SEMIPARAMETRIC GENERALIZED LINEAR BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study finite-armed semiparametric generalized linear bandits (SGLBs), where each arm's mean reward follows a generalized linear model with an unknown common shift that may vary arbitrarily over time. This model extends semiparametric linear bandits to nonlinear links, where standard feature centering no longer removes the common shift exactly. We propose Localized Elimination for Semiparametric GLBs (`LE-SGLB`), which progressively localizes the active arms and uses centered experimental design to eliminate suboptimal arms. `LE-SGLB` achieves a leading round-dependent regret of $\tilde{\mathcal{O}}(\sqrt{dT\min\{d,\log K\}})$, up to initialization and lower-order terms, where d is the feature dimension, K is the number of arms, and T is the horizon. Importantly, this leading term is free of the global inverse-curvature parameter κ and is independent of the magnitude and variation of the common shifts. For logistic SGLBs, we further prove an $\Omega(d\sqrt{T})$ lower bound for $K = 2^d$, equivalently $\Omega(\sqrt{dT\log K})$ in this regime, matching the leading term of our upper bound up to logarithmic factors. Moreover, the lower bound remains $\Omega(d\sqrt{T})$ even as $\kappa^\star$ grows to order $\sqrt{T}/d$, where $\kappa^\star$ denotes the inverse average slope of the logistic link at the optimal arm. Thus, the favorable $1/\sqrt{\kappa^\star}$ local-curvature dependence of ordinary logistic bandits does not extend to SGLBs.

1 INTRODUCTION

Generalized linear bandits (GLBs) extend linear bandits by modeling the mean reward as a nonlinear function of a linear predictor. Given an action feature x, the conditional mean reward is modeled as $\mu(x^\top\theta^\star)$, where μ is a known link function and $\theta^\star$ is an unknown parameter. This framework covers a broad class of exponential-family reward models and has been extensively studied (Filippi et al., 2010; Li et al., 2017; Jun et al., 2017; Faury et al., 2020; Abeille et al., 2021; Faury et al., 2022; Lee et al., 2024; Clerico et al., 2025; Sawarni et al., 2024; Zhang et al., 2025; Lee & Oh, 2026). A common premise of standard GLBs is that, once $\theta^\star$ is known, the expected reward of an action is determined by its observed feature. In many sequential decision problems, however, outcomes may also depend on an unknown and unobserved condition that varies over time and affects all actions simultaneously. For example, a patient's latent health condition may shift the response to all candidate treatments in sequential treatment selection, and a user's latent engagement level may shift the response to all candidate items in a recommendation system.

Motivated by such settings, we study finite-armed *semiparametric generalized linear bandits* (SGLBs). Let $x_a \in \mathbb{R}^d$ denote the known feature vector of arm a, $\theta^\star \in \mathbb{R}^d$ an unknown parameter, and $\mu : \mathbb{R} \to \mathbb{R}$ a strictly increasing link function. At round t, the conditional mean reward of arm a is

$$\mathbb{E}[Y_t \mid \mathcal{G}_t, A_t = a] = \mu\big(\nu_t + x_a^\top\theta^\star\big)\,,$$

where $\mathcal{G}_t$ is the conditioning sigma-field containing the history and the current latent shift. The key additional component is $\nu_t \in \mathbb{R}$, an *unknown* and *unobserved* common shift shared by all arms at round t. It may vary adversarially over time and depend on the past and on the learner's current sampling distribution π_t, provided it is chosen before the current randomized action is drawn. Nevertheless, since μ is strictly increasing and ν_t is common to all arms, the optimal arm remains fixed and is determined solely by $x_a^\top\theta^\star$.

Semiparametric linear bandits, whose common-shift structure is closely related to our setting, allow an arbitrary action-independent baseline in addition to the linear treatment effect (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim & Paik, 2019; Chowdhury et al., 2023; Choi et al., 2023;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Kim et al., 2025). A central idea in this literature is *feature centering*: because the baseline enters additively, centering the randomized action features removes it exactly, allowing the action-dependent parameter to be learned without estimating the nuisance itself. This cancellation no longer holds in SGLBs. The common shift enters inside the nonlinear link together with the action-dependent score, so centering leaves a residual dependence on the unknown ν_t. Moreover, since the shifts are unobserved and may vary arbitrarily, they cannot simply be estimated or tracked as a separate process.

This difficulty is further compounded by the role of *curvature* in GLBs. The hidden shift changes the local slope μ' of the link around the current natural parameter and can therefore substantially alter how informative each observation is. Standard GLB analyses often involve the global inverse-curvature parameter $\kappa := (\inf_z \mu'(z))^{-1}$, which can be prohibitively large, and reducing this dependence has been a central goal in the GLB literature (Faury et al., 2020; Abeille et al., 2021; Faury et al., 2022; Lee et al., 2024; Sawarni et al., 2024; Zhang et al., 2025). The issue is especially pronounced in SGLBs: an unrestricted ν_t can move the natural parameter into regions where μ' is arbitrarily small, making a global curvature lower bound unsuitable for controlling the main learning process. This leads to the central question of this paper:

Can we achieve a κ-free leading regret guarantee for SGLBs
under uniform mean and variance bounds, but without a global slope lower bound?

We answer this question affirmatively by proposing Localized Elimination for Semiparametric GLBs (`LE-SGLB`). The algorithm first localizes the active arms so that the nonlinear link can be accurately approximated over the remaining score range. In the subsequent elimination procedure, each phase uses a fixed centered experimental design, allowing arm comparisons to preserve the underlying *ranking* despite the time-varying local slopes induced by the hidden shifts. This enables `LE-SGLB` to eliminate suboptimal arms without estimating either the hidden shifts or the local slopes themselves. An adaptive stopping rule determines when these comparisons are sufficiently accurate for safe elimination, progressively refining the active set. Crucially, smaller local slopes may make each elimination phase last longer, but they also shrink the reward gaps between active arms, reducing the regret incurred during these additional samples. This information–regret coupling removes the need for a global lower bound on μ' after initialization. Our main contributions are summarized as follows:

- **First κ-free leading regret guarantee for SGLBs.** `LE-SGLB` satisfies, with high probability, a regret upper bound of $\widetilde{\mathcal{O}}\big(\kappa_{\mathrm{init}}^2 Kd + \sqrt{dT\min\{d,\log K\}} + d\min\{d,\log K\}\big)$ (Theorem 1). The leading term is independent of both the magnitude and variation of the common shifts $(\nu_t)_{t=1}^T$ and the global inverse-curvature parameter κ. A curvature lower bound is needed only during initialization through κ_{init}; after localization, no positive global lower bound on μ' is required. To the best of our knowledge, this is the first SGLB regret guarantee whose leading round-dependent term is independent of both the scale of the common shifts and κ.
- **Matching lower bound across local-curvature levels.** We prove a matching lower bound of $\Omega(d\sqrt{T})$ for $K = 2^d$, equivalently $\Omega(\sqrt{dT\log K})$, up to logarithmic factors (Theorem 2). Unlike ordinary logistic bandits, which can enjoy a $d\sqrt{T/\kappa^\star}$ dependence (Abeille et al., 2021), the SGLB lower bound remains $\Omega(d\sqrt{T})$ for every $\kappa^\star$ up to order $\sqrt{T}/d$. The unobserved common shifts ν_t can attenuate the observable differences between nearby instances without a comparable reduction in reward variability, making them harder to distinguish (see Appendix G.3).
- **Improved initial localization.** We further provide an improved initialization method that reduces the initialization regret from $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 Kd)$ to $\widetilde{\mathcal{O}}\big(\min\{\kappa_{\mathrm{init}}^2 K\sqrt{d},\ \kappa_{\mathrm{init}}^4 d^{9/2}\}\big)$ (Theorem I.1), making the initialization guarantee suitable for large action sets.

2 Related Work

Semiparametric linear bandits and experimental design. Semiparametric bandit models learn action-dependent effects in the presence of potentially complex action-independent baselines. Greenewald et al. (2017) introduced action-centered contextual bandits, using randomized treatment assignment to separate a linear treatment effect from a time-varying baseline, while Krishnamurthy et al. (2018) developed orthogonalized estimation and obtained $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret. Related randomized and orthogonalized approaches were studied in (Kim & Paik, 2019; Chowdhury et al., 2023; Choi

et al., 2023). For fixed finite action sets, Kim et al. (2025) introduced a centered experimental-design approach and obtained minimax-optimal regret up to logarithmic factors, together with PAC and best-arm identification guarantees, building on optimal experimental design for linear best-arm identification (Soare et al., 2014; Fiez et al., 2019). Related design-based ideas have also been studied for transductive best-arm identification (Kim, 2026). Our setting shares the action-independent nuisance structure and builds on the centered-design approach of Kim et al. (2025), but the nuisance enters *inside* a nonlinear link. Consequently, feature centering no longer yields a nuisance-free estimating equation, and the effective sensitivity of reward differences varies with the unobserved common shift. Thus, the orthogonalized estimation methods used in semiparametric linear bandits do not directly extend to our setting.

Nonlinear confounding and semiparametric inference. Beyond semiparametric linear bandits, Kirschner & Krause (2021) study optimization of a kernelized objective under an adversarial action-independent additive bias, $Y_t = f(x_t) + \nu_t + \epsilon_t$. Their randomization over two actions removes the additive nuisance exactly, yielding an unbiased pairwise signal proportional to $f(x_i) - f(x_j)$. In our setting, however, the nuisance enters *inside* the nonlinear link: the analogous randomized comparison has conditional mean $\mu(\nu_t + x_i^\top \theta^\star) - \mu(\nu_t + x_j^\top \theta^\star)$, which still depends on the unobserved shift. Thus, unlike the exact debiasing available in Kirschner & Krause (2021), our analysis relies on localization to make pairwise comparisons share an unknown positive scale while preserving their ranking. Related semiparametric work studies generalized linear regression with nonparametric nuisance components under adaptive data collection (Lin et al., 2025), as well as inference and policy evaluation in adaptive experiments (Cook et al., 2024; Hadad et al., 2021). These works focus on estimation and inference rather than cumulative regret under arbitrary unobserved common shifts.

