Nearly Optimal Best Arm Identification for Semiparametric Bandits

Seok-Jin Kim
Columbia University

Abstract

We study fixed-confidence Best Arm Identification (BAI) in semiparametric bandits, where rewards are linear in arm features plus an unknown additive baseline shift. Unlike linear-bandit BAI, this setting requires orthogonalized regression, and its instance-optimal sample complexity has remained open. For the transductive setting, we establish an attainable instance-dependent lower bound characterized by the corresponding linear-bandit complexity on shifted features. We then propose a computationally efficient phase-elimination algorithm based on a new $\mathcal{XY}$-design for orthogonalized regression. Our analysis yields a nearly optimal high-probability sample-complexity upper bound, up to log factors and an additive d^2 term, and experiments on synthetic instances and the Jester dataset show clear gains over prior baselines.

1 INTRODUCTION

In the multi-armed bandit (MAB) framework, a learner sequentially selects actions and receives rewards from unknown distributions (Lattimore and Szepesvári, 2019). Beyond regret minimization, a major line of work studies *Best Arm Identification* (BAI) (Soare et al., 2014; Kaufmann et al., 2016; Jamieson and Nowak, 2014), where the objective is to identify the optimal action. The two standard formulations are the fixed-confidence setting (Fiez et al., 2019; Soare et al., 2014) and the fixed-budget setting (Jedra and Proutiere, 2020; Yang and Tan, 2022). We focus on the *fixed-confidence* setting: given a risk parameter δ, the learner must return the best arm with probability at least $1-\delta$. Performance is measured by the resulting *sample complexity*.

Classical MAB models treat each arm as an unrelated option (Auer et al., 2002; Langford and Zhang, 2007). In many applications, however, arms are associated with feature vectors and rewards are structured through those features. When the mean reward is linear in the feature vector, one obtains the *linear bandit* model, which has been studied extensively (Goldenshluger and Zeevi, 2013; Abbasi-Yadkori et al., 2011; Li et al., 2019; Rusmevichientong and Tsitsiklis, 2010).

Purely linear models are often too restrictive for practice. In mobile health, for example, a user's baseline health state may vary in complex and non-stationary ways that are unrelated to the treatment effect (Greenewald et al., 2017). In such treatment-regime problems, each arm corresponds to a treatment, and "do nothing" often acts as a reference arm. The reward of that reference arm is a nuisance component that may be highly complex, while the *treatment effect* remains comparatively simple. Motivated by this setting, Greenewald et al. (2017) studied a model in which the baseline reward is arbitrary but the treatment effect, namely the difference between the reward of treatment x_i and the reference x_1, follows a linear parametric model. Related semistructured models, where the nuisance component is complex but the effect of interest is simple, also appear in treatment-effect estimation (Kennedy, 2023; Kennedy et al., 2024) and sequential decision-making (Wen et al., 2025).

This idea is captured by the **semiparametric reward model** of Krishnamurthy et al. (2018). The model has been studied in contextual bandits (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim and Paik, 2019) and, more recently, in the fixed-feature setting by Kim et al. (2025). Formally, the reward r_t at time t for action a_t with feature vector $x_{a_t} \in \mathbb{R}^d$ is

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

where $\theta^\star \in \mathbb{R}^d$ is an unknown parameter vector, ν_t is an arbitrary baseline shift chosen before action selection, and η_t is random noise. This extends the linear bandit model (the special case $\nu_t = 0$) by allowing

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

complex baseline dynamics such as those arising in recommender systems (Kim and Paik, 2019) and clinical trials. Because BAI is typically carried out over a fixed set of candidate actions, robust experimental design under such baseline shifts is essential. Kim et al. (2025) referred to this fixed-feature setting as **semiparametric bandits** and established a $\tilde{\mathcal{O}}(\sqrt{dT})$ regret bound together with the first BAI results for this setting.

Motivating Examples. The baseline-shift model naturally captures various real-world scenarios:

- *Clinical Trials:* Trials often involve fixed treatments administered to varying subjects. Here, ν_t represents the complex, non-stationary baseline response of the t-th subject (e.g., the response under "do nothing"), while the treatment effect remains stable across the population (Greenewald et al., 2017).
- *Ad Selection:* In landing page optimization without personal data, ν_t can represent the baseline clicking propensity of a user arriving at time t, reflecting external trends or user heterogeneity (Kim et al., 2025).
- *Human Ratings (Jester):* In the Jester joke rating dataset (Goldberg et al., 2001), users exhibit distinct tendencies to rate all items high or low, regardless of the specific joke. The term ν_t captures this evaluator-specific baseline, while the intrinsic joke quality remains parametric. We return to this example in Section 7.

Transductive Fixed-Confidence BAI. We study BAI in the more general *transductive* setting (Fiez et al., 2019), where data collection and final recommendation need not share the same action set. The learner can sample from a *source* feature set $\mathcal{X} = \{x_1, \dots, x_K\} \subset \mathbb{R}^d$, but must identify the best arm in a possibly different *target* feature set $\mathcal{Z} = \{z_1, \dots, z_H\} \subset \mathbb{R}^d$. After collecting data from $\mathcal{X}$, the learner recommends $\hat{z} \in \mathcal{Z}$, where the true target optimum $z^\star = \arg\max_{z \in \mathcal{Z}} z^\top \theta^\star$ is assumed unique. This formulation models settings in which experimentation is constrained to a design set, while the downstream decision is taken over a broader or different evaluation set. For example, in drug development many compounds and dosages may be testable in the lab ($\mathcal{X}$), but only a subset is approved for deployment ($\mathcal{Z}$) (Fiez et al., 2019). The usual non-transductive setting is recovered by taking $\mathcal{Z} = \mathcal{X}$, so all of our results immediately specialize to the standard case.

1.1 Research Questions

We study fixed-confidence transductive BAI for semiparametric bandits. While Kim et al. (2025) initiated the study of experimental design in this setting, it remains unclear whether their sample-complexity guarantees are optimal. More fundamentally, instance-dependent lower bounds for semiparametric BAI were not previously known. This leads to two central questions:

Q1. What is the attainable lower bound on the sample complexity for fixed-confidence BAI in semiparametric bandits?

Q2. Can we design an algorithm that matches this lower bound?

In linear bandits, instance-optimal BAI relies on $\mathcal{XY}$-designs rather than G-optimal designs (Fiez et al., 2019). Semiparametric bandits, however, require **orthogonalized regression** to obtain consistent estimates; see Section 2.3. Although G-optimal design for orthogonalized regression is known (Kim et al., 2025), the corresponding $\mathcal{XY}$-design problem has not been developed. Answering Q1 and Q2 therefore requires us to solve two challenges: (C1) identify an attainable instance-dependent lower bound, and (C2) construct an effective $\mathcal{XY}$-design for orthogonalized regression.

1.2 Result Overview and Contribution

Our main contributions are as follows:

- **Lower bound:** We prove the first attainable instance-dependent lower bound for BAI in semiparametric bandits.
- **Algorithm:** We propose an efficient phase-elimination algorithm for transductive semiparametric BAI.
- **Sample-complexity guarantee:** We prove a high-probability upper bound controlled by the shifted linear-bandit benchmark up to logarithmic factors and an additive d^2 term. The same benchmark also appears in our lower bound, and the guarantee immediately extends to the standard non-transductive case $\mathcal{Z} = \mathcal{X}$.
- **$\mathcal{XY}$-design for orthogonalized regression:** We introduce a new design construction that controls contrast variances for orthogonalized regression, which is the key technical ingredient behind our upper bound.

Comparison with Prior Work. To compare with prior semiparametric BAI results, we specialize here to the standard non-transductive setting $\mathcal{Z} = \mathcal{X}$ and identify the common arms by $z_i = x_i$ for $i \in [K]$. Define $q_i := z_i^\top \theta^\star$, let $q^\star := \max_{i\in[K]} q_i$, and let $q_{(1)} \geq q_{(2)} \geq \cdots$ denote the sorted expected rewards. Define the gaps $\Delta_j := q^\star - q_{(j)}$, so $0 = \Delta_1 \leq \Delta_2 \leq \cdots \leq \Delta_K$ and Δ_2 is the smallest nonzero gap. In this non-transductive setting, prior semiparametric BAI algorithms such as SBE and G-Opt (Kim et al., 2025) guarantee sample complexity $\tilde{\mathcal{O}}(d/\Delta_2^2)$. By contrast, our algorithm scales as $\tilde{\mathcal{O}}(\tau^*_{\mathrm{lin}}(\mathcal{X} - x_1))$, where $\tau^*_{\mathrm{lin}}(\mathcal{X} - x_1)$ is the optimal fixed-confidence sample complexity of the corresponding linear bandit instance with shifted feature set $\mathcal{X} - x_1$. Since $\tau^*_{\mathrm{lin}}(\mathcal{X} - x_1) \lesssim d/\Delta_2^2$ always holds and can be much smaller on favorable instances, our guarantee is never worse and can be substantially tighter. Appendix A gives a broader comparison with related semiparametric, linear-bandit, and corruption-robust literatures.

1.3 Notations

We write $[n] := \{1, 2, \ldots, n\}$ for a positive integer n and denote the n-dimensional simplex by $\Delta^{(n)}$. For a vector $x \in \mathbb{R}^d$, $\|x\|_p$ denotes the ℓ_p norm. For any positive semidefinite matrix $\mathbf{A}$, we define the weighted norm $\|x\|_{\mathbf{A}} = \sqrt{x^\top \mathbf{A} x}$. We use $\mathcal{O}(\cdot)$ or $\lesssim$ to suppress absolute constants, and $\tilde{\mathcal{O}}(\cdot)$ to further suppress logarithmic factors in the natural problem parameters when this causes no confusion. The notation $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$. *Constants $c, C, c_1, \ldots$ may change from line to line.* For sets $A, B \subset \mathbb{R}^n$ and a vector $a \in \mathbb{R}^n$, we define $A - a := \{a' - a \mid a' \in A\}$ and $A - B := \{a - b \mid a \in A, b \in B\}$.

