t + \xi \mathbf{I}_d) \succeq \frac{1}{2c_0}(\mathbf{V}_t + \xi \mathbf{I}_d) \succeq \frac{1}{2c_0}(\mathbf{V}_t + \beta \mathbf{I}_d).$$

Similarly, we have

$$\widehat{\mathbf{V}}_t + \beta \mathbf{I}_d \preceq 2c_0(\mathbf{V}_t + \xi \mathbf{I}_d).$$

Hence there exists absolute constant $c > 0$ with probability at least $1 - \frac{\delta}{10}$,

$$\frac{1}{c}(\mathbf{V}_t + \beta \mathbf{I}_d) \preceq \widehat{\mathbf{V}}_t + \beta \mathbf{I}_d \preceq c(\mathbf{V}_t + \beta \mathbf{I}_d).$$

■

Remark 25 *We can choose $\beta = c\log(t/\delta)$ for any absolute constant c, but for simplicity, we just choose $\beta = \log(\frac{t}{\delta})$.*

Lemma 26 (Corollary E.1 from Wang 2023) *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. random elements in a separable Hilbert space $\mathbb{H}$ with $\boldsymbol{\Sigma} = \mathbb{E}\left(\boldsymbol{x}_i \otimes \boldsymbol{x}_i\right)$ being trace class. Define $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \otimes \boldsymbol{x}_i$. Choose any constant $\gamma \in (0,1)$ and define an event $\mathcal{A} = \{(1-\gamma)(\boldsymbol{\Sigma}+\lambda \boldsymbol{I}) \preceq \widehat{\boldsymbol{\Sigma}}+\lambda \boldsymbol{I} \preceq (1+\gamma)(\boldsymbol{\Sigma}+\lambda \boldsymbol{I})\}$. 1. If $\|\boldsymbol{x}_i\|_{\mathbb{H}} \leq M$ holds almost surely for some constant M, then there exists a constant $C \geq 1$ determined by γ such that $\mathbb{P}(\mathcal{A}) \geq 1-\delta$ holds so long as $\delta \in (0, 1/14]$ and $\lambda \geq \frac{CM^2\log(n/\delta)}{n}$.*

Lemma 27 (Lemma E.3 from Wang 2023) *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. random elements in a separable Hilbert space $\mathbb{H}$ with $\boldsymbol{\Sigma} = \mathbb{E}\left(\boldsymbol{x}_i \otimes \boldsymbol{x}_i\right)$ being trace class. Define $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \otimes \boldsymbol{x}_i$. 1. If $\|\boldsymbol{x}_i\|_{\mathbb{H}} \leq S$ holds almost surely for some constant S, then for any $v^2 \geq \|\boldsymbol{\Sigma}\|$ and $r \geq \mathrm{Tr}(\boldsymbol{\Sigma})/v^2$,*

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} \leq \sqrt{\frac{8S^2v^2\log(r/\delta)}{n}} + \frac{6S^2\log(r/\delta)}{n}\right) \geq 1-\delta, \quad \forall \delta \in (0, r/14]$$