Standard generalized linear bandits. Generalized linear bandits (GLBs) model the conditional mean reward through a known nonlinear link applied to a linear predictor. Following the optimistic approach of Filippi et al. (2010), subsequent work developed sharper confidence bounds and regret guarantees for GLBs (Li et al., 2017; Jun et al., 2017; Abeille et al., 2021; Faury et al., 2022; Zhang & Sugiyama, 2024; Lee et al., 2024; Clerico et al., 2025; Lee & Oh, 2024; 2025; Zhang et al., 2025; Sawarni et al., 2024), with related extensions to unknown but fixed monotone reward functions (Kang et al., 2026). A central issue is the dependence on the curvature of the link function. For logistic bandits, a special case of GLBs, Faury et al. (2020) removed the global inverse-curvature parameter from the leading regret term, while Abeille et al. (2021) established a favorable $\widetilde{\mathcal{O}}(d\sqrt{T/\kappa^\star})$ dependence on the local curvature at the optimal arm, with $\kappa^\star = 1/\sigma'(x_{a^\star}^\top \theta^\star)$. In contrast, our lower bound (Theorem 2) shows that this favorable $1/\sqrt{\kappa^\star}$ dependence does not extend to the semiparametric setting: even when $\kappa^\star$ grows to order $\sqrt{T}/d$, the regret can remain $\Omega(d\sqrt{T})$.

3 Problem Setting

Notation. For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$, and let $\Delta(\mathcal{A})$ denote the probability simplex over $\mathcal{A}$. For $V \succeq 0$, write $\|x\|_V := \sqrt{x^\top V x}$, and let $V^\dagger$ denote its Moore–Penrose pseudoinverse. We use $\widetilde{\mathcal{O}}(\cdot)$ to suppress logarithmic factors and fixed problem-scale constants.

Semiparametric generalized linear model. In this paper, we consider a finite-armed bandit problem with a fixed action set $\mathcal{X} = \{x_1, \dots, x_K\} \subset \mathbb{R}^d$, where each arm $a \in [K]$ is represented by a known feature vector x_a.[1] Let $\mathcal{H}_{t-1} := \sigma(A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$ denote the learner's observable history. For the analysis, we define the enlarged sigma-field $\mathcal{G}_t := \sigma(\mathcal{H}_{t-1}, \pi_t, \nu_1, \dots, \nu_t)$, which additionally includes the unobserved common shifts. At each round, the learner chooses an $\mathcal{H}_{t-1}$-measurable distribution $\pi_t \in \Delta(K)$. After observing π_t, the environment chooses an *unobserved* scalar common shift $\nu_t \in \mathbb{R}$, possibly as an arbitrary function of the past and π_t, but before the current action is drawn. The learner then draws $A_t \sim \pi_t$ and observes a reward Y_t. Conditional on $\mathcal{G}_t$ and $A_t = a$, the reward follows a canonical exponential-family distribution with natural parameter $z_{t,a} = \nu_t + x_a^\top \theta^\star$, whose conditional density or mass function is

$$\Pr(y \mid \mathcal{G}_t, A_t = a) = \exp\big(y z_{t,a} - m(z_{t,a}) + h(y)\big). \tag{1}$$

[1] Without loss of generality, assume $\operatorname{span}\{x_a - x_b : a, b \in [K]\} = \mathbb{R}^d$; otherwise, redefine d accordingly.

Algorithm 1 Localized Elimination for Semiparametric GLBs (LE-SGLB)

Require: $\mathcal{X}, T, \delta, \rho, \kappa_{\text{init}}, M_\mu, L_\mu, R$

I. INITIAL LOCALIZATION

Tournament: Repeatedly run pairwise tests on the remaining arms until one arm $\widehat{a}$ remains.
Filtering: Compare $\widehat{a}$ with every other arm and remove those confidently worse than $\widehat{a}$.
Let $\mathcal{A}_1$ be the surviving set and set $\rho_1 \leftarrow \rho/4$.

II. LOCALIZED ELIMINATION

for $m = 1, 2, \ldots$ while rounds remain **do**
  if $|\mathcal{A}_m| = 1$ or $\rho_m \leqslant (L_\mu T)^{-1}$ **then**
    Play any arm in $\mathcal{A}_m$ for all remaining rounds and **stop**.
  end if
  Design: Compute a centered G-optimal design p_m and keep it fixed throughout phase m.
  Sampling: Sample $A_t \sim p_m$ and update $\widehat{w}_{m,n}$ and $\widehat{q}_m(n)$ until Equation 5 is satisfied.
  Elimination: Form $\mathcal{A}_{m+1}$ using Equation 6 and set $\rho_{m+1} \leftarrow \rho_m/4$.
end for

Here, $\theta^\star \in \mathbb{R}^d$ is an unknown parameter, m is the log-partition function, and h is determined by the exponential-family model. Writing $\mu(z) = m'(z)$ for the known mean function, we have $\mathbb{E}[Y_t \mid \mathcal{G}_t, A_t = a] = \mu(\nu_t + x_a^\top \theta^\star)$ and $\mathrm{Var}(Y_t \mid \mathcal{G}_t, A_t = a) = \mu'(\nu_t + x_a^\top \theta^\star)$.

The common shift ν_t is shared by all arms at round t, may vary arbitrarily over time, and is unobserved by the learner. We assume only that ν_t is selected before the current randomized action A_t is drawn. When μ is affine, the conditional mean induced by Equation 1 reduces to the semiparametric linear-bandit model (Greenewald et al., 2017; Krishnamurthy et al., 2018; Kim & Paik, 2019; Choi et al., 2023; Kim et al., 2025). When $\nu_t \equiv 0$, Equation 1 reduces to the standard generalized linear bandit model (Filippi et al., 2010; Li et al., 2017; Faury et al., 2020; Abeille et al., 2021).

Assumption 1 (Boundedness)**.** *For every $a \in [K]$, $\|x_a\|_2 \leqslant 1$ and $\|\theta^\star\|_2 \leqslant S$. Let $\mathcal{Z}_T := \bigcup_{t=1}^T (\nu_t + [-S, S])$. For some known $M_\mu > 0$, $|\mu(z)| \leqslant M_\mu$ for every $z \in \mathcal{Z}_T$.*

Assumption 2 (Smoothness)**.** *The natural-parameter domain is $\mathbb{R}$, and μ is twice continuously differentiable on $\mathbb{R}$. For some known $L_\mu > 0$, $0 < \mu'(z) \leqslant L_\mu$ for every $z \in \mathcal{Z}_T$.*

Assumption 3 (Self-concordance)**.** *The link function satisfies $|\mu''(z)| \leqslant R \cdot \mu'(z)$ for every $z \in \mathbb{R}$.*

Assumptions 1 and 2 are standard in the GLB literature (Filippi et al., 2010). We only require the conditional mean reward μ to be uniformly bounded, which is weaker than the bounded-reward assumptions in Filippi et al. (2010); Li et al. (2017); Sawarni et al. (2024). For unbounded links, these conditions implicitly restrict the allowable common shifts. Assumption 3 holds for many canonical GLMs, including logistic, Poisson, and Gaussian models with $R = 1, 1$, and 0, respectively, and follows from bounded rewards (Sawarni et al., 2024, Lemma 2.2). Moreover, without loss of generality, we assume $R \geqslant 1$, since any smaller valid value of R can be replaced by 1.

Optimal arm and Regret. We define an optimal arm as $a^\star \in \arg\max_{a \in [K]} x_a^\top \theta^\star$. Since μ is strictly increasing and the common shift ν_t is shared by all arms, we have $\arg\max_{a \in [K]} x_a^\top \theta^\star = \arg\max_{a \in [K]} \mu(\nu_t + x_a^\top \theta^\star)$ for every $t \in [T]$. Our goal is to minimize the cumulative regret $\textbf{Regret}(T) := \sum_{t=1}^T \left[\mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_{A_t}^\top \theta^\star)\right]$.

4 LOCALIZED ELIMINATION FOR SEMIPARAMETRIC GLBS

In this section, we present Localized Elimination for Semiparametric GLBs (LE-SGLB). The algorithm first uses randomized pairwise comparisons to obtain an active set with similar true scores (Subsection 4.1). It then proceeds in phases, keeping a centered experimental design fixed within each phase (Subsection 4.2). This makes the unknown time-varying slopes enter all arm comparisons through a common positive scale, preserving the *ranking* of the arms. Rather than estimating this scale, the algorithm collects samples until the comparisons are reliable enough for elimination. This coupling gives the κ-free leading regret after initialization. Algorithm 1 summarizes the procedure.

4.1 STEP 1. INITIAL LOCALIZATION

The goal of initialization is to construct an active set whose true scores $x_a^\top \theta^\star$ are close to one another. Write $W(\mathcal{A}) := \max_{a,b\in\mathcal{A}} |(x_a - x_b)^\top \theta^\star|$ for the score width, and set $\rho = 1/(64R\sqrt{d})$. Specifically, we seek to retain an optimal arm while reducing this width to at most $\rho/4$.

Randomized pairwise comparisons. Rewards collected from two arms at different times cannot be compared directly because the common shift may change between observations. For a pair (i, j), we instead set $\pi_t(i) = \pi_t(j) = 1/2$ and, after ν_t is chosen, draw $B_t \in \{-1, 1\}$ uniformly. We play i if $B_t = 1$ and j otherwise, and record the signed reward $C_t = 2B_tY_t$. This gives an unbiased comparison under the same current shift:

$$\mathbb{E}[C_t \mid \mathcal{G}_t] = \mu(\nu_t + x_i^\top \theta^\star) - \mu(\nu_t + x_j^\top \theta^\star).$$

Since μ is strictly increasing, this conditional mean has the same sign as $(x_i - x_j)^\top \theta^\star$. Its magnitude may vary over time, but its sign remains unchanged, so averaging these conditional means preserves the direction of the comparison.