Following Kim et al. (2025), we generalize the matrix-inverse-weighted norm $\|x\|_{\mathbf{A}^{-1}}$ to positive semidefinite matrices $\mathbf{A}$ (regardless of invertibility) as:

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

This definition coincides with the standard definition when $\mathbf{A}$ is full rank and remains well-defined and finite if x lies in the column space of $\mathbf{A}$. If x does not lie in the column space of $\mathbf{A}$, then $\|x\|_{\mathbf{A}^{-1}} = +\infty$ under this convention.

2 SETUP AND PRELIMINARIES

2.1 Fixed-confidence BAI

For a confidence level $\delta \in (0, 1)$, a fixed-confidence algorithm sequentially selects source arms, stops at a random time $\boldsymbol{\tau}(\delta)$, and outputs an estimate $\hat{z} \in \mathcal{Z}$ of the best target arm $z^\star$. We require the algorithm to be δ-correct:

$$\mathbb{P}[\hat{z} \neq z^\star] \leq \delta.$$

When this condition holds, we refer to the stopping time $\boldsymbol{\tau}(\delta)$ as the *sample complexity*.

Lower bound for transductive linear bandits. Consider a linear bandit problem with a *source* feature set $\mathcal{X}$ and a *target* feature set $\mathcal{Z}$. Let $z^\star$ be the unique maximizer of $z^\top \theta^\star$ over $z \in \mathcal{Z}$. The optimal sample complexity for transductive BAI admits the following instance-dependent lower bound (Fiez et al., 2019):

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X}) := \min_{\mathbf{p}\in\Delta^{(K)}} \max_{z\in\mathcal{Z}\setminus\{z^\star\}} \frac{\|z^\star - z\|^2_{(\sum_{i=1}^K p_i x_i x_i^\top)^{-1}}}{|(z^\star - z)^\top \theta^\star|^2}. \tag{1}$$

The numerator captures the statistical resolution required to estimate the contrast $(z^\star - z)^\top \theta^\star$ under design $\mathbf{p}$. Thus (1) has the familiar "resolution versus gap" form: to certify that $z^\star$ beats a competitor z, the learner must estimate the corresponding contrast to an accuracy commensurate with the gap size. The design $\mathbf{p}$ controls the contrast variance through the inverse information matrix $(\sum_i p_i x_i x_i^\top)^{-1}$. Under the generalized inverse-norm convention above, this benchmark also captures identifiability. In particular, if some relevant contrast $z^\star - z$ lies outside $\mathrm{span}(\mathcal{X})$, then the numerator is $+\infty$ for every design $\mathbf{p}$, and hence $\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X}) = +\infty$.

As a special case, setting $\mathcal{Z} = \mathcal{X}$ recovers the standard (non-transductive) linear-bandit BAI lower bound:

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X}) := \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} : \mathcal{X}) \tag{2}$$

$$= \min_{\mathbf{p}\in\Delta^{(K)}} \max_{x\in\mathcal{X}\setminus\{x^\star\}} \frac{\|x^\star - x\|^2_{(\sum_{i=1}^K p_i x_i x_i^\top)^{-1}}}{|(x^\star - x)^\top \theta^\star|^2}. \tag{3}$$

Later, for the shifted non-transductive benchmark, we use the shorthand

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} - x_1) := \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} : \mathcal{X} - x_1).$$

Attainability via $\mathcal{XY}$-design. To achieve nearly optimal sample complexity in linear-bandit BAI, one uses $\mathcal{XY}$-design rather than classical G-optimal design (Fiez et al., 2019). For an active target set $\mathcal{A} \subseteq \mathbb{R}^d$ (typically $\mathcal{A} \subset \mathcal{Z}$ during elimination), the $\mathcal{XY}$-design problem is

$$\min_{\mathbf{p}\in\Delta^{(K)}} \max_{u,u'\in\mathcal{A}} (u - u')^\top \Big(\sum_{i=1}^K p_i x_i x_i^\top\Big)^{-1} (u - u').$$

Algorithms based on this design achieve nearly optimal complexity (Fiez et al., 2019).

The reason is that G-optimal design minimizes $\max_{x\in\mathcal{X}} x^\top (\sum_i p_i x_i x_i^\top)^{-1} x$, which is appropriate for

prediction over $\mathcal{X}$. In BAI, the relevant quantities are instead the *pairwise contrasts* within the surviving set $\mathcal{A} \subseteq \mathcal{Z}$. $\mathcal{XY}$-designs are tailored precisely to control those contrast variances. This distinction becomes even more important in semiparametric bandits, where estimation is based on orthogonalized, centered features.

2.2 Our Setup: Semiparametric Bandits

The semiparametric bandit model with fixed source features was formalized by Kim et al. (2025). Let $\mathcal{X} = \{x_1, \dots, x_K\} \subset \mathbb{R}^d$ denote the *source* feature set. At each time $t = 1, 2, \dots$, the learner selects a source arm $a_t \in [K]$ and observes

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

where $\nu_t \in \mathbb{R}$ is an arbitrary bounded shift and η_t is independent sub-Gaussian noise with proxy 1. Let $\mathcal{H}_{t-1}$ be the history sigma-algebra generated by $\{a_1, r_1, \dots, a_{t-1}, r_{t-1}\}$. We assume that ν_t is $\mathcal{H}_{t-1}$-measurable, so the baseline may be fully adaptive to the past.

In the transductive BAI problem, we are also provided with a target feature set $\mathcal{Z} = \{z_1, \dots, z_H\} \subset \mathbb{R}^d$. The objective is to identify the unique best arm $z^\star = \arg\max_{z \in \mathcal{Z}} z^\top \theta^\star$ with probability at least $1 - \delta$.

Following prior work (Kim et al., 2025; Krishnamurthy et al., 2018; Kim and Paik, 2019; Choi et al., 2023), we impose the following boundedness assumption:

Assumption 1 (Boundedness)**.** *For all $t \geq 1$, $i \in [K]$, and $h \in [H]$, we assume $|\nu_t| \leq 1$, $\|x_i\|_2 \leq 1$, and $\|z_h\|_2 \leq 1$. We also assume $\|\theta^\star\|_2 \leq 1$.*

The term ν_t represents a complex baseline or drift that is *not* modeled parametrically. Because it may adapt to past observations, standard least-squares regression on (x_{a_t}, r_t) is not appropriate. We therefore use **orthogonalized regression**, which mitigates the effect of ν_t by regressing on centered features $x_{a_t} - \bar{x}_{\mathbf{p}}$.

2.3 Orthogonalized Regression and Policy Design

Orthogonalized regression was developed precisely for the semiparametric reward model (Krishnamurthy et al., 2018; Kim and Paik, 2019; Kim et al., 2025). Given data $\{x_{a_s}, r_s\}_{s=1}^t$ generated under a fixed policy $\mathbf{p} = (p_1, \dots, p_K) \in \Delta^{(K)}$, define the policy mean $\bar{x}_{\mathbf{p}} := \sum_{i=1}^K p_i x_i$. The estimator is the ridge-regression fit obtained from centered covariates $x_{a_s} - \bar{x}_{\mathbf{p}}$:

$$(\beta \mathbf{I}_d + \sum_{s=1}^t (x_{a_s} - \bar{x}_{\mathbf{p}})(x_{a_s} - \bar{x}_{\mathbf{p}})^\top)^{-1} \sum_{s=1}^t (x_{a_s} - \bar{x}_{\mathbf{p}}) r_s.$$

To see why the baseline does not bias the estimator, fix a policy $\mathbf{p}$ and write

$$\tilde{x}_{a_s} := x_{a_s} - \bar{x}_{\mathbf{p}}, \qquad \widehat{\mathbf{V}}_t := \sum_{s=1}^t \tilde{x}_{a_s} \tilde{x}_{a_s}^\top.$$

Since $x_{a_s} = \tilde{x}_{a_s} + \bar{x}_{\mathbf{p}}$, the estimation error admits the decomposition

$$\hat{\theta}_t - \theta^\star = (\widehat{\mathbf{V}}_t + \beta \mathbf{I}_d)^{-1} \sum_{s=1}^t \tilde{x}_{a_s} \big(\bar{x}_{\mathbf{p}}^\top \theta^\star + \nu_s + \eta_s\big) - \beta (\widehat{\mathbf{V}}_t + \beta \mathbf{I}_d)^{-1} \theta^\star.$$

Under the fixed policy $\mathbf{p}$, we have $\mathbb{E}[\tilde{x}_{a_s} \mid \mathcal{H}_{s-1}] = 0$. Therefore

$$\mathbb{E}\big[\tilde{x}_{a_s} \big(\bar{x}_{\mathbf{p}}^\top \theta^\star + \nu_s\big) \mid \mathcal{H}_{s-1}\big] = 0,$$

because $\bar{x}_{\mathbf{p}}^\top \theta^\star + \nu_s$ is $\mathcal{H}_{s-1}$-measurable. In particular, the nuisance shift ν_s cancels from the estimating equation in conditional expectation. The remaining term $\sum_{s=1}^t \tilde{x}_{a_s} \eta_s$ is the usual mean-zero noise term, while the final term is the regularization bias. This is the basic reason orthogonalized regression remains consistent even when the baseline shift is fully adaptive to the past.

Concentration Bounds. Kim et al. (2025) established sharp concentration bounds for this estimator. Define the covariance of a policy $\mathbf{p}$ by

$$\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}} := \sum_{i=1}^K p_i (x_i - \bar{x}_{\mathbf{p}})(x_i - \bar{x}_{\mathbf{p}})^\top,$$
$$\text{where } \bar{x}_{\mathbf{p}} := \sum_{i=1}^K p_i x_i.$$

Suppose t samples are collected under policy $\mathbf{p}$. If the regularizer satisfies $\beta \asymp \log(t/\delta)$, and for a vector y of interest there exist constants $L, M > 0$ such that

$$y^\top \mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}}^{-1} y \leq L, \qquad \max_{i \in [K]} \|x_i - \bar{x}_{\mathbf{p}}\|^2_{\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}}^{-1}} \leq M,$$

then

$$|y^\top (\hat{\theta}_t - \theta^\star)| \leq c_1 \frac{\sqrt{L \log(t/\delta)}}{\sqrt{t}} + c_2 \frac{M \sqrt{L} \log(d/\delta)}{t}, \tag{4}$$

for absolute constants $c_1, c_2 > 1$. Consequently, to resolve a gap Δ, it suffices to take

$$n \gtrsim R_1 \frac{L}{\Delta^2} \log\left(\frac{L}{\Delta \delta}\right) + R_2 \frac{M \sqrt{L}}{\Delta} \log\left(\frac{d}{\delta}\right), \tag{5}$$

to guarantee $|y^\top (\hat{\theta}_t - \theta^\star)| \leq \Delta$ for some absolute constants $R_1, R_2 > 0$. For a finite family of contrasts,

the same bound holds uniformly after a union bound. The proof-level values of these constants are in fact moderate (well below 10^2), but as in many modern ML and bandit works, we treat their practical counterparts as tunable hyperparameters.

The dominant term in (5) scales with L/Δ^2, where L is the inverse-covariance norm of the relevant contrast under policy $\mathbf{p}$. The second term, involving the source-only stability quantity M, is lower-order in t.

Approximate G-optimal design. Since the leading term in (5) is governed by L, a natural design objective is

$$\max_{x\in\mathcal{X}} x^\top \mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}}^{-1} x.$$

Kim et al. (2025) provided a procedure to find a policy $\mathbf{p}_G$ satisfying:

$$\max_{x\in\mathcal{X}} x^\top \mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_G}^{-1} x \le 4d. \tag{6}$$

While $\mathbf{p}_G$ controls prediction error uniformly over $\mathcal{X}$, fixed-confidence BAI requires accurate estimation of *pairwise* target contrasts. Attaining instance-optimal sample complexity therefore calls for an $\mathcal{XY}$-style objective adapted to orthogonalized regression. This is nontrivial because $\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}}$ depends on $\mathbf{p}$ through both the weights and the centering term $\bar{x}_{\mathbf{p}}$.

Challenge: $\mathcal{XY}$-design for orthogonalized regression. Introducing an active target set $\mathcal{A}\subseteq\mathcal{Z}$, the optimal design objective becomes:

$$\max_{u,u'\in\mathcal{A}} (u-u')^\top \mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}}^{-1}(u-u').$$

This quantity is exactly the relevant L term in Eq. (4) when the vectors of interest are pairwise differences $u-u'$. The difficulty is that the optimization problem is non-convex, so even understanding its optimum is nontrivial. Our subsequent algorithmic development gives a constant-factor construction that is sufficient for the final high-probability upper bound.

3 LOWER BOUNDS

We now establish a lower bound on the sample complexity of fixed-confidence transductive BAI in semiparametric bandits. We begin with the second-moment matrix induced by shifted source features.

Definition 1. *For a policy $\mathbf{p}=(p_1,\dots,p_K)\in\Delta^{(K)}$, we define the second moment of the shifted features $\mathcal{X}-x_1$ as:*

$$\mathbf{\Sigma}_{-1,\mathbf{p}} = \sum_{i=1}^{K} p_i (x_i-x_1)(x_i-x_1)^\top. \tag{7}$$

Similarly, for any $j\in[K]$, let $\mathbf{\Sigma}_{-j,\mathbf{p}}$ denote the second moment matrix for the features shifted by x_j (i.e., $\mathcal{X}-x_j$).

For the corresponding linear bandit problem with source features $\mathcal{X}-x_1$ and target set $\mathcal{Z}$, the optimal instance-dependent sample complexity for transductive BAI is (Fiez et al., 2019)

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_1) = \min_{\mathbf{p}\in\Delta^{(K)}} \max_{z\in\mathcal{Z}\setminus\{z^\star\}} \frac{\|z^\star-z\|^2_{\mathbf{\Sigma}_{-1,\mathbf{p}}^{-1}}}{|(z^\star-z)^\top\theta^\star|^2}. \tag{8}$$

The next proposition shows that this shifted linear benchmark is unavoidable in the worst case over admissible baseline-shift sequences.

Proposition 1 (Lower bound for transductive BAI)**.** *Under Assumption 1, consider any algorithm for fixed-confidence transductive BAI in semiparametric bandits with source features $\mathcal{X}$ and target features $\mathcal{Z}$, required to be δ-correct for every admissible shift sequence $\{\nu_t\}$. Then there exists an admissible deterministic shift sequence for which the expected stopping time $\boldsymbol{\tau}(\delta)$ satisfies:*

$$\mathbb{E}[\boldsymbol{\tau}(\delta)] \gtrsim \log\left(\frac{1}{\delta}\right) \times \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_1).$$

Discussion. Proposition 1 identifies the fundamental worst-case benchmark for semiparametric BAI over admissible baseline-shift sequences. Two questions remain: whether the bound is attainable, and whether the specific anchor arm x_1 matters. The next result addresses the second issue by showing that different anchors change the benchmark by at most a constant factor.

Proposition 2 (Compatibility of lower bounds)**.** *For any indices $i,j\in[K]$, the following inequality holds:*

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_i) \le 4\cdot\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_j).$$

Discussion. Although (8) is written with anchor x_1, Proposition 2 shows that the benchmark is effectively *anchor-invariant*. Conceptually, this means that the hardness of the problem does not depend on an arbitrary reference arm. Algorithmically, it lets us choose whichever anchor is most convenient for design and analysis.

Since the standard setting $\mathcal{Z}=\mathcal{X}$ is a special case, we immediately obtain the following corollary.

Corollary 1 (Lower bound: Non-transductive case)**.** *By setting $\mathcal{Z}=\mathcal{X}$ in Proposition 1, we recover the corresponding worst-case lower bound for the standard*

(non-transductive) semiparametric BAI problem:

$$\mathbb{E}[\boldsymbol{\tau}(\delta)] \gtrsim \log\left(\frac{1}{\delta}\right) \times \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} : \mathcal{X} - x_1)$$
$$= \log\left(\frac{1}{\delta}\right) \times \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} - x_1).$$

4 ALGORITHM

We now present our algorithm for transductive BAI in semiparametric bandits. It follows the standard phase-elimination template from linear bandits (Fiez et al., 2019), but replaces the linear-bandit design step with a new $\mathcal{XY}$-design tailored to orthogonalized regression.

Roadmap. The key technical ingredient is a new $\mathcal{XY}$-design for orthogonalized regression, so we present it first. It controls $\max_{u,u' \in \mathcal{A}} \|u-u'\|^2_{\boldsymbol{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}}}$, the leading variance term governing pairwise contrasts in orthogonalized regression. Once this contrast-focused design is in place, the rest of the algorithm follows the usual phase-elimination template, with an additional mixture with a semiparametric analogue of G-optimal design used only to control the lower-order source-stability term in (5).

4.1 $\mathcal{XY}$-Design for Orthogonalized Regression

Fix an active target set $\mathcal{A} \subset \mathbb{R}^d$ (typically $\mathcal{A} \subseteq \mathcal{Z}$). In analogy with linear-bandit $\mathcal{XY}$-design, we define the semiparametric design objective as follows.

Problem 1 ($\mathcal{XY}$-design for orthogonalized regression)**.** *Define the maximum pairwise variance over the set $\mathcal{A}$ under policy $\mathbf{p}$ as:*

$$\mathcal{V}_{\mathrm{cov}}(\mathcal{A} : \mathcal{X}, \mathbf{p}) := \max_{u,u' \in \mathcal{A}} \|u - u'\|^2_{\boldsymbol{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}}}.$$

Our objective is to find a policy that minimizes this quantity:

$$\mathcal{V}^\star_{\mathrm{cov}}(\mathcal{A} : \mathcal{X}) := \min_{\mathbf{p} \in \Delta^{(K)}} \mathcal{V}_{\mathrm{cov}}(\mathcal{A} : \mathcal{X}, \mathbf{p}).$$

This problem is non-convex because $\boldsymbol{\Sigma}_{\mathrm{cov},\mathbf{p}}$ depends on $\mathbf{p}$ through both the weights and the centering term $\bar{x}_{\mathbf{p}}$, leading to a cubic dependence on $(p_1, \ldots, p_K)$. Fortunately, exact optimization is unnecessary for our purposes. A constant-factor approximation with respect to the right linear-bandit benchmark is enough for our final high-probability upper bound.

Discussion: Sufficiency of Approximation. By Proposition 1, the lower-bound benchmark is expressed through the *linear* transductive BAI problem on shifted features $\mathcal{X} - x_1$. Therefore, it is enough to build a semiparametric policy whose contrast variances are within a constant factor of the optimal linear $\mathcal{XY}$-design value on this shifted instance. This naturally suggests a reduction: compute a linear $\mathcal{XY}$-design on $\mathcal{X} - x_1$ and convert it into a valid semiparametric policy.

Linear $\mathcal{XY}$-design benchmark. For the corresponding linear instance on shifted features, define

$$\mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A} : \mathcal{X} - x_1) := \min_{\mathbf{p} \in \Delta^{(K)}} \max_{u,u' \in \mathcal{A}} \|u - u'\|^2_{\boldsymbol{\Sigma}^{-1}_{-1,\mathbf{p}}}.$$

Proposed Method for $\mathcal{XY}$-Design. The construction is simple:

1. Compute an optimal linear-bandit $\mathcal{XY}$-design for the shifted source features $\mathcal{X} - x_1$ and active target set $\mathcal{A}$. Denote the resulting policy by $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_K)$; necessarily $\tilde{p}_1 = 0$ because $x_1 - x_1 = 0$ carries no information.