Assumption 4 (Initialization curvature)**.** *There exists a known constant $\kappa_{\text{init}} > 0$ such that $\mu'(\nu_t + u) \geqslant 1/\kappa_{\text{init}}$ for every $u \in [-S, S]$ and every initialization round t.*

This condition is required only during initialization, where it allows uncertainty in mean-reward comparisons to be translated into closeness of the underlying scores. Without such a lower bound, a nearly flat link can make a large score gap appear arbitrarily small in reward space.

Pairwise test. Given tolerance ε and failure probability δ_{pair}, collect $N_{\text{pair}} = \tilde{\mathcal{O}}\left(\frac{\kappa_{\text{init}}^2}{\varepsilon^2}\right)$ comparisons, suppressing fixed reward and link constants and lower-order terms. Index these comparison rounds by $s = 1, \ldots, N_{\text{pair}}$, and compute $\bar{C} = \frac{1}{N_{\text{pair}}} \sum_{s=1}^{N_{\text{pair}}} C_s$. The test selects i if $\bar{C} > \frac{\varepsilon}{2\kappa_{\text{init}}}$, selects j if $\bar{C} < -\frac{\varepsilon}{2\kappa_{\text{init}}}$, and otherwise leaves the comparison unresolved. Under Assumption 4, with probability at least $1 - \delta_{\text{pair}}$, any selected arm has a strictly larger true score than the other arm, while an unresolved comparison certifies $|(x_i - x_j)^\top \theta^\star| \leqslant \varepsilon$.

Tournament (Line 1) and filtering (Line 2). Both stages use the pairwise test described above, with failure probability $\delta_{\text{pair}} = \frac{\delta}{6K}$ for each test. The *tournament* repeatedly pairs the remaining arms and runs this test on each pair with tolerance $\varepsilon = \frac{\rho}{8\lceil \log_2 K \rceil}$. For each pair, we retain the arm selected by the test, or either arm if the comparison is unresolved (an unpaired arm advances without a test). We repeat until a single reference arm $\widehat{a}$ remains. On the event that all tests are correct, this arm satisfies $(x_{a^\star} - x_{\widehat{a}})^\top \theta^\star \leqslant \frac{\rho}{8}$.

The *filtering* stage then compares $\widehat{a}$ with every other arm in the original action set, again using the pairwise test above, now with tolerance $\varepsilon = \frac{\rho}{8}$. An arm is removed only if the test selects $\widehat{a}$ over it; otherwise, it is retained. Let $\mathcal{A}_1$ consist of $\widehat{a}$ and all arms retained by filtering. This yields the following guarantee.

Proposition 1 (Initial localization)**.** *Under the model assumptions and Assumption 4, with probability at least $1 - \frac{\delta}{3}$, whenever initialization completes, its output satisfies*

$$a^\star \in \mathcal{A}_1, \qquad W(\mathcal{A}_1) \leqslant \frac{\rho}{4}. \tag{2}$$

The initializer uses $\widetilde{\mathcal{O}}(\kappa_{\text{init}}^2 Kd)$ samples.

Details of the pairwise test and proofs of the tournament, filtering, and sample complexity guarantees are given in Appendix C. Appendix I presents improved initializers that achieve the same localization guarantee with smaller initialization regret. The subsequent main procedure needs only the localization guarantee in Equation 2 from initialization and requires no uniform lower bound on μ'.

4.2 STEP 2. LOCALIZED PHASED ELIMINATION

Phase m starts with $a^\star \in \mathcal{A}_m$ and $W(\mathcal{A}_m) \leqslant \rho_m$, where $\rho_1 = \rho/4$. We keep the design fixed within each phase so that the unknown time-varying slopes multiply a common covariance matrix.

Fixed-design comparisons (Lines 8–9). Choose a centered G-optimal design p_m over $\mathcal{A}_m$, following Kim et al. (2025), and define

$$\bar{x}_m = \sum_{a \in \mathcal{A}_m} p_m(a) x_a, \qquad z_{m,a} = x_a - \bar{x}_m, \qquad \Sigma_m = \sum_{a \in \mathcal{A}_m} p_m(a) z_{m,a} z_{m,a}^\top.$$

Such a design exists and controls estimation error across all active-arm differences (Lemma B.1). Index the observations within phase m by $s = 1, \dots, n$. We expand the mean function around $c_s := \nu_s + \bar{x}_m^\top \theta^\star$. Writing $\alpha_s := \mu'(c_s) > 0$, a first-order Taylor expansion gives $\mu(c_s + z_{m,a}^\top \theta^\star) = \mu(c_s) + \alpha_s z_{m,a}^\top \theta^\star + \mathcal{R}_{s,a}$, where $\mathcal{R}_{s,a}$ is the Taylor remainder. Localization ensures $|z_{m,a}^\top \theta^\star| \leqslant \rho_m$, so self-concordance controls this remainder relative to the first-order term.

Since $\pi_s = p_m$, we have $\mathbb{E}[z_{m,A_s} \mid \mathcal{G}_s] = 0$. The common term $\mu(c_s)$ therefore cancels, giving

$$\mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s] = \alpha_s \Sigma_m \theta^\star + r_s, \qquad r_s := \mathbb{E}[z_{m,A_s} \mathcal{R}_{s,A_s} \mid \mathcal{G}_s].$$

The slope α_s remains unknown. However, keeping Σ_m fixed throughout the phase makes the first-order terms accumulate as $(\sum_{s=1}^n \alpha_s)\Sigma_m \theta^\star$. This motivates the estimator

$$\widehat{w}_{m,n} = \Sigma_m^\dagger \sum_{s=1}^n z_{m,A_s} Y_s. \tag{3}$$

For every pair $a, b \in \mathcal{A}_m$, it gives

$$(x_a - x_b)^\top \widehat{w}_{m,n} \approx \left(\sum_{s=1}^n \alpha_s \right) (x_a - x_b)^\top \theta^\star,$$

where the approximation accounts for stochastic noise and the controlled nonlinear remainder. The accumulated slope $\sum_{s=1}^n \alpha_s > 0$ is positive and common to all arm comparisons, so the comparison targets preserve the true arm ranking. Thus, the algorithm can compare arms without recovering $\theta^\star$ or estimating the accumulated slope.

Observable uncertainty. Let $\widehat{G}_m(n) := \max_{a,b \in \mathcal{A}_m} (x_a - x_b)^\top \widehat{w}_{m,n}$ be the estimated width on the common scale, and let $\beta_m(n)$ denote the stochastic confidence radius in Equation D.7. Localization bounds the relative nonlinear error by $\eta_m := 2\sqrt{d}\,(e^{R\rho_m} - 1) \leqslant \frac{1}{64}$ (Lemma D.2).

By Corollary D.1, with probability at least $1 - \frac{\delta}{3}$, the following bounds hold simultaneously for every phase m, every sample count n reached in that phase, and every pair $a, b \in \mathcal{A}_m$:

$$\begin{aligned} &\left| (x_a - x_b)^\top \widehat{w}_{m,n} - \left(\sum_{s=1}^n \alpha_s \right) (x_a - x_b)^\top \theta^\star \right| \\ &\qquad \leqslant \beta_m(n) + \eta_m W(\mathcal{A}_m) \sum_{s=1}^n \alpha_s \leqslant \frac{\beta_m(n) + \eta_m \widehat{G}_m(n)}{1 - \eta_m} =: \widehat{q}_m(n). \end{aligned} \tag{4}$$

We denote this event by $\mathcal{E}_{\mathrm{conf}}$. The observable radius $\widehat{q}_m(n)$ therefore bounds all pairwise comparison errors on the accumulated common scale. The derivation is given in Appendix D.

Adaptive phase lengths (Line 9). We stop at the first n satisfying

$$\widehat{G}_m(n) \geqslant 64\, \beta_m(n). \tag{5}$$

This rule waits until the observable comparison signal dominates the stochastic confidence radius. On $\mathcal{E}_{\mathrm{conf}}$, the comparisons are then sufficiently accurate for safe elimination (Lemmas D.6 and D.7). Small slopes may prolong a phase, but they also shrink the reward gaps between active arms. This coupling is central to the κ-free leading regret guarantee proved below.

Elimination (Line 10). When the stopping condition is met, choose $\widehat{a}_m \in \arg\max_{a \in \mathcal{A}_m} x_a^\top \widehat{w}_{m,n}$ and retain the arms whose estimated scores are within $2\widehat{q}_m(n)$ of the largest estimated score:

$$\mathcal{A}_{m+1} = \left\{ a \in \mathcal{A}_m : (x_{\widehat{a}_m} - x_a)^\top \widehat{w}_{m,n} \leqslant 2\widehat{q}_m(n) \right\}. \tag{6}$$

On $\mathcal{E}_{\mathrm{conf}}$, Lemma D.7 guarantees $a^\star \in \mathcal{A}_{m+1}$ and $W(\mathcal{A}_{m+1}) \leqslant \frac{W(\mathcal{A}_m)}{4}$. We then set $\rho_{m+1} = \frac{\rho_m}{4}$ and proceed to the next phase, subject to the termination conditions in Algorithm 1. All sampling stops when the horizon T is reached.

5 MAIN RESULTS

In this section, we first prove an upper bound on the regret of LE-SGLB (Subsection 5.1) and then establish a matching lower bound (Subsection 5.2).