2. Form the semiparametric policy $\mathbf{p}_{\mathrm{xor}}$ by mixing $\tilde{\mathbf{p}}$ with the anchor arm:

$$\mathbf{p}_{\mathrm{xor}} = \left(\frac{1}{2}, \frac{1}{2}\tilde{p}_2, \ldots, \frac{1}{2}\tilde{p}_K\right).$$

Since $\tilde{p}_1 = 0$, $\mathbf{p}_{\mathrm{xor}}$ is a valid distribution. The pseudocode appears in Algorithm 1.

Algorithm 1 `XOR`: $\mathcal{XY}$-Design for **O**rthogonalized **R**egression

Require: Target set $\mathcal{A} \subset \mathbb{R}^d$, Source feature set $\mathcal{X}$.
1: Choose an anchor source arm, denoted by x_1.
2: Compute the linear bandit $\mathcal{XY}$-design policy $(\tilde{p}_1, \ldots, \tilde{p}_K)$ for the active set $\mathcal{A}$ using source features $\mathcal{X} - x_1$, and set $\tilde{p}_1 = 0$.
3: **Return** the policy $\mathbf{p}_{\mathrm{xor}}(\mathcal{A}) = \left(\frac{1}{2}, \frac{\tilde{p}_2}{2}, \ldots, \frac{\tilde{p}_K}{2}\right)$.

Proposition 3 (Performance of `XOR` design)**.** *The design $\mathbf{p}_{\mathrm{xor}}$ returned by Algorithm 1 satisfies:*

$$\mathcal{V}_{\mathrm{cov}}(\mathcal{A} : \mathcal{X}, \mathbf{p}_{\mathrm{xor}}(\mathcal{A})) \le 4 \cdot \mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A} : \mathcal{X} - x_1).$$

Implications of Proposition 3. Proposition 3 is the key design result. Although Problem 1 is non-convex, `XOR` achieves a constant-factor approximation to the *optimal linear* $\mathcal{XY}$-design on the shifted instance. Because our lower bound is stated in terms of exactly that benchmark, this approximation is sufficient for the final upper bound to track the benchmark up to logarithmic factors.

4.2 Algorithm for Fixed-Confidence Transductive BAI

We now describe the full elimination algorithm. In phase ℓ, with active target set $\mathcal{A}_\ell \subseteq \mathcal{Z}$, we sample n_ℓ source arms according to the mixture policy

$$\mathbf{p}_\ell = \frac{1}{2}\big(\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell) + \mathbf{p}_G\big), \tag{9}$$

where $\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)$ comes from Algorithm 1 and $\mathbf{p}_G$ is the approximate G-optimal design satisfying (6). We set the confidence radius to $\varepsilon_\ell = 2^{-\ell}$ and define

$$\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell) = \max_{u,u' \in \mathcal{A}_\ell} \|u - u'\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_\ell}}.$$

Rationale for the Mixture Policy. The mixture $\mathbf{p}_\ell$ is designed to control both terms in the orthogonalized-regression concentration bound (4). The $\mathcal{XY}$-component $\mathbf{p}_{\mathrm{xor}}$ controls contrast variances within the active set and therefore the leading $1/\Delta^2$ contribution (the L-term in (4)). The G-optimal component $\mathbf{p}_G$ controls prediction variance over the source set and stabilizes the lower-order source-stability contribution (the M-term). Combining them yields contrast efficiency without sacrificing uniform stability.

Based on $\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell)$, we determine the phase length n_ℓ as:

$$n_\ell = \left\lceil R_1 \frac{\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell)}{\varepsilon_\ell^2} \log\!\left(\frac{\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell)}{\varepsilon_\ell \delta_\ell}\right) + R_2 \frac{32d\sqrt{\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell)}}{\varepsilon_\ell} \log\!\left(\frac{d}{\delta_\ell}\right) \right\rceil, \tag{10}$$

where $\varepsilon_\ell = 2^{-\ell}$ and $\delta_\ell = \frac{\delta}{|\mathcal{A}_\ell|^2 \ell(\ell+1)}$. The constants R_1, R_2 correspond to those in the concentration bound (5). In each phase, the algorithm collects n_ℓ samples under $\mathbf{p}_\ell$, fits an orthogonalized-regression estimator $\hat{\theta}_\ell$, eliminates targets whose estimated gap from the empirical best exceeds ε_ℓ, and then repeats on the surviving set. The full pseudocode appears in Algorithm 2.

Computational Efficiency. Each phase of Algorithm 2 requires two convex optimization subroutines: a linear $\mathcal{XY}$-design on the shifted instance (to obtain $\mathbf{p}_{\mathrm{xor}}$) and an approximate G-optimal design (to obtain $\mathbf{p}_G$). Both can be solved efficiently, for example by Frank-Wolfe or multiplicative-weights methods, so the overall procedure is computationally practical.

5 UPPER BOUNDS AND NEAR-OPTIMALITY

We now state the sample-complexity guarantee for Algorithm 2.

Algorithm 2 `SP-BAI`: Fixed-Confidence **BAI** for **S**emi**P**arametric Bandits

Input: Source features $\mathcal{X} = \{x_1, \dots, x_K\}$, Target features $\mathcal{Z} = \{z_1, \dots, z_H\}$, Confidence $\delta \in (0,1)$. Hyperparameters $R_1, R_2 > 0$.
Initialize: Active set $\mathcal{A}_1 = \mathcal{Z}$, phase index $\ell = 1$.
while $|\mathcal{A}_\ell| > 1$ **do**
  Set confidence radius $\varepsilon_\ell \leftarrow 2^{-\ell}$.
  Set phase confidence $\delta_\ell \leftarrow \delta/(|\mathcal{A}_\ell|^2 \ell(\ell+1))$.
  Compute the sampling policy $\mathbf{p}_\ell$ for $\mathcal{A}_\ell$ via (9).
  Calculate phase length n_ℓ via (10).
  Data Collection:
    Initialize $\mathbf{B} \leftarrow \mathbf{0}_{d\times d}$, $\mathbf{b} \leftarrow \mathbf{0}_d$.
  for $s = 1$ **to** n_ℓ **do**
    Sample $a_s \sim \mathbf{p}_\ell$ and observe reward r_s.
    Compute centered feature: $\tilde{x}_{a_s} \leftarrow x_{a_s} - \sum_{i=1}^K \mathbf{p}_\ell(i) x_i$.
    Update statistics: $\mathbf{B} \leftarrow \mathbf{B} + \tilde{x}_{a_s}\tilde{x}_{a_s}^\top$, $\mathbf{b} \leftarrow \mathbf{b} + \tilde{x}_{a_s} r_s$.
  end for
  Estimation and Elimination:
  Compute estimator: $\hat{\theta}_\ell \leftarrow \big(\mathbf{B} + \log(n_\ell/\delta_\ell)\,\mathbf{I}_d\big)^{-1}\mathbf{b}$.
  Identify empirical best: $\hat{z}_\ell \in \arg\max_{z \in \mathcal{A}_\ell} z^\top \hat{\theta}_\ell$.
  Update active set: $\mathcal{A}_{\ell+1} \leftarrow \{\, z \in \mathcal{A}_\ell \mid (\hat{z}_\ell - z)^\top \hat{\theta}_\ell < \varepsilon_\ell \,\}$.
  Increment phase: $\ell \leftarrow \ell + 1$.
end while
Output: The single arm remaining in $\mathcal{A}_\ell$.

Theorem 1 (High-probability sample complexity bound)**.** *Under Assumption 1, with probability at least $1-\delta$, Algorithm 2 correctly identifies the best target arm $z^\star$, and its stopping time satisfies*

$$\boldsymbol{\tau}(\delta) \lesssim \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1) + d^2,$$

where the notation $\lesssim$ suppresses absolute constants and logarithmic factors.

Discussion. The theorem shows that, whenever the shifted-linear benchmark is finite, our algorithm attains a high-probability sample-complexity upper bound of order $\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1) + d^2$ up to logarithmic factors. Combined with Proposition 1, this shows that $\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1)$ is unavoidable in general and therefore the right benchmark to compare against.

The additive d^2 term comes from stability requirements for orthogonalized regression. Importantly, it is independent of the instance-specific gaps. On the hard instances that dominate fixed-confidence complexity, the leading term $\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1)$ therefore governs the sample complexity.

The standard non-transductive guarantee is recovered

immediately by taking $\mathcal{Z} = \mathcal{X}$.

Corollary 2 (Sample complexity: non-transductive case)**.** *By setting $\mathcal{Z} = \mathcal{X}$ in Theorem 1, we recover the guarantee for the standard (non-transductive) setting:*

$$\boldsymbol{\tau}(\delta) \lesssim \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} : \mathcal{X} - x_1) + d^2 = \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{X} - x_1) + d^2,$$

where the notation $\lesssim$ suppresses absolute constants and logarithmic factors.

Thus, in both the transductive and standard settings, semiparametric BAI has essentially the same instance-dependent complexity landscape as linear-bandit BAI on the shifted feature space $\mathcal{X} - x_1$.

Normalization and General Scales. For clarity we present the normalized case $|\nu_t| \le 1$ with unit sub-Gaussian noise. More generally, if $|\nu_t| \le R$ and η_t is α-sub-Gaussian, rescaling rewards by $1/(R+\alpha)$ reduces to this setting, so the confidence width scales with $R+\alpha$ and the sample-complexity bound by $(R+\alpha)^2$; see also Kim et al. (2025).

6 EXPERIMENTS WITH SYNTHETIC DATA

6.1 Two Feature Instances

We consider two synthetic instance families in the standard non-transductive setting $\mathcal{Z} = \mathcal{X}$. In all synthetic experiments, the reward is generated according to the semiparametric model

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

where $\eta_t \sim \mathcal{N}(0,1)$, $\nu_t = 1 + \sin(2t)$ as in Kim et al. (2025). All reported numbers are averages over 100 independent runs. Unless stated otherwise, we use $\delta = 0.1$ throughout, including the real-data experiments below.

Feature set 1: small-gap instance. We set $\mathcal{X} = \{e_1, \dots, e_d, z\}$ with $z = \cos(\alpha)\, e_1 + \sin(\alpha)\, e_2$. We also set $\theta^\star = 2e_1$. This type of instance is standard in linear-bandit BAI experiments (Fiez et al., 2019). We set $\alpha = 0.2$, which yields a small suboptimality gap $\Delta_2 \approx 0.04$. For this family, we report results for $d = 10$ with $\delta = 0.1$.

Feature set 2: uniform-feature instance. For the second family, we sample $\mathcal{X}$ with $|\mathcal{X}| = K$ uniformly from the unit sphere $\mathbb{S}^{d-1}$ and again set $\theta^\star = 2e_1$. We consider $d = 10$, $K = 100$, and $\delta = 0.1$. Detailed standard deviations are reported in the appendix.

6.2 Baselines and Hyperparameters

We compare `SP-BAI` with two semiparametric baselines and one misspecified linear-bandit reference: `SBE` (Kim et al., 2025), `G-Opt`, and `RAGE` (Fiez et al., 2019).

SBE. This is the arm-elimination procedure of Kim et al. (2025), which uses G-optimal design for orthogonalized regression. All other implementation details follow Kim et al. (2025).

G-Opt. This oracle one-shot baseline assumes the smallest nonzero gap Δ_2 is known. We compute an approximate G-optimal design over $\mathcal{X}$ using the semiparametric design routine of Kim et al. (2025), sample according to that design for the corresponding one-shot budget from their fixed-confidence analysis, then fit orthogonalized regression and output $\arg\max_x x^\top \hat{\theta}$.

RAGE. RAGE (Fiez et al., 2019) is a nearly optimal algorithm for linear-bandit BAI, but it does not model the round-wise shared baseline shift. We therefore include it only as a misspecified reference point.

Hyperparameters. We use the same constants R_1, R_2 in Eq. (5) throughout and set $R_1 = R_2 = 1/3$ in every experiment. The confidence level is $\delta = 0.1$ in every experiment. In our implementation of `SP-BAI`, the anchor arm used inside the `XOR` design is chosen phase by phase as the current empirical best from the previous phase. Since the theory allows any anchor up to constant factors, this choice is simply a practical heuristic.

Table 1: Small-gap instance, $d = 10$ (100 runs).

	Avg. τ	Avg. Error Prob.
`SBE`	1 881 513.94	0.0000
`G-Opt`	387 106.00	0.0000
`SP-BAI`	**187924.68**	**0.0000**
`RAGE` ($\sigma = 1$)	22 212.00	1.0000
`RAGE` ($\sigma = 3$)	199 908.00	1.0000

6.3 Simulation Results

Table 1 reports the small-gap result. On this instance, `SP-BAI` achieves the smallest average stopping time among the semiparametric methods while maintaining zero empirical error over 100 runs. By contrast, `RAGE` stops much earlier at small σ but fails completely, confirming that the linear model is badly misspecified in this setting.

Table 2 gives a complementary random uniform-feature instance. This example highlights calibration rather than raw stopping time. The oracle one-shot `G-Opt` baseline is fastest on average, but its error probability rises to 0.29. `SP-BAI` is more conservative, yet it achieves the smallest empirical error (0.01) among all reported methods.

Table 2: Uniform-feature instance, $d = 10$, $K = 100$ (100 runs).

	Avg. τ	Avg. Error Prob.
SBE	1 523 534.64	0.1000
G-Opt	323 509.88	0.2900
SP-BAI	**2038886.49**	**0.0100**
RAGE ($\sigma = 1$)	4 367 940.79	0.9500
RAGE ($\sigma = 3$)	6 065 302.82	0.9400

7 REAL-WORLD DATA: JESTER JOKE RATINGS

We next test the practical utility of our method on the *Jester* dataset (Goldberg et al., 2001), which contains continuous user ratings for jokes. Jester exhibits substantial *user-specific baseline variation*: some users systematically rate most jokes high, while others rate most jokes low, regardless of the joke identity.

Experimental Setup. We focus on $K = 8$ jokes, namely ARM_IDS $= \{7, 8, 13, 15, 16, 17, 18, 19\}$. For this joke set, we restrict attention to the maximal complete-case subset, consisting of the 50,699 users who rated all eight jokes. This lets each pull correspond to sampling a user and then querying any joke without introducing an additional missing-data model. Since the action set is finite, we use one-hot features $x_i = e_i$. Each pull samples a user uniformly with replacement from these 50,699 users and observes the corresponding joke rating. The ground-truth best arm, Joke #19, is defined by the empirical mean over the full subset. As in the synthetic experiments, we use $R_1 = R_2 = 1/3$ and $\delta = 0.1$.

7.1 Justification of Semiparametric Reward Model: Impact of Baseline Correction

To isolate the value of baseline correction, we first compare a **Uniform** sampling strategy with **DEO** (the approximate G-optimal design method for semiparametric bandits from Kim et al. 2025) in a single-phase setting. Both methods collect a fixed budget of samples and then rank arms by their estimated values, but DEO uses orthogonalized regression to remove the user baseline ν_t. We report 100 runs at budgets $T \in \{3000, 5000\}$.

Results. Accounting for the user baseline matters already in this simple ranking task. As shown in Table 3, DEO identifies the best joke in 65% of runs at budget 3000, versus 58% for uniform sampling, and the gap widens to 69% versus 58% at budget 5000. The naive estimator absorbs the large variance of ν_t into the arm means, making it harder to separate the two leading jokes, 19 and 17.

Table 3: Jester toy ranking: probability of ranking Joke #19 first over 100 runs.

Budget	DEO	Uniform
3000	**0.65**	0.58
5000	**0.69**	0.58

Table 4: Jester (fixed-confidence BAI): Success probability and sample complexity over 100 runs. For LUCB and AE, we report the smallest-sample choice among σ-values whose empirical error probability is below 10%; the full grid appears in the appendix.

Method	Success Prob.	Avg. Pulls
SP-BAI **(Ours)**	**0.97**	61 546
SBE	0.98	266 706
LUCB ($\sigma = 3.5$)	0.96	136 698
AE ($\sigma = 1.5$)	0.98	105 939

7.2 Fixed-confidence BAI on Jester

We next evaluate fixed-confidence BAI on Jester. We compare our SP-BAI algorithm against (i) SBE (Kim et al., 2025), which uses approximate G-optimal design but not our $\mathcal{XY}$-optimization, and (ii) standard MAB baselines following Jamieson and Nowak (2014), namely LUCB and AE, run over the noise-proxy grid $\sigma \in \{1, 1.5, 2, 2.5, 3, 3.5, 4\}$. In the main table, for each MAB baseline we report the *smallest-sample* choice among those whose empirical error probability is below 10%; the full grid is deferred to Appendix F.

Results. Table 4 reports averages over 100 runs. SP-BAI achieves a **97% success rate** with only **61,546** average pulls, while SBE attains **98%** success but requires **266,706** pulls, about 4.3× more samples. Under the "error < 0.1" rule, the best standard MAB baselines are LUCB with $\sigma = 3.5$ and AE with $\sigma = 1.5$, and both remain less sample-efficient than SP-BAI.

8 CONCLUSION

We studied fixed-confidence BAI in semiparametric bandits, where rewards contain an arbitrary baseline shift in addition to a linear treatment effect. We identified the correct shifted-linear complexity benchmark, developed a new $\mathcal{XY}$-design for orthogonalized regression, and proved a high-probability sample-complexity upper bound controlled by that benchmark up to logarithmic factors and an additive d^2 term. Two natural directions for future work are fixed-budget BAI in semiparametric bandits and top-m identification under the fixed-confidence criterion.

Bibliography

Abbasi-Yadkori, Y., Bartlett, P., Gabillon, V., Malek, A., and Valko, M. (2018). Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on learning theory*, pages 918–949. PMLR.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Choi, Y.-G., Kim, G.-S., Paik, S., and Paik, M. C. (2023). Semi-parametric contextual bandits with graph-laplacian regularization. *Information Sciences*, 645:119367.

Degenne, R., Ménard, P., Shang, X., and Valko, M. (2020). Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR.

Fiez, T., Jain, L., Jamieson, K. G., and Ratliff, L. (2019). Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151.

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

Greenewald, K., Tewari, A., Murphy, S., and Klasnja, P. (2017). Action centered contextual bandits. *Advances in neural information processing systems*, 30.

Jamieson, K. and Nowak, R. (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE.

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

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049.

Kennedy, E. H., Balakrishnan, S., Robins, J. M., and Wasserman, L. (2024). Minimax rates for heterogeneous causal effect estimation. *Annals of statistics*, 52(2):793.

Kim, G.-S. and Paik, M. C. (2019). Contextual multi-armed bandit algorithm for semiparametric reward model. In *International Conference on Machine Learning*, pages 3389–3397. PMLR.

Kim, S.-J., Kim, G.-S., and Oh, M.-h. (2025). Experimental design for semiparametric bandits. In Haghtalab, N. and Moitra, A., editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3215–3252. PMLR.

Krishnamurthy, A., Wu, Z. S., and Syrgkanis, V. (2018). Semiparametric contextual bandits. In *International Conference on Machine Learning*, pages 2776–2785. PMLR.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1.

Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms*. Cambridge University Press (preprint).

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

Lykouris, T., Mirrokni, V., and Paes Leme, R. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122. ACM.

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

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27.

Tao, C., Blanco, S., and Zhou, Y. (2018). Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR.

Wen, Y., Han, Y., and Zhou, Z. (2025). Joint value estimation and bidding in repeated first-price auctions. *arXiv preprint arXiv:2502.17292*.

Xu, L., Honda, J., and Sugiyama, M. (2018). A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR.