5.1 REGRET UPPER BOUND FOR LE-SGLB

Theorem 1 (Regret of LE-SGLB)**.** *Suppose Assumptions 1–4 hold. Set $\rho = 1/(64R\sqrt{d})$. Then, with probability at least $1-\delta$, Algorithm 1 satisfies*

$$\textbf{Regret}(T) = \widetilde{\mathcal{O}}\Big(\kappa_{\text{init}}^2 Kd + \sqrt{dT\min\{d, \log K\}} + d\min\{d, \log K\}\Big).$$

Discussion of Theorem 1. The proof is provided in Appendix E. The leading round-dependent regret is $\widetilde{\mathcal{O}}(\sqrt{dT\min\{d, \log K\}})$ and, importantly, is independent of both the magnitude and variation of the common shifts $(\nu_t)_{t=1}^T$ and the global inverse-curvature parameter $\kappa = (\inf_{z\in\mathcal{Z}_T}\mu'(z))^{-1}$. The factor $\min\{d, \log K\}$ follows from taking the sharper of the concentration bounds over active-arm pairs and over the active difference span. After initialization, the algorithm neither requires a positive global lower bound on μ' nor knowledge of κ. The only curvature requirement appears in the one-time initialization cost $\widetilde{\mathcal{O}}(\kappa_{\text{init}}^2 Kd)$, which is used to reach the localized regime and does not scale with T, up to logarithmic factors. To the best of our knowledge, this is the first SGLB regret guarantee whose leading round-dependent term is independent of both the scale of the common shifts and κ.

Remark 1 (Why not reduce to a semiparametric linear bandit?)**.** *A careful reader may wonder whether SGLBs could be reduced directly to semiparametric linear bandits by simply paying a factor of κ. Although the regret can indeed be related to the corresponding linear-score regret, the hidden shift enters inside the nonlinear link, so neither feature centering nor a standard nuisance-free GLM likelihood recovers the estimating structure required by existing analyses. See Appendix H for details.*

Remark 2 (Improved initial localization)**.** *The initialization cost in Theorem 1 can be reduced from $\widetilde{\mathcal{O}}(\kappa_{\text{init}}^2 Kd)$ to $\widetilde{\mathcal{O}}\big(\min\{\kappa_{\text{init}}^2 K\sqrt{d},\, \kappa_{\text{init}}^4 d^{9/2}\}\big)$ without changing the subsequent elimination procedure. See Appendix I for the algorithms and formal guarantees.*

Proof sketch of Theorem 1. We use $\rho = 1/(64R\sqrt{d})$ and divide the proof into four steps.

1. Initial localization. The initialization procedure returns an active set satisfying $a^\star \in \mathcal{A}_1$ and $W(\mathcal{A}_1) \leqslant \rho/4$ with high probability. Moreover, since its number of rounds is $N_{\text{init}}(\rho, \delta/3)$ and the instantaneous regret is at most $2M_\mu$, its regret is $\widetilde{\mathcal{O}}(\kappa_{\text{init}}^2 Kd)$ (Proposition C.1).

2. Confidence bound. Suppose phase m starts with $a^\star \in \mathcal{A}_m$ and $W(\mathcal{A}_m) \leqslant \rho_m$. Define the true scaled score width $G_m(n) := (\sum_{s=1}^n \alpha_s) W(\mathcal{A}_m)$. The uniform pairwise confidence bound implies

$$\left|\widehat{G}_m(n) - G_m(n)\right| \leqslant \beta_m(n) + \eta_m G_m(n),$$

where $\eta_m \leqslant 1/64$ and $\beta_m(n) = \widetilde{\mathcal{O}}\big(\sqrt{dn\min\{d, \log K\}} + d\min\{d, \log K\}\big)$ (Corollary D.1 and Lemma D.5). Before stopping, $\widehat{G}_m(n) < 64\beta_m(n)$, and hence $(1-\eta_m)G_m(n) \leqslant \widehat{G}_m(n) + \beta_m(n) < 65\beta_m(n)$, which gives $G_m(n) \leqslant 67\beta_m(n)$. Thus, for a phase ending after n_m samples,

$$G_m(n_m) \leqslant 67\beta_m(n_m) + L_\mu \rho_m. \qquad \text{(Lemma D.8)}$$

3. Width contraction. When the stopping rule is met, we have $\widehat{q}_m(n) \leqslant G_m(n)/16$ (Lemma D.6). For every retained arm $a \in \mathcal{A}_{m+1}$, the observable confidence bound and the retention rule in Equation 6 give $(\sum_{s=1}^n \alpha_s)\,(x_{a^\star} - x_a)^\top \theta^\star \leqslant 3\widehat{q}_m(n) \leqslant \frac{3}{16} G_m(n)$. Since $a^\star \in \mathcal{A}_{m+1}$ and $G_m(n) = (\sum_{s=1}^n \alpha_s) W(\mathcal{A}_m)$, taking the maximum over retained arms yields

$$W(\mathcal{A}_{m+1}) \leqslant \frac{3}{16} W(\mathcal{A}_m) < \frac{1}{4} W(\mathcal{A}_m) \leqslant \rho_{m+1}. \qquad \text{(Lemma D.7)}$$

4. Regret aggregation. The same local slopes that govern the comparison signal also scale the instantaneous reward gaps. Since all active-arm scores lie within ρ_m of $\bar{x}_m^\top \theta^\star$, self-concordance gives, for every $a \in \mathcal{A}_m$, we have

$$\mu(\nu_s + x_{a^\star}^\top \theta^\star) - \mu(\nu_s + x_a^\top \theta^\star) \leqslant e^{R\rho_m} \alpha_s (x_{a^\star} - x_a)^\top \theta^\star \leqslant e^{R\rho_m} \alpha_s W(\mathcal{A}_m).$$

Thus, small slopes make the comparison signal accumulate slowly, but also reduce the regret incurred in each sampling round. Summing these reward gaps shows that phase regret is controlled by the same accumulated signal $G_m(n_m)$ that appears in Step 2. Specifically, let Reg_m denote the cumulative regret incurred during phase m. Then, we have

$$\begin{aligned}\mathrm{Reg}_m &\leqslant e^{R\rho_m}\left(\sum_{s=1}^{n_m}\alpha_s\right)W(\mathcal{A}_m) = e^{R\rho_m}G_m(n_m) && \text{(Lemma D.9)}\\ &\leqslant e^{R\rho_m}\big(67\beta_m(n_m) + L_\mu\rho_m\big). && \text{(Lemma D.8)}\end{aligned}$$

The last bound also holds when the phase is truncated by the horizon. Since $e^{R\rho_m} = \mathcal{O}(1)$ and $\rho_m = \rho/4^m$ decreases geometrically, summing over phases gives

$$\sum_m \mathrm{Reg}_m = \widetilde{\mathcal{O}}\Big(\sum_m \beta_m(n_m)\Big) = \widetilde{\mathcal{O}}\Big(\sqrt{dT\min\{d,\log K\}} + d\min\{d,\log K\}\Big). \quad \text{(Lemma E.1)}$$

Combining it with the initialization regret proves Theorem 1.

5.2 A LOWER BOUND UNDER LOCAL CURVATURE

We next ask whether the favorable local-curvature improvement established for ordinary logistic bandits (Abeille et al., 2021; Faury et al., 2022; Zhang & Sugiyama, 2024; Lee et al., 2024) continues to hold in the semiparametric model. Throughout this subsection, let $\sigma(z) = (1+e^{-z})^{-1}$ denote the logistic mean function, and define $\kappa^\star := \big(\frac{1}{T}\sum_{t=1}^{T}\sigma'(\nu_t + x_{a^\star}^\top\theta^\star)\big)^{-1}$. This is the inverse average slope at the optimal arm. In ordinary logistic bandits, local curvature can yield a $1/\sqrt{\kappa^\star}$ improvement in the leading statistical regret. The following result shows that this improvement does not extend in general to our setting with unobserved time-varying common shifts.

Theorem 2 (Lower bound)**.** *There exist universal constants $c_0, C > 0$ such that the following holds. For every $d \geqslant 1$, $T \geqslant Cd^2$, $S \geqslant 1/8$, every $4 \leqslant \kappa_0 \leqslant c_0\sqrt{T}/d$, and every policy, there exists a logistic semiparametric bandit instance on the $K = 2^d$ hypercube action set, with $\|\theta^\star\|_2 \leqslant S$ and a deterministic common-shift sequence, such that $\kappa^\star = \kappa_0$ and $\mathbb{E}[\mathbf{Regret}(T)] = \Omega\big(d\sqrt{T}\big)$.*

Discussion of Theorem 2. The proof is provided in Appendix G. Since $K = 2^d$, we have $\log K = \Theta(d)$, so the lower bound $d\sqrt{T} = \Theta(\sqrt{dT\log K})$ matches the d- and T-dependence of the leading term in Theorem 1, up to logarithmic factors. More importantly, for every prescribed $4 \leqslant \kappa_0 \leqslant c_0\sqrt{T}/d$, there exists an instance with $\kappa^\star = \kappa_0$ and regret $\Omega(d\sqrt{T})$. Hence, throughout this range, increasing $\kappa^\star$ does not reduce the worst-case statistical difficulty of the semiparametric problem. In particular, the favorable $d\sqrt{T/\kappa^\star}$ statistical dependence established for ordinary logistic bandits (Abeille et al., 2021) cannot hold uniformly over our semiparametric class.

The reason for this difference lies in how local curvature affects the statistical distinguishability provided by each reward observation. Write $\alpha = 1/\kappa^\star$. In an ordinary logistic bandit, a score perturbation of size Δ changes the mean by order $\alpha\Delta$, while the Bernoulli variance is also of order α, leading to a one-step KL divergence of order $\alpha\Delta^2$. In our hard construction, averaging over the unobserved shift sign yields the observable mean $q_B(u) = \{\sigma(B+u) + \sigma(-B+u)\}/2$, for which $q_B(0) = 1/2$ and $q_B'(0) = \alpha$. Thus, a score perturbation still changes the observable mean by order $\alpha\Delta$, while the observable variance $q_B(u)(1-q_B(u)) = \Theta(1)$ over the relevant score range. Consequently, the one-step KL divergence scales as $\alpha^2\Delta^2$, smaller by a factor of α than in the ordinary logistic model. This loss of observable information eliminates the usual $1/\sqrt{\kappa^\star}$ improvement and ultimately yields the $\Omega(d\sqrt{T})$ worst-case regret in Theorem 2 (refer Appendix G.3).