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

Zhong, Z., Cheung, W. C., and Tan, V. (2021). Probabilistic sequential shrinking: A best arm identification algorithm for stochastic bandits with corruptions. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12772–12781. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Nearly Optimal Best Arm Identification for Semiparametric Bandits: Supplementary Materials

Contents

A Additional Discussion of Related Work

Semiparametric Bandits The semiparametric reward model was introduced for contextual decision making, where the observed reward consists of a structured treatment effect plus an unrestricted baseline shift (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim and Paik, 2019). That literature primarily studies regret minimization or policy learning under changing contexts. The closest precursor to our work is the recent fixed-feature study of Kim et al. (2025). They developed the experimental-design viewpoint for semiparametric bandits, established sharp concentration for orthogonalized regression, and proposed a G-optimal-design-based procedure together with an initial BAI guarantee. Our work builds directly on that foundation, but addresses a question left open there: what is the correct instance-dependent benchmark for fixed-confidence BAI, and can it be attained?

Linear-Bandit BAI and Experimental Design The broader linear-bandit literature begins with linearly parameterized exploration and regret minimization (Rusmevichientong and Tsitsiklis, 2010; Goldenshluger and Zeevi, 2013). Within that literature, pure exploration and best-arm identification have developed into a distinct direction. Early fixed-confidence guarantees for linear-bandit BAI were established by Soare et al. (2014), who already highlighted the role of design matrices and confidence ellipsoids. Subsequent works sharpened this picture in several directions: improved adaptivity and computational efficiency (Tao et al., 2018; Xu et al., 2018), asymptotically or instance-optimal fixed-confidence procedures (Degenne et al., 2020; Jedra and Proutiere, 2020), fixed-budget minimax formulations (Yang and Tan, 2022), and a broader understanding that, unlike in classical MAB, the geometry of the feature vectors is the main determinant of statistical difficulty.

For our paper, the most relevant thread is transductive and instance-dependent experimental design for linear BAI. Fiez et al. (2019) showed that in the transductive setting the correct complexity is governed by the variances

of pairwise target contrasts, which leads to $\mathcal{XY}$-design rather than G-optimal design. This point is central for understanding our contribution. G-optimal design controls uniform prediction error over the sampling set $\mathcal{X}$, whereas BAI depends on the specific directions needed to separate the best candidate from its competitors. In particular, once elimination begins, only a small subset of contrast directions matters, so a design that is optimal for prediction need not be optimal for identification.

Our lower bound is expressed exactly through the shifted transductive linear instance $(\mathcal{Z} : \mathcal{X} - x_1)$, and our algorithm can therefore be viewed as importing the linear-bandit design philosophy into the semiparametric model. At the same time, the transfer is not formal. In linear bandits, the relevant covariance matrix is simply the second moment of the chosen design. In semiparametric bandits, orthogonalized regression instead works with centered covariance matrices that themselves depend on the policy through $\bar{x}_{\mathbf{p}}$. Thus, even though the right benchmark remains linear-bandit-like, the design problem and the analysis are genuinely new.

Corrupted and Adversarially Perturbed Bandits Another adjacent line of work studies stochastic bandits with adversarial corruptions (Lykouris et al., 2018), corrupted best-arm identification (Zhong et al., 2021), and best-arm identification in mixed stochastic-adversarial regimes (Abbasi-Yadkori et al., 2018). These models allow perturbations that are effectively arm-dependent and are typically quantified by a corruption budget C or an adversarial contamination level, with guarantees that deteriorate as the environment becomes less stochastic.

Our semiparametric model is structurally different. The baseline shift ν_t is shared across all arms at round t and is chosen before the learner samples the current action. If one treated this shared shift as an arbitrary corruption, the induced corruption budget could be as large as order T, since a nontrivial perturbation may appear on every round. The problem is tractable here not because the cumulative perturbation is small, but because the perturbation has a special common-component structure. Orthogonalized regression is designed precisely to cancel this shared term in conditional expectation, so the learner can still recover linear-bandit-type instance dependence even when $\sum_{t=1}^{T} |\nu_t|$ is large. In this sense, corruption-robust methods are related but target a different benchmark: they provide generic robustness to unstructured contamination, whereas our algorithm exploits a specific semiparametric structure that those methods do not use.

B Proof of Proposition 1

Proof. The proof is based on an observation-law equivalence between a carefully chosen semiparametric hard instance and a shifted linear bandit instance.

Consider a semiparametric bandit environment defined by the source feature set $\mathcal{X}$, the target feature set $\mathcal{Z}$, the parameter $\theta^\star$, and a deterministic shift sequence $\nu_t \equiv -x_1^\top \theta^\star$ for all $t \geq 1$. First, we verify that this environment satisfies Assumption 1. Since $\|x_1\|_2 \leq 1$ and $\|\theta^\star\|_2 \leq 1$, we have $|\nu_t| \leq 1$, which is a valid shift sequence.

In this environment, the reward observed at round t after playing arm $a_t \in [K]$ is

$$\begin{aligned} r_t &= x_{a_t}^\top \theta^\star + \nu_t + \eta_t \\ &= x_{a_t}^\top \theta^\star - x_1^\top \theta^\star + \eta_t \\ &= (x_{a_t} - x_1)^\top \theta^\star + \eta_t. \end{aligned}$$

Let $\tilde{x}_i := x_i - x_1$ denote the shifted feature vectors. The resulting observation process is statistically identical to that of a linear bandit with source feature set $\tilde{\mathcal{X}} = \{\tilde{x}_1, \ldots, \tilde{x}_K\}$ and parameter $\theta^\star$.

Now consider the transductive identification problem. The goal is to identify

$$z^\star = \arg\max_{z \in \mathcal{Z}} z^\top \theta^\star.$$

Any δ-correct algorithm `Alg` for the semiparametric bandit must identify $z^\star$ with probability at least $1 - \delta$ for every admissible shift sequence $\{\nu_t\}$. Therefore, `Alg` must in particular succeed on the hard instance where $\nu_t \equiv -x_1^\top \theta^\star$.

Because the history distribution $\{(a_s, r_s)\}_{s \leq t}$ generated by the semiparametric model with $\nu_t = -x_1^\top \theta^\star$ is identical to that of the linear bandit with source features $\mathcal{X} - x_1$, any stopping time $\boldsymbol{\tau}$ and recommendation rule valid for the semiparametric problem are also valid for this linear bandit instance.

We invoke the standard information-theoretic lower bound for transductive linear bandit BAI. For a linear bandit with source features $\tilde{\mathcal{X}}$ and target features $\mathcal{Z}$, the expected sample complexity is lower-bounded by $\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \tilde{\mathcal{X}}) \log(1/\delta)$ (Fiez et al., 2019), where:

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \tilde{\mathcal{X}}) = \min_{\mathbf{p} \in \Delta^{(K)}} \max_{z \in \mathcal{Z} \setminus \{z^\star\}} \frac{\|z^\star - z\|^2_{(\sum_{i=1}^K p_i \tilde{x}_i \tilde{x}_i^\top)^{-1}}}{((z^\star - z)^\top \theta^\star)^2}.$$

Substituting $\tilde{x}_i = x_i - x_1$, we conclude that for any δ-correct semiparametric algorithm:

$$\mathbb{E}[\boldsymbol{\tau}(\delta)] \gtrsim \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1) \log\left(\frac{1}{\delta}\right).$$

□

C Proof of Proposition 2

Proof. Without loss of generality, it suffices to prove

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_2) \leq 4\, \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1).$$

Recall that

$$\boldsymbol{\Sigma}_{-k,\mathbf{p}} := \sum_{i=1}^K p_i (x_i - x_k)(x_i - x_k)^\top.$$

Fix any $\mathbf{p} \in \Delta^{(K)}$. Since

$$x_i - x_1 = (x_i - x_2) + (x_2 - x_1),$$

the matrix inequality $(a+b)(a+b)^\top \preceq 2aa^\top + 2bb^\top$ gives

$$\begin{aligned}
\boldsymbol{\Sigma}_{-1,\mathbf{p}} &= \sum_{i=1}^K p_i (x_i - x_1)(x_i - x_1)^\top \\
&\preceq 2 \sum_{i=1}^K p_i (x_i - x_2)(x_i - x_2)^\top + 2(x_1 - x_2)(x_1 - x_2)^\top \\
&= 2\boldsymbol{\Sigma}_{-2,\mathbf{p}} + 2(x_1 - x_2)(x_1 - x_2)^\top.
\end{aligned}$$

Now define a nonnegative weight vector $\mathbf{q}$ by

$$q_1 := 2 + 2p_1, \qquad q_i := 2p_i \ \text{ for } i \geq 2.$$

Then $\sum_{i=1}^K q_i = 4$, and

$$\sum_{i=1}^K q_i (x_i - x_2)(x_i - x_2)^\top = 2\boldsymbol{\Sigma}_{-2,\mathbf{p}} + 2(x_1 - x_2)(x_1 - x_2)^\top.$$

Hence, for every vector y,

$$\|y\|^2_{\boldsymbol{\Sigma}^{-1}_{-1,\mathbf{p}}} \geq \|y\|^2_{\left(\sum_{i=1}^K q_i (x_i - x_2)(x_i - x_2)^\top\right)^{-1}}.$$

Let $\bar{\mathbf{q}} := \mathbf{q}/4 \in \Delta^{(K)}$. By homogeneity of inverse norms,

$$\|y\|^2_{\left(\sum_{i=1}^K q_i (x_i - x_2)(x_i - x_2)^\top\right)^{-1}} = \frac{1}{4} \|y\|^2_{\left(\sum_{i=1}^K \bar{q}_i (x_i - x_2)(x_i - x_2)^\top\right)^{-1}}.$$

Therefore, for this fixed $\mathbf{p}$,

$$\begin{aligned}\max_{z\in\mathcal{Z}\setminus\{z^\star\}} \frac{\|z^\star - z\|^2_{\mathbf{\Sigma}^{-1}_{-1,\mathbf{p}}}}{|(z^\star - z)^\top\theta^\star|^2} &\geq \frac{1}{4}\max_{z\in\mathcal{Z}\setminus\{z^\star\}} \frac{\|z^\star - z\|^2_{\left(\sum_{i=1}^K \bar{q}_i(x_i-x_2)(x_i-x_2)^\top\right)^{-1}}}{|(z^\star - z)^\top\theta^\star|^2} \\ &\geq \frac{1}{4}\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_2),\end{aligned}$$

where the second step uses only that $\bar{\mathbf{q}} \in \Delta^{(K)}$, so the value at $\bar{\mathbf{q}}$ is at least the minimum over all designs. Since this lower bound holds for every $\mathbf{p} \in \Delta^{(K)}$, taking the minimum over $\mathbf{p}$ yields