6 NUMERICAL EXPERIMENTS

Setup. We consider a fixed action set with $K = 20$ and $d = 5$, set $\|\theta^\star\|_2 = S = 3$, and run each algorithm for $T = 30{,}000$ rounds. Rewards are Bernoulli with mean $\mu(x_a^\top\theta^\star + \nu_t)$, where $\nu_t \in [-B, B]$ is an unobserved common shift with $B = 5$. We evaluate both the logistic link $\mu(z) = (1+e^{-z})^{-1}$ and the probit link $\mu(z) = \Phi(z)$. We report mean cumulative regret over 20 independent seeds. We consider three common-shift environments. 1) In *fixed-period Beta*, ν_t follows

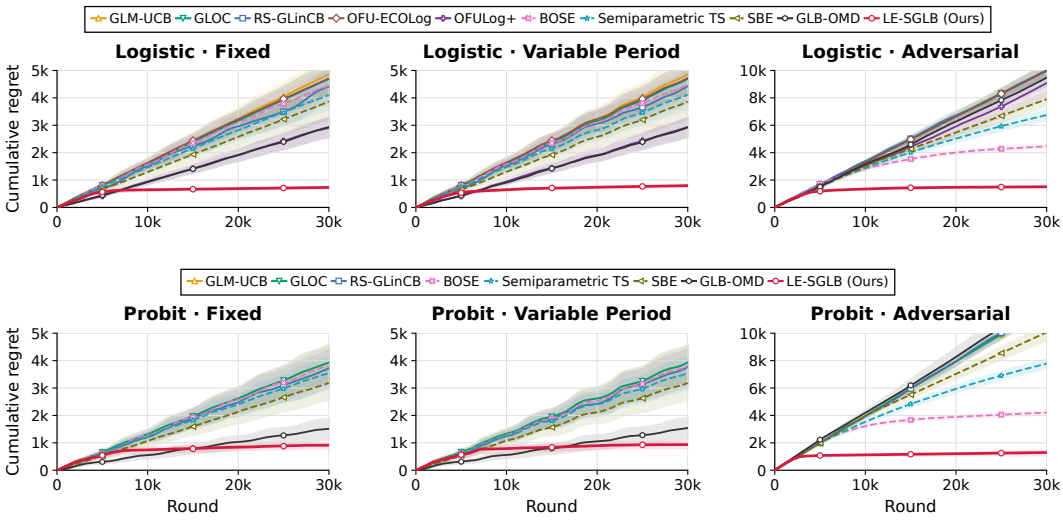

Figure 1: Cumulative regret over $T = 30{,}000$ rounds under the three common-shift environments. Curves show the mean over 20 seeds and shaded regions indicate one standard error.

a periodic trajectory obtained by transforming a triangular wave through the $\mathrm{Beta}(2, 10)$ quantile function and rescaling it to $[-B, B]$, with period $5{,}000$. 2) *Variable-period Beta* uses the same $\mathrm{Beta}(2, 10)$-based construction, but independently varies each cycle length between $4{,}000$ and $6{,}000$ rounds. Both processes are generated independently of the learner. 3) In the *adversarial environment*, after observing the learner's current action distribution p_t, the adversary chooses a common shift $\nu_t \in [-B, B]$ before the action is sampled so that the learner's marginal expected reward is $1/2$. Further details of the nuisance constructions and the adversary are provided in Appendix J.

Baselines. We compare against representative GLB algorithms, `GLM-UCB` (Filippi et al., 2010), `GLOC` (Jun et al., 2017), `RS-GLinCB` (Sawarni et al., 2024), and `GLB-OMD` (Zhang et al., 2025); logistic-bandit algorithms, `OFU-ECOLog` (Faury et al., 2022) and `OFULog+` (Lee et al., 2024); and semiparametric linear-bandit methods, `BOSE` (Krishnamurthy et al., 2018), `Semiparametric TS` (Kim & Paik, 2019), and `SBE` (Kim et al., 2025). `OFU-ECOLog` and `OFULog+` are specific to the logistic link and are therefore included only in the logistic experiments. For each algorithm, we tune an exploration coefficient $c \in \{0.05, 0.1, 0.5, 1\}$, which controls the exploration strength (e.g., the confidence radius or phase sample budget), and report the best-performing value. For `LE-SGLB`, we cap the initialization budget at $2{,}500$ rounds, since the theoretical initialization budget is overly conservative in practice.

Results. Figure 1 shows that `LE-SGLB` consistently and substantially outperforms all baselines across both logistic and probit links and all three nuisance environments. The results also illustrate the benefit of avoiding global curvature dependence. With $B = 5$ and $S = 3$, the natural parameter lies in $[-8, 8]$. Hence, the global inverse-curvature parameters are $\kappa = 1/\sigma'(8) \approx 2.98 \times 10^3$ for the logistic link and $\kappa = 1/\Phi'(8) \approx 1.98 \times 10^{14}$ for the probit link. Despite these extremely large curvature parameters, `LE-SGLB` remains effective across both settings. Moreover, as shown in Figure J.1, increasing $\kappa^\star$ does not lead to a corresponding improvement in empirical performance. This is consistent with our lower-bound result (Theorem 2), which shows that the favorable $1/\sqrt{\kappa^\star}$ dependence of ordinary logistic bandits does not extend to SGLBs.

7 Conclusion

In this paper, we studied finite-armed semiparametric generalized linear bandits with unobserved, time-varying common shifts and proposed `LE-SGLB`, which learns without estimating the nuisance shifts. `LE-SGLB` achieves a leading regret of $\widetilde{\mathcal{O}}(\sqrt{dT \min\{d, \log K\}})$ that is independent of the global inverse-curvature parameter κ and the common-shift scale, together with a matching lower bound. To the best of our knowledge, this is the first regret guarantee for generalized linear bandits with unobserved, time-varying common shifts.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, including polishing language and drafting portions of the manuscript. We did not use generative AI to develop the core research ideas, formulate the main theoretical contributions, determine the proof strategies, or shape the main scientific arguments of the paper. We also used generative AI for auxiliary tasks, including rapid checks of mathematical derivations and proofs, improving notation and presentation, and assisting with the implementation and debugging of experimental code. We carefully reviewed all AI-assisted work. All mathematical arguments and proofs were independently verified by the authors, all AI-assisted writing was reviewed and revised as needed, and all AI-assisted code was inspected and tested for correctness. We take full responsibility for the final content of this work, including all text, claims, proofs, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the complete assumptions, algorithmic details, and theoretical statements in the main paper, together with full proofs of all results in the appendix. The appendix also contains detailed descriptions of the initialization procedures, confidence bounds, lower-bound construction, and additional technical lemmas needed to reproduce the theoretical analysis. For the numerical experiments, we provide the full experimental setup, nuisance-process constructions, baseline implementations, and hyperparameter tuning protocol in Appendix J. Source code for reproducing all experiments and figures is included in the supplementary material.

REFERENCES

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3691–3699. PMLR, 2021.

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.

Eugenio Clerico, Hamish Flynn, Gergely Neu, et al. Confidence sequences for generalized linear models via regret analysis. *arXiv preprint arXiv:2504.16555*, 2025.

Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *Causal Learning and Reasoning*, pp. 1033–1064. PMLR, 2024.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pp. 3052–3060. PMLR, 2020.

Louis Faury, Marc Abeille, Kwang-Sung Jun, and Clément Calauzènes. Jointly efficient and optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 546–580. PMLR, 2022.

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.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 586–594, Red Hook, NY, USA, 2010. Curran Associates Inc.

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

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

Kwang-Sung Jun, Aniruddha Bhargava, Robert Nowak, and Rebecca Willett. Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems*, 30, 2017.

Yue Kang, Mingshuo Liu, Bongsoo Yi, Jing Lyu, Zhi Zhang, Doudou Zhou, and Yao Li. Single index bandits: Generalized linear contextual bandits with unknown reward functions. In *International Conference on Learning Representations*, volume 2026, pp. 11062–11101, 2026.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

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

Seok-Jin Kim. Nearly optimal best arm identification for semiparametric bandits. *arXiv preprint arXiv:2604.03969*, 2026.

Seok-Jin Kim, Gi-Soo Kim, and Min-hwan Oh. Experimental design for semiparametric bandits. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 3215–3252. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/kim25a.html.

Johannes Kirschner and Andreas Krause. Bias-robust bayesian optimization via dueling bandits. In *International Conference on Machine Learning*, pp. 5595–5605. PMLR, 2021.

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

Joongkyu Lee and Min-hwan Oh. Nearly minimax optimal regret for multinomial logistic bandit. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Joongkyu Lee and Min-hwan Oh. Improved online confidence bounds for multinomial logistic bandits. In *Forty-second International Conference on Machine Learning*, 2025.

Joongkyu Lee and Min-hwan Oh. Nonstationary generalized linear bandits with discounted online mirror descent. *arXiv preprint arXiv:2605.25590*, 2026.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. Improved regret bounds of (multinomial) logistic bandits via regret-to-confidence-set conversion. In *International Conference on Artificial Intelligence and Statistics*, pp. 4474–4482. PMLR, 2024.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Licong Lin, Koulik Khamaru, and Martin J Wainwright. Semiparametric inference based on adaptively collected data. *The Annals of Statistics*, 53(3):989–1014, 2025.

Ayush Sawarni, Nirjhar Das, Siddharth Barman, and Gaurav Sinha. Generalized linear bandits with limited adaptivity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FTPDBQuT4G.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in neural information processing systems*, 27, 2014.

Yu-Jie Zhang and Masashi Sugiyama. Online (multinomial) logistic bandit: Improved regret and constant computation cost. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized linear bandits: Almost optimal regret with one-pass update. *arXiv preprint arXiv:2507.11847*, 2025.