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_1) \geq \frac{1}{4}\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_2).$$

Rearranging yields

$$\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_2) \leq 4\,\boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z}:\mathcal{X}-x_1).$$

□

D Proof of Proposition 3

Proof. Let $(\tilde{p}_1,\dots,\tilde{p}_K)$ be an optimal $\mathcal{XY}$-design for the shifted linear instance with source features $\mathcal{X}-x_1$ and target set $\mathcal{A}$. By construction, $\tilde{p}_1 = 0$. Define

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

Applying Lemma 9 from Kim et al. (2025) to the mixture policy $\mathbf{p}_{\mathrm{xor}}(\mathcal{A})$, we obtain

$$\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_{\mathrm{xor}}(\mathcal{A})} \succeq \frac{1}{4}\widetilde{\mathbf{\Sigma}}.$$

Since inverse quadratic forms reverse the PSD order, for every contrast $y \in \mathbb{R}^d$,

$$\|y\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_{\mathrm{xor}}(\mathcal{A})}} \leq 4\|y\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}}.$$

Now take $y = x - x'$ for arbitrary $x, x' \in \mathcal{A}$. Then

$$\begin{aligned}\|x-x'\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_{\mathrm{xor}}(\mathcal{A})}} &\leq 4\|x-x'\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}} \\ &= 4\|(x-x_1)-(x'-x_1)\|^2_{\widetilde{\mathbf{\Sigma}}^{-1}} \\ &\leq 4\mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A}:\mathcal{X}-x_1),\end{aligned}$$

where the last step is exactly the defining optimality property of $\tilde{\mathbf{p}}$ for the shifted linear instance. Taking the maximum over $x, x' \in \mathcal{A}$, we conclude that

$$\mathcal{V}_{\mathrm{cov}}(\mathcal{A}:\mathcal{X},\mathbf{p}_{\mathrm{xor}}(\mathcal{A})) \leq 4\mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A}:\mathcal{X}-x_1).$$

□

E Proof of Theorem 1

E.1 Good Events and Correctness

For each target arm $z \in \mathcal{Z}\setminus\{z^\star\}$, define the gap

$$\Delta_z := (z^\star - z)^\top\theta^\star, \qquad \Delta_{\min} := \min_{z\in\mathcal{Z}\setminus\{z^\star\}} \Delta_z.$$

We also define the deterministic phase cutoff

$$L_\star := \max\left\{1, \left\lceil \log_2\left(\frac{4}{\Delta_{\min}}\right)\right\rceil\right\}.$$

Definition 2 (Good events)**.** *For each phase $\ell \in [L_\star]$, let $\mathscr{E}_\ell$ be the event that*

$$\left|(z - z')^\top(\hat{\theta}_\ell - \theta^\star)\right| \le \varepsilon_\ell \qquad \text{for all } z, z' \in \mathcal{A}_\ell.$$

Define $\mathscr{E}_\star := \bigcap_{\ell=1}^{L_\star} \mathscr{E}_\ell$.

Lemma 1. *For every phase $\ell \in [L_\star]$, $\mathbb{P}[\mathscr{E}_\ell] \ge 1 - \delta/(\ell(\ell+1))$. Consequently,*

$$\mathbb{P}[\mathscr{E}_\star] \ge 1 - \delta.$$

Proof. Fix a phase ℓ, and condition on the history at the start of that phase. Under this conditioning, the active set $\mathcal{A}_\ell$, the phase confidence level $\delta_\ell = \delta/(|\mathcal{A}_\ell|^2\ell(\ell+1))$, the sampling distribution $\mathbf{p}_\ell$, and the phase length n_ℓ are all deterministic. We prove a conditional probability bound that is uniform over the conditioning event, and then remove the conditioning at the end.

Write $\mathbf{p}_\ell = \frac{1}{2}(\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell) + \mathbf{p}_G)$, and let

$$\bar{x}_{\mathrm{xor}} := \sum_i \mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)(i)x_i, \qquad \bar{x}_G := \sum_i \mathbf{p}_G(i)x_i.$$

Viewing $\mathbf{p}_\ell$ as the mixture that first chooses between $\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)$ and $\mathbf{p}_G$ with probability $1/2$ each, the law of total covariance gives

$$\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_\ell} = \frac{1}{2}\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)} + \frac{1}{2}\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_G} + \frac{1}{4}(\bar{x}_{\mathrm{xor}} - \bar{x}_G)(\bar{x}_{\mathrm{xor}} - \bar{x}_G)^\top,$$

where the final PSD term is the covariance of the two component means. In particular, the mixture covariance dominates each component covariance up to a factor of $1/2$, so

$$\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_\ell} \succeq \frac{1}{2}\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)}, \qquad \mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_\ell} \succeq \frac{1}{2}\mathbf{\Sigma}_{\mathrm{cov},\mathbf{p}_G}. \tag{11}$$

Hence, for every contrast $y = z - z'$ with $z, z' \in \mathcal{A}_\ell$,

$$\begin{aligned}
\|y\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_\ell}} &\le 2\|y\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)}} \\
&\le 2\,\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)) \\
&\le 8\,\mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A}_\ell : \mathcal{X} - x_1),
\end{aligned}$$

where the last step uses Proposition 3. Thus, in the concentration bound (4), the leading variance parameter L for the contrast set $\mathcal{A}_\ell - \mathcal{A}_\ell$ is controlled, up to an absolute constant, by $\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell)$.

We next control the source-only quantity M in (4). By (11) and the definition of $\mathbf{p}_G$,

$$\|x_i\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_\ell}} \le 2\|x_i\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_G}} \le 8d \qquad \text{for every } i \in [K].$$

Moreover,

$$\begin{aligned}
\|x_i - \bar{x}_{\mathbf{p}_\ell}\|_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_\ell}} &\le \sqrt{2}\,\|x_i - \bar{x}_{\mathbf{p}_\ell}\|_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_G}} \\
&= \sqrt{2}\left\|\sum_{j=1}^K \mathbf{p}_\ell(j)(x_i - x_j)\right\|_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_G}} \\
&\le \sqrt{2}\sum_{j=1}^K \mathbf{p}_\ell(j)\big(\|x_i\|_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_G}} + \|x_j\|_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_G}}\big) \\
&\le \sqrt{2}\cdot 2\sqrt{4d} = 4\sqrt{2d},
\end{aligned}$$

and therefore

$$\max_{i\in[K]} \|x_i - \bar{x}_{\mathbf{p}_\ell}\|^2_{\mathbf{\Sigma}^{-1}_{\mathrm{cov},\mathbf{p}_\ell}} \le 32d. \tag{12}$$

Applying the concentration bound (4) conditionally on the phase-ℓ history to each vector in the finite set $\mathcal{A}_\ell - \mathcal{A}_\ell$, using the L-control above together with the source-only M-bound (12), and taking a union bound over at most $|\mathcal{A}_\ell|^2$ contrasts, we obtain

$$\mathbb{P}[\mathscr{E}_\ell] \geq 1 - |\mathcal{A}_\ell|^2 \delta_\ell = 1 - \frac{\delta}{\ell(\ell+1)}$$

conditionally on the phase-ℓ history, provided the absolute constants R_1, R_2 in (10) are chosen large enough. Since the bound is uniform over the conditioning event, it also holds unconditionally. Finally,

$$\sum_{\ell=1}^{\infty} \frac{\delta}{\ell(\ell+1)} \leq \delta,$$

so $\mathbb{P}[\mathscr{E}_\star] \geq 1 - \delta$. □

Lemma 2 (Correctness)**.** *Under $\mathscr{E}_\star$, the best arm $z^\star$ is never eliminated, and after phase $L_\star$ the active set is exactly $\{z^\star\}$.*

Proof. We first show that $z^\star$ is never eliminated. Suppose $z^\star \in \mathcal{A}_\ell$. If $z^\star$ were removed in phase ℓ, then by the elimination rule there would exist some $\hat{z}_\ell \in \mathcal{A}_\ell$ such that

$$(\hat{z}_\ell - z^\star)^\top \hat{\theta}_\ell \geq \varepsilon_\ell.$$

Since $z^\star$ is optimal, $(\hat{z}_\ell - z^\star)^\top \theta^\star < 0$, hence

$$(\hat{z}_\ell - z^\star)^\top (\hat{\theta}_\ell - \theta^\star) \geq \varepsilon_\ell,$$

which contradicts $\mathscr{E}_\ell$. Therefore $z^\star$ survives every phase.