Appendix

Table of Contents

A Notation **12**

B Centered G-Optimal Design **12**

C Proof of the Initial Localization Guarantee **14**

D Analysis of a Localized Elimination Phase **16**
D.1 Localized Taylor error 17
D.2 Confidence bounds 18
D.3 Observable confidence radius and elimination 21
D.4 Regret in a localized phase 22

E Proof of the Regret Upper Bound **24**

F Technical Lemmas for the Upper Bound **27**
F.1 Self-concordance and localization 27
F.2 Martingale inequalities 28

G Proof of the Regret Lower Bound **29**
G.1 Hard instance construction. 29
G.2 Proof of Theorem 2 30
G.3 Comparison with ordinary logistic bandits 33

H Why a direct reduction to semiparametric linear bandits fails **34**

I Improved Initial Localization **37**
I.1 Successive tournament initialization 38
I.2 Geometric initialization 40
I.3 Guarantee for geometric initialization 41
I.4 Proofs of the geometric initialization guarantees 43

J Experiment Details and Additional Results **48**
J.1 Experimental setup 48
J.2 Additional results: varying curvature 49

A Notation

We summarize the notation used in the appendix. Throughout, t denotes a global interaction round, while s denotes a locally reindexed sample within a fixed phase or pairwise test.

B Centered G-Optimal Design

The covariance Σ_m may be singular when the active arms lie in a proper affine subspace. This is harmless because elimination only uses directions of the form $x_a - x_b$.

Lemma B.1 (Existence of a centered G-optimal design)**.** *For every finite active set $\mathcal{A} \subseteq [K]$, there exists a distribution p over $\mathcal{A}$ such that, with*

$$\bar{x} = \sum_{a \in \mathcal{A}} p(a) x_a, \qquad z_a = x_a - \bar{x}, \qquad \Sigma = \sum_{a \in \mathcal{A}} p(a) z_a z_a^\top,$$

Table A.1: Summary of notation

Symbol	Definition	Meaning
Problem parameters		
S	$\|\theta^\star\|_2 \leqslant S$	Parameter norm bound.
$M_\mu,\ L_\mu$	$\|\mu(z)\| \leqslant M_\mu, \quad 0 < \mu'(z) \leqslant L_\mu, \quad z \in \mathcal{Z}_T$	Mean and variance bounds.
R	$\|\mu''(z)\| \leqslant R\mu'(z), \quad z \in \mathbb{R}$	Self-concordance coefficient.
Initial localization		
κ_{init}	$\mu'(\nu_t + u) \geqslant \dfrac{1}{\kappa_{\text{init}}}, \quad u \in [-S, S]$	Initialization curvature.
$B_s,\ C_s$	$B_s \sim \text{Rad}(1/2), \qquad C_s = 2B_sY_s$	Randomized pairwise comparison.
$N_{\text{pair}}(\varepsilon, \delta_{\text{pair}})$	$\left\lceil \left(\dfrac{1024(L_\mu + M_\mu^2)\kappa_{\text{init}}^2}{\varepsilon^2} + \dfrac{32(R + M_\mu + \sqrt{L_\mu})\kappa_{\text{init}}}{\varepsilon} \right) \log \dfrac{2}{\delta_{\text{pair}}} \right\rceil$	Samples per pairwise test.
$\rho,\ \rho_1$	$\rho = \dfrac{1}{64R\sqrt{d}}, \qquad \rho_1 = \dfrac{\rho}{4}$	Initial localization radii.
Localized elimination		
$W(\mathcal{A}_m)$	$\max\limits_{a,b \in \mathcal{A}_m} \left\lvert (x_a - x_b)^\top \theta^\star \right\rvert$	Active-arm score width.
ρ_m	$W(\mathcal{A}_m) \leqslant \rho_m, \qquad \rho_{m+1} = \rho_m/4$	Width bound.
$\bar{x}_m,\ z_{m,a}$	$\bar{x}_m = \sum\limits_{a \in \mathcal{A}_m} p_m(a) x_a, \qquad z_{m,a} = x_a - \bar{x}_m$	Centered features.
Σ_m	$\Sigma_m = \sum\limits_{a \in \mathcal{A}_m} p_m(a) z_{m,a} z_{m,a}^\top$	Centered design covariance.
p_m	$p_m \in \operatorname*{arg\,min}\limits_{\substack{p \in \Delta(\mathcal{A}_m) \\ \text{range}(\Sigma(p)) = \mathcal{V}_m}} \max\limits_{a \in \mathcal{A}_m} \|x_a - \bar{x}(p)\|^2_{\Sigma(p)^\dagger}$	Centered G-optimal design, $\mathcal{V}_m := \text{span}\{x_a - x_b : a, b \in \mathcal{A}_m\}$.
$c_s,\ \alpha_s$	$c_s = \nu_s + \bar{x}_m^\top \theta^\star, \qquad \alpha_s = \mu'(c_s)$	Local expansion point and slope.
$\mathcal{R}_{s,a}$	$\mu(c_s + z_{m,a}^\top \theta^\star) - \mu(c_s) - \alpha_s z_{m,a}^\top \theta^\star$	Taylor remainder.
r_s	$r_s = \mathbb{E}[z_{m,A_s} \mathcal{R}_{s,A_s} \mid \mathcal{G}_s]$	Centered Taylor remainder.
$\widehat{w}_{m,n}$	$\Sigma_m^\dagger \sum\limits_{s=1}^{n} z_{m,A_s} Y_s$	Pairwise-score estimator.
$\widehat{G}_m(n)$	$\max\limits_{a,b \in \mathcal{A}_m} (x_a - x_b)^\top \widehat{w}_{m,n}$	Estimated scaled score width.
η_m	$2\sqrt{d}\left(e^{R\rho_m} - 1\right)$	Relative nonlinear error.
$\widehat{q}_m(n)$	$\dfrac{\beta_m(n) + \eta_m \widehat{G}_m(n)}{1 - \eta_m}$	Observable confidence radius.
Confidence and proof quantities		
D_s	$z_{m,A_s} Y_s - \mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s]$	Vector martingale difference.
$\beta_{K,m}(n)$	$\sqrt{64(L_\mu + M_\mu^2) dn \log\left(\dfrac{12T^2\lvert\mathcal{A}_m\rvert^2}{\delta}\right)} + 8(R + M_\mu + \sqrt{L_\mu}) d \log\left(\dfrac{12T^2\lvert\mathcal{A}_m\rvert^2}{\delta}\right)$	Pairwise confidence radius.
$\beta_{d,m}(n)$	$4\sqrt{16(L_\mu + M_\mu^2) dn \left(d \log 5 + \log \dfrac{12T^2}{\delta}\right)} + 16(R + M_\mu + \sqrt{L_\mu}) d \left(d \log 5 + \log \dfrac{12T^2}{\delta}\right)$	Span-wise confidence radius.
$\beta_m(n)$	$\min\{\beta_{K,m}(n), \beta_{d,m}(n)\}$	Confidence radius for `LE-SGLB`.
$G_m(n)$	$\left(\sum\limits_{s=1}^{n} \alpha_s\right) W(\mathcal{A}_m)$	True scaled score width.

we have $\text{range}(\Sigma) = \text{span}\{x_a - x_b : a, b \in \mathcal{A}\}$ *and* $\max_{a \in \mathcal{A}} z_a^\top \Sigma^\dagger z_a \leqslant d$.

Proof of Lemma B.1. Let $r = \dim \text{span}\{x_a - x_b : a, b \in \mathcal{A}\}$. If $r = 0$, the claim is trivial. Otherwise, fix $a_0 \in \mathcal{A}$ and let $Q \in \mathbb{R}^{d \times r}$ have orthonormal columns spanning the active-arm difference space. We write

$$y_a = Q^\top (x_a - x_{a_0}), \qquad \widetilde{y}_a = \begin{bmatrix} 1 \\ y_a \end{bmatrix}.$$

Let $p^\star$ be a D-optimal design for the augmented features $\widetilde{y}_a$. By the Kiefer–Wolfowitz equivalence theorem (Kiefer & Wolfowitz, 1960), we have

$$\widetilde{y}_a^\top \mathbf{M}^{-1} \widetilde{y}_a \leqslant r + 1, \qquad \text{where } \mathbf{M} = \sum_a p^\star(a) \widetilde{y}_a \widetilde{y}_a^\top = \begin{bmatrix} 1 & \bar{y}^\top \\ \bar{y} & \Sigma_y + \bar{y}\bar{y}^\top \end{bmatrix}.$$

Applying the Schur complement to $\mathbf{M}$ yields

$$(y_a - \bar{y})^\top \Sigma_y^{-1} (y_a - \bar{y}) \leqslant r \leqslant d.$$

Since $x_a - \bar{x} = Q(y_a - \bar{y})$ and $\Sigma = Q\Sigma_y Q^\top$, it follows that

$$\mathrm{range}(\Sigma) = \mathrm{span}\{x_a - x_b : a, b \in \mathcal{A}\},$$

and

$$z_a^\top \Sigma^\dagger z_a = (y_a - \bar{y})^\top \Sigma_y^{-1} (y_a - \bar{y}) \leqslant d.$$

This proves the lemma. □

Remark B.1 (Approximate design)**.** *For simplicity, we state the algorithm using an exact centered G-optimal design. Exact optimality, however, is not essential. The analysis only requires that the design span all active-arm differences and satisfy* $\max_{a \in \mathcal{A}_m} z_{m,a}^\top \Sigma_m^\dagger z_{m,a} \leqslant C \cdot d$ *for some universal constant* C*. Such a constant-factor guarantee changes only numerical constants in the confidence radii and regret bound. Thus, in practice, one may use an approximate design satisfying the same leverage guarantee, as is standard in experimental-design-based bandit algorithms (Soare et al., 2014; Fiez et al., 2019; Kim et al., 2025).*

C Proof of the Initial Localization Guarantee

In this section, we first formalize the randomized pairwise test in Step 1 (Subsection 4.1) and then prove the tournament and filtering guarantees.

For a tolerance $\varepsilon > 0$ and pairwise failure probability $\delta_{\mathrm{pair}} \in (0, 1)$, set

$$N_{\mathrm{pair}}(\varepsilon, \delta_{\mathrm{pair}}) := \left\lceil \left(\frac{1024(L_\mu + M_\mu^2)\kappa_{\mathrm{init}}^2}{\varepsilon^2} + \frac{32(R + M_\mu + \sqrt{L_\mu})\kappa_{\mathrm{init}}}{\varepsilon} \right) \log \frac{2}{\delta_{\mathrm{pair}}} \right\rceil. \quad \text{(C.1)}$$

For a pair (i, j), let $N = N_{\mathrm{pair}}(\varepsilon, \delta_{\mathrm{pair}})$ and index the comparison rounds by $s = 1, \ldots, N$. Collect the signed rewards $C_s = 2B_s Y_s$ and let $\overline{C} := \frac{1}{N} \sum_{s=1}^N C_s$ denote their empirical average. We select i if $\overline{C} > \varepsilon/(2\kappa_{\mathrm{init}})$ and select j if $\overline{C} < -\varepsilon/(2\kappa_{\mathrm{init}})$. If neither condition holds, we leave the comparison unresolved.

Lemma C.1 (Correctness of the randomized pairwise test)**.** *Under Assumptions 1–4, with probability at least* $1 - \delta_{\mathrm{pair}}$*, the randomized pairwise test satisfies:*

(i) the selected arm has a strictly larger linear score than the other arm: if the test selects $a \in \{i, j\}$ *and* a' *denotes the other arm, then* $(x_a - x_{a'})^\top \theta^\star > 0$*;*

(ii) if the comparison remains unresolved, then $|(x_i - x_j)^\top \theta^\star| \leqslant \varepsilon$*.*

Consequently, whenever $|(x_i - x_j)^\top \theta^\star| > \varepsilon$*, the test selects the arm with the larger linear score.*

Proof of Lemma C.1. Let $\Delta_{i,j} = (x_i - x_j)^\top \theta^\star$ and $m_s = \mathbb{E}[C_s \mid \mathcal{G}_s]$. By the mean-value theorem,

$$m_s = \alpha_{s,ij}(x_i - x_j)^\top \theta^\star = \alpha_{s,ij}\Delta_{i,j} \qquad \text{and} \qquad \alpha_{s,ij} \geqslant \frac{1}{\kappa_{\mathrm{init}}}. \qquad \text{(Assumption 4)}$$

Since $\alpha_{s,ij} > 0$, all m_s have the same sign as $\Delta_{i,j}$ and the following holds:

$$\left| \frac{1}{N} \sum_{s=1}^N m_s \right| \geqslant \frac{|\Delta_{i,j}|}{\kappa_{\mathrm{init}}}. \quad \text{(C.2)}$$

By Lemma F.5 with weights 2 and -2 (and hence $H = 2$ and $V = 4$), with probability at least $1 - \delta_{\mathrm{pair}}$,

$$\left| \overline{C} - \frac{1}{N} \sum_{s=1}^N m_s \right| \leqslant \sqrt{\frac{64(L_\mu + M_\mu^2)\log(2/\delta_{\mathrm{pair}})}{N}} + \frac{8(R + M_\mu + \sqrt{L_\mu})\log(2/\delta_{\mathrm{pair}})}{N}.$$

The two terms are each at most $\varepsilon/(4\kappa_{\text{init}})$ under Equation C.1. Therefore,

$$\left|\overline{C} - \frac{1}{N}\sum_{s=1}^{N} m_s\right| \leqslant \frac{\varepsilon}{2\kappa_{\text{init}}}.$$

On this event, if the test selects i (i.e., $\overline{C} > \varepsilon/(2\kappa_{\text{init}})$), then $\Delta_{i,j} > 0$, and if it selects j (i.e., $\overline{C} < -\varepsilon/(2\kappa_{\text{init}})$), then $\Delta_{i,j} < 0$. This proves (i). If the comparison remains unresolved, then $|\overline{C}| \leqslant \varepsilon/(2\kappa_{\text{init}})$, so

$$\left|\frac{1}{N}\sum_{s=1}^{N} m_s\right| \leqslant \frac{\varepsilon}{\kappa_{\text{init}}}.$$

Combining this with Equation C.2 gives $|\Delta_{i,j}| \leqslant \varepsilon$, proving (ii). □

Lemma C.2 (Tournament guarantee)**.** *Run the tournament in Algorithm 1 using tolerance ε in every pairwise test. For each pair, advance the selected arm if the test selects one; otherwise, advance either arm. If all pairwise tests satisfy Lemma C.1, then the winner of the tournament $\widehat{a}$ satisfies*

$$x_{a^\star}^\top \theta^\star - x_{\widehat{a}}^\top \theta^\star \leqslant \lceil \log_2 K \rceil \varepsilon.$$

Proof of Lemma C.2. Follow the path from an optimal arm through the tournament. If a comparison selects one arm, Lemma C.1 ensures that the selected arm has the larger linear score. If the comparison is unresolved, either arm may advance, but their scores differ by at most ε. Since the tournament has at most $\lceil \log_2 K \rceil$ levels, the total possible decrease in linear score is at most $\lceil \log_2 K \rceil \varepsilon$. □

Lemma C.3 (Filtering guarantee)**.** *Suppose $x_{a^\star}^\top \theta^\star - x_{\widehat{a}}^\top \theta^\star \leqslant \frac{\rho}{8}$. Compare $\widehat{a}$ with every other arm using the pairwise test with tolerance $\varepsilon = \frac{\rho}{8}$, and remove an arm a only when the test selects $\widehat{a}$ over a. If all pairwise tests satisfy Lemma C.1, the surviving set $\mathcal{A}_1$ contains $a^\star$ (i.e., $a^\star \in \mathcal{A}_1$) and satisfies*

$$W(\mathcal{A}_1) = \max_{a,b \in \mathcal{A}_1} |(x_a - x_b)^\top \theta^\star| \leqslant \frac{\rho}{4}.$$

Proof of Lemma C.3. The optimal arm cannot be removed, since Lemma C.1 precludes selecting $\widehat{a}$ over $a^\star$ as $x_{\widehat{a}}^\top \theta^\star \leqslant x_{a^\star}^\top \theta^\star$.

Now consider any surviving arm a. If the test selects a over $\widehat{a}$, then

$$x_a^\top \theta^\star > x_{\widehat{a}}^\top \theta^\star.$$

Otherwise, neither arm is selected, and Lemma C.1, applied with $\varepsilon = \rho/8$, gives

$$\left|(x_a - x_{\widehat{a}})^\top \theta^\star\right| \leqslant \frac{\rho}{8},$$

and hence $x_a^\top \theta^\star \geqslant x_{\widehat{a}}^\top \theta^\star - \frac{\rho}{8}$. Thus, every surviving arm a satisfies

$$x_a^\top \theta^\star \geqslant x_{\widehat{a}}^\top \theta^\star - \frac{\rho}{8} \geqslant x_{a^\star}^\top \theta^\star - \frac{\rho}{4},$$

where the second inequality follows from the assumed $\rho/8$-optimality of $\widehat{a}$.

Hence, since $a^\star \in \mathcal{A}_1$ and $a^\star$ has the largest linear score, we have

$$\begin{aligned} W(\mathcal{A}_1) &= \max_{a,b \in \mathcal{A}_1} \left|(x_a - x_b)^\top \theta^\star\right| \\ &= \max_{a \in \mathcal{A}_1} \left(x_{a^\star}^\top \theta^\star - x_a^\top \theta^\star\right) \leqslant \frac{\rho}{4}. \end{aligned}$$

This proves the lemma. □

Proposition C.1 (Sample complexity of initial localization)**.** *For $K = 1$, return $[K]$ with $N_{\text{init}} = 0$. Otherwise, run the tournament with tolerance $\rho/(8\lceil \log_2 K \rceil)$ and then filtering with tolerance $\rho/8$. Assign failure probability $\delta/(2K)$ to each pairwise test. Under Assumptions 1–4, with probability*

at least $1-\delta$*, whenever initialization is completed, its output satisfies Equation 2. Moreover, the number of initialization rounds is at most*

$$N_{\mathrm{init}}(\rho,\delta) := (K-1)N_{\mathrm{pair}}\left(\frac{\rho}{8\lceil\log_2 K\rceil},\frac{\delta}{2K}\right) + KN_{\mathrm{pair}}\left(\frac{\rho}{8},\frac{\delta}{2K}\right). \tag{C.3}$$

Consequently,

$$N_{\mathrm{init}}(\rho,\delta) = \mathcal{O}\left(K\left[1+\frac{(L_\mu+M_\mu^2)\kappa_{\mathrm{init}}^2\lceil\log_2 K\rceil^2}{\rho^2} + \frac{(R+M_\mu+\sqrt{L_\mu})\kappa_{\mathrm{init}}\lceil\log_2 K\rceil}{\rho}\right]\log\frac{2K}{\delta}\right).$$

Proof of Proposition C.1. The tournament uses exactly $K-1$ pairwise tests, since each test removes one arm from further tournament consideration until a single arm remains. The filtering step then compares the tournament winner $\widehat{a}$ with each arm in the original action set, requiring at most K additional pairwise tests. Thus, there are fewer than $2K$ tests in total, and a union bound implies that all of them satisfy Lemma C.1 with probability at least $1-\delta$.

On this event, Lemma C.2, with $\varepsilon = \rho/(8\lceil\log_2 K\rceil)$, gives

$$x_{a^\star}^\top\theta^\star - x_{\widehat{a}}^\top\theta^\star \leqslant \frac{\rho}{8}.$$

Lemma C.3 then yields Equation 2.