Next consider any suboptimal arm $z \in \mathcal{A}_\ell$ with $\Delta_z > 2\varepsilon_\ell$. Since $\hat{z}_\ell$ maximizes $u^\top \hat{\theta}_\ell$ over $u \in \mathcal{A}_\ell$,

$$(\hat{z}_\ell - z)^\top \hat{\theta}_\ell \geq (z^\star - z)^\top \hat{\theta}_\ell \geq \Delta_z - \varepsilon_\ell > \varepsilon_\ell,$$

where the second step uses $\mathscr{E}_\ell$. Thus every arm with gap larger than $2\varepsilon_\ell$ is removed in phase ℓ, so

$$\mathcal{A}_{\ell+1} \subseteq \{z \in \mathcal{Z} : \Delta_z \leq 2\varepsilon_\ell\}.$$

For the base case $\ell = 1$, we have $\mathcal{A}_1 = \mathcal{Z}$, and for every $z \in \mathcal{Z}$,

$$\Delta_z = (z^\star - z)^\top \theta^\star \leq \|z^\star - z\|_2 \|\theta^\star\|_2 \leq 2 = 4\varepsilon_1.$$

Hence

$$\mathcal{A}_1 \subseteq \{z \in \mathcal{Z} : \Delta_z \leq 4\varepsilon_1\} \cup \{z^\star\}.$$

Using $\mathcal{A}_{\ell+1} \subseteq \{z \in \mathcal{Z} : \Delta_z \leq 2\varepsilon_\ell\}$ and $\varepsilon_{\ell+1} = \varepsilon_\ell/2$, an immediate induction gives, for every $\ell \geq 1$,

$$\mathcal{A}_\ell \subseteq \{z \in \mathcal{Z} : \Delta_z \leq 4\varepsilon_\ell\} \cup \{z^\star\}.$$

Finally, by the definition of $L_\star$, we have $4\varepsilon_{L_\star} \leq \Delta_{\min}$. Therefore no suboptimal arm can remain after phase $L_\star$, and the active set is $\{z^\star\}$. □

E.2 Bounding Sample Complexity

Let

$$\tau_\star := \boldsymbol{\tau}^\star_{\mathrm{lin}}(\mathcal{Z} : \mathcal{X} - x_1).$$

Fix an optimal design $\mathbf{p}^\star \in \Delta^{(K)}$ for $\tau_\star$, and write

$$\boldsymbol{\Sigma}^\star := \sum_{i=1}^{K} p_i^\star (x_i - x_1)(x_i - x_1)^\top.$$

Under $\mathscr{E}_\star$, Lemma 2 shows that every arm $z \in \mathcal{A}_\ell \setminus \{z^\star\}$ satisfies

$$\Delta_z \le 4\varepsilon_\ell.$$

We first compare the linear design value for $\mathcal{A}_\ell$ with the full-instance lower-bound benchmark. For any policy $\mathbf{p}$,

$$\max_{u,u' \in \mathcal{A}_\ell} \|u - u'\|^2_{\mathbf{\Sigma}^{-1}_{-1,\mathbf{p}}} \le 4 \max_{z \in \mathcal{A}_\ell \setminus \{z^\star\}} \|z^\star - z\|^2_{\mathbf{\Sigma}^{-1}_{-1,\mathbf{p}}},$$

because every pairwise difference can be written as

$$u - u' = (u - z^\star) - (u' - z^\star)$$

and $\|a - b\|^2_{\mathbf{A}^{-1}} \le 2\|a\|^2_{\mathbf{A}^{-1}} + 2\|b\|^2_{\mathbf{A}^{-1}}$. Evaluating this bound at $\mathbf{p}^\star$, we obtain

$$\begin{aligned}
\mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A}_\ell : \mathcal{X} - x_1) &\le 4 \max_{z \in \mathcal{A}_\ell \setminus \{z^\star\}} \|z^\star - z\|^2_{(\mathbf{\Sigma}^\star)^{-1}} \\
&\le 4\tau_\star \max_{z \in \mathcal{A}_\ell \setminus \{z^\star\}} \Delta_z^2 \qquad (13) \\
&\le 64\tau_\star \varepsilon_\ell^2. \qquad (14)
\end{aligned}$$

Equation (14) is the key step that connects elimination to the benchmark $\tau_\star$. The first inequality converts arbitrary pairwise contrasts inside $\mathcal{A}_\ell$ into contrasts against the true best arm $z^\star$. The second inequality uses the defining property of the optimal benchmark design $\mathbf{p}^\star$: for every competitor $z \ne z^\star$,

$$\|z^\star - z\|^2_{(\mathbf{\Sigma}^\star)^{-1}} \le \tau_\star \Delta_z^2.$$

The last step then uses the consequence of the good event $\mathscr{E}_\star$, namely that every suboptimal arm surviving into phase ℓ must satisfy $\Delta_z \le 4\varepsilon_\ell$. Thus, once the active set has shrunk to near-optimal arms, its entire linear $\mathcal{X}\mathcal{Y}$-design value is automatically of order $\tau_\star \varepsilon_\ell^2$. This is exactly what allows the phase length to track the instance-dependent benchmark rather than a cruder worst-case quantity.

By (11) and Proposition 3,

$$\begin{aligned}
\mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_\ell) &\le 2\, \mathcal{V}_{\mathrm{cov}}(\mathcal{A}_\ell : \mathcal{X}, \mathbf{p}_{\mathrm{xor}}(\mathcal{A}_\ell)) \\
&\le 8\, \mathcal{V}^\star_{\mathrm{lin}}(\mathcal{A}_\ell : \mathcal{X} - x_1) \le 512\tau_\star \varepsilon_\ell^2. \qquad (15)
\end{aligned}$$

Substituting (15) into (10), we find

$$n_\ell \lesssim \tau_\star \log\left(\frac{\tau_\star \varepsilon_\ell}{\delta_\ell}\right) + d\sqrt{\tau_\star} \log\left(\frac{d}{\delta_\ell}\right).$$

Since $\delta_\ell = \delta/(|\mathcal{A}_\ell|^2 \ell(\ell+1))$, $|\mathcal{A}_\ell| \le H$, $\varepsilon_\ell = 2^{-\ell}$, and $L_\star = \mathcal{O}(\max\{1, \log(1/\Delta_{\min})\})$, all logarithmic factors above are polylogarithmic in d, H, $1/\delta$, and $1/\Delta_{\min}$. Therefore,

$$n_\ell \le \tilde{\mathcal{O}}(\tau_\star + d\sqrt{\tau_\star}).$$

Summing over $\ell = 1, \dots, L_\star$ yields

$$\begin{aligned}
\sum_{\ell=1}^{L_\star} n_\ell &\le \tilde{\mathcal{O}}(L_\star(\tau_\star + d\sqrt{\tau_\star})) \\
&\le \tilde{\mathcal{O}}\big(L_\star(\tau_\star + d^2)\big),
\end{aligned}$$

where the second step uses $d\sqrt{\tau_\star} \le \frac{1}{2}(d^2 + \tau_\star)$. Finally, on $\mathscr{E}_\star$ the algorithm stops by phase $L_\star$ by Lemma 2, so

$$\boldsymbol{\tau}(\delta) \le \sum_{\ell=1}^{L_\star} n_\ell \le \tilde{\mathcal{O}}(\tau_\star + d^2).$$

This proves Theorem 1.

F More on Experiments

F.1 Detailed Synthetic Results

We summarize the synthetic experiments again, now including empirical standard deviations of the stopping time. All tables below are based on 100 independent runs with $R_1 = R_2 = 1/3$. Tables 5 and 6 correspond to the synthetic instances in the main text, and include the full `RAGE` grid used to diagnose linear-model misspecification.

Table 5: Simulation Results Summary: Small-gap, $d = 10$

Algorithm	Avg. $\boldsymbol{\tau}$	Std. $\boldsymbol{\tau}$	Avg. Error Prob.
`SBE`	1 881 513.94	1 228 628.36	0.0000
`G-Opt`	387 106.00	0.00	0.0000
`SP-BAI`	187 924.68	112 295.59	0.0000
`RAGE` ($\sigma = 1$)	22 212.00	0.00	1.0000
`RAGE` ($\sigma = 2$)	88 848.00	0.00	1.0000
`RAGE` ($\sigma = 3$)	199 908.00	0.00	1.0000
`RAGE` ($\sigma = 4$)	355 392.00	0.00	1.0000

Table 6: Simulation Results Summary: Uniform-feature, $d = 10$, $K = 100$

Algorithm	Avg. $\boldsymbol{\tau}$	Std. $\boldsymbol{\tau}$	Avg. Error Prob.
`SBE`	1 523 534.64	2 822 607.36	0.1000
`G-Opt`	323 509.88	904 131.81	0.2900
`SP-BAI`	2 038 886.49	2 930 198.90	0.0100
`RAGE` ($\sigma = 1$)	4 367 940.79	1 924 666.73	0.9500
`RAGE` ($\sigma = 2$)	5 415 498.52	1 952 295.28	0.9500
`RAGE` ($\sigma = 3$)	6 065 302.82	2 243 377.14	0.9400
`RAGE` ($\sigma = 4$)	6 428 781.60	2 434 510.77	0.9600

The integrated `RAGE` rows make the misspecification pattern easy to read across instances. On the reported small-gap instance, `RAGE` fails uniformly across all tested noise scales $\sigma \in \{1, 2, 3, 4\}$: the empirical error probability is 1.0, while larger σ only increases the stopping time. The uniform-feature instance shows the same qualitative behavior, though less starkly, with error probabilities remaining around 0.94 to 0.96.

F.2 Additional Jester Results

Table 7 reports the full σ-grid used for the standard MAB baselines in the Jester fixed-confidence experiment. In the main text, we report only the smallest-sample configuration whose empirical error rate is below 10%; the full grid is included here for transparency.

Table 7: Jester fixed-confidence experiment: full σ-grid over 100 runs at $\delta = 0.1$.

Method	Success Prob.	Avg. Pulls
`SP-BAI`	0.970	61 546.4
`SBE`	0.980	266 706.2
`LUCB` ($\sigma = 1$)	0.300	64.0
`LUCB` ($\sigma = 1.5$)	0.460	3742.8
`LUCB` ($\sigma = 2$)	0.540	12 677.2
`LUCB` ($\sigma = 2.5$)	0.740	31 959.7
`LUCB` ($\sigma = 3$)	0.900	82 030.7
`LUCB` ($\sigma = 3.5$)	0.960	136 698.2
`LUCB` ($\sigma = 4$)	0.960	181 116.9
`AE` ($\sigma = 1$)	0.660	14 286.9
`AE` ($\sigma = 1.5$)	0.980	105 938.8
`AE` ($\sigma = 2$)	1.000	185 485.2
`AE` ($\sigma = 2.5$)	1.000	340 073.2
`AE` ($\sigma = 3$)	1.000	495 855.5
`AE` ($\sigma = 3.5$)	1.000	627 211.7
`AE` ($\sigma = 4$)	1.000	870 003.1