Finally, each tournament comparison uses $N_{\mathrm{pair}}\left(\frac{\rho}{8\lceil\log_2 K\rceil},\frac{\delta}{2K}\right)$ samples, while each filtering comparison uses $N_{\mathrm{pair}}\left(\frac{\rho}{8},\frac{\delta}{2K}\right)$ samples. Multiplying these per-test sample requirements by the numbers of tests, $K-1$ and at most K, respectively, gives Equation C.3. The stated order bound then follows from Equation C.1. □

D Analysis of a Localized Elimination Phase

In this section, we analyze a single localized elimination phase of Step 2 (Subsection 4.2). We show that localization controls the nonlinear Taylor error, which allows the centered estimator to accurately compare active arms and leads to safe elimination and a phase-wise regret bound.

Fix phase m and condition on the history at its start. Index the samples collected in this phase by $s = 1, 2, \ldots$, and use $\nu_s, A_s, Y_s, \mathcal{G}_s$ for the corresponding round-wise quantities. Throughout this section, we assume

$$a^\star \in \mathcal{A}_m, \qquad W(\mathcal{A}_m) = \max_{a,b\in\mathcal{A}_m} |(x_a - x_b)^\top\theta^\star| \leqslant \rho_m \leqslant \rho_1. \tag{D.1}$$

The distribution p_m, the centered features $z_{m,a}$, and the covariance Σ_m remain fixed throughout the phase. The technical inequalities used below are collected in Appendix F.

Because $\bar{x}_m^\top\theta^\star$ is a convex combination of the active-arm scores, Equation D.1 implies

$$|z_{m,a}^\top\theta^\star| = |x_a^\top\theta^\star - \bar{x}_m^\top\theta^\star| \leqslant W(\mathcal{A}_m) \leqslant \rho_m, \qquad a \in \mathcal{A}_m. \tag{D.2}$$

For the s-th sample, define

$$c_s := \nu_s + \bar{x}_m^\top\theta^\star, \qquad \alpha_s := \mu'(c_s),$$

and

$$\mathcal{R}_{s,a} := \mu(c_s + z_{m,a}^\top\theta^\star) - \mu(c_s) - \alpha_s z_{m,a}^\top\theta^\star.$$

Finally, we define

$$G_m(n) := \left(\sum_{s=1}^n \alpha_s\right) W(\mathcal{A}_m),$$

the true scaled pairwise score width on the common scale after n samples.

D.1 Localized Taylor error

Lemma D.1 (Centered local moment)**.** *For every sample s in phase m,*

$$\mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s] = \alpha_s \Sigma_m \theta^\star + r_s,$$

where $r_s := \mathbb{E}[z_{m,A_s} \mathcal{R}_{s,A_s} \mid \mathcal{G}_s]$. Moreover, for every $a \in \mathcal{A}_m$, we have

$$|\mathcal{R}_{s,a}| \leqslant \alpha_s W(\mathcal{A}_m)\big(e^{R\rho_m} - 1\big).$$

Proof of Lemma D.1. Since $A_s \sim p_m$ within the phase,

$$\begin{aligned}\mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s] &= \sum_{a \in \mathcal{A}_m} p_m(a) z_{m,a} \mu(c_s + z_{m,a}^\top \theta^\star) \\ &= \sum_{a \in \mathcal{A}_m} p_m(a) z_{m,a} \left(\mu(c_s) + \alpha_s z_{m,a}^\top \theta^\star + \mathcal{R}_{s,a}\right).\end{aligned}$$

By the definition of the centered features, we have $\sum_{a \in \mathcal{A}_m} p_m(a) z_{m,a} = 0$ and $\sum_{a \in \mathcal{A}_m} p_m(a) z_{m,a} z_{m,a}^\top = \Sigma_m$. Therefore, we can write

$$\mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s] = \alpha_s \Sigma_m \theta^\star + \sum_{a \in \mathcal{A}_m} p_m(a) z_{m,a} \mathcal{R}_{s,a},$$

which gives the first claim.

For the remainder, Lemma F.2 gives

$$|\mathcal{R}_{s,a}| \leqslant \alpha_s |z_{m,a}^\top \theta^\star| \left(e^{R|z_{m,a}^\top \theta^\star|} - 1\right).$$

Using Equation D.2, we have

$$|z_{m,a}^\top \theta^\star| \leqslant W(\mathcal{A}_m) \leqslant \rho_m,$$

and hence we obtain

$$|\mathcal{R}_{s,a}| \leqslant \alpha_s W(\mathcal{A}_m)\big(e^{R\rho_m} - 1\big).$$

This proves the lemma. □

Lemma D.2 (Localized Taylor error)**.** *For every $n \geqslant 1$ and every $a, b \in \mathcal{A}_m$, we have*

$$\left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^n r_s\right| \leqslant \eta_m \underbrace{\left(\sum_{s=1}^n \alpha_s\right) W(\mathcal{A}_m)}_{=: G_m(n)} \tag{D.3}$$

where $\eta_m := 2\sqrt{d}\big(e^{R\rho_m} - 1\big)$ and satisfies $\eta_m \leqslant 1/64$.

Proof of Lemma D.2. Since $x_a - x_b = z_{m,a} - z_{m,b}$, Lemma B.1 implies

$$\|x_a - x_b\|_{\Sigma_m^\dagger} \leqslant \|z_{m,a}\|_{\Sigma_m^\dagger} + \|z_{m,b}\|_{\Sigma_m^\dagger} \leqslant 2\sqrt{d}. \tag{D.4}$$

For a single sample s, we have

$$\begin{aligned}|(x_a - x_b)^\top \Sigma_m^\dagger r_s| &= \left|\sum_{c \in \mathcal{A}_m} p_m(c)(x_a - x_b)^\top \Sigma_m^\dagger z_{m,c} \mathcal{R}_{s,c}\right| && \text{(definition of } r_s\text{)} \\ &\leqslant \alpha_s W(\mathcal{A}_m)\big(e^{R\rho_m} - 1\big) \sum_{c \in \mathcal{A}_m} p_m(c) |(x_a - x_b)^\top \Sigma_m^\dagger z_{m,c}| && \text{(Lemma D.1)} \\ &\leqslant \alpha_s W(\mathcal{A}_m)\big(e^{R\rho_m} - 1\big) \left(\sum_{c \in \mathcal{A}_m} p_m(c)\big((x_a - x_b)^\top \Sigma_m^\dagger z_{m,c}\big)^2\right)^{1/2} \\ & && \text{(Cauchy–Schwarz)} \\ &= \alpha_s W(\mathcal{A}_m)\big(e^{R\rho_m} - 1\big)\|x_a - x_b\|_{\Sigma_m^\dagger} && \text{(definition of } \Sigma_m\text{)} \\ &\leqslant 2\sqrt{d}\,\alpha_s W(\mathcal{A}_m)\big(e^{R\rho_m} - 1\big). && \text{(Equation D.4)}\end{aligned}$$

Summing over $s = 1, \dots, n$ and applying the triangle inequality gives

$$\begin{aligned}\left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} r_s\right| &= \left|\sum_{s=1}^{n} (x_a - x_b)^\top \Sigma_m^\dagger r_s\right| \\ &\leqslant \sum_{s=1}^{n} \left|(x_a - x_b)^\top \Sigma_m^\dagger r_s\right| \\ &\leqslant 2\sqrt{d}\left(e^{R\rho_m} - 1\right)\left(\sum_{s=1}^{n} \alpha_s\right) W(\mathcal{A}_m) \\ &= \eta_m G_m(n),\end{aligned}$$

which proves Equation D.3. Finally, Equation D.1 gives $\rho_m \leqslant \rho_1$, so Lemma F.3 yields $\eta_m \leqslant 1/64$. □

D.2 Confidence bounds

The concentration bounds below do not require the localization invariant. All sample counts refer to observations actually collected in the phase; after its termination, the martingale increments are set to zero. Define the vector martingale difference

$$D_s := z_{m,A_s} Y_s - \mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s].$$

For $n \geqslant 1$, define

$$\begin{aligned}\beta_{K,m}(n) := &\sqrt{64(L_\mu + M_\mu^2) dn \log\left(\frac{12T^2|\mathcal{A}_m|^2}{\delta}\right)} \\ &+ 8(R + M_\mu + \sqrt{L_\mu}) d \log\left(\frac{12T^2|\mathcal{A}_m|^2}{\delta}\right), \qquad \text{(D.5)} \\ \beta_{d,m}(n) := &\, 4\sqrt{16(L_\mu + M_\mu^2) dn \left(d \log 5 + \log\frac{12T^2}{\delta}\right)} \\ &+ 16(R + M_\mu + \sqrt{L_\mu}) d \left(d \log 5 + \log\frac{12T^2}{\delta}\right), \qquad \text{(D.6)} \\ \beta_m(n) := &\, \min\{\beta_{K,m}(n), \beta_{d,m}(n)\}. \qquad \text{(D.7)}\end{aligned}$$

These confidence radii are fully observable and do not depend on the unobserved common shifts ν_t or on the initialization curvature κ_{init}.

Lemma D.3 (Pairwise concentration)**.** *Define $D_s := z_{m,A_s} Y_s - \mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s]$. Conditional on the start of phase m, with failure probability at most $\delta/(6T)$, simultaneously for every $n \leqslant T$ and every $a, b \in \mathcal{A}_m$, we have*

$$\left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} D_s\right| \leqslant \beta_{K,m}(n).$$

Proof of Lemma D.3. Fix $a, b \in \mathcal{A}_m$ and define $X_s := (x_a - x_b)^\top \Sigma_m^\dagger D_s$. By the definition of D_s,

$$\mathbb{E}[X_s \mid \mathcal{G}_s] = (x_a - x_b)^\top \Sigma_m^\dagger \mathbb{E}[D_s \mid \mathcal{G}_s] = 0,$$

so (X_s) is a scalar martingale-difference sequence (MDS). We apply Lemma F.5 with $h_{s,c} = (x_a - x_b)^\top \Sigma_m^\dagger z_{m,c}$. The design remains fixed within the phase, and

$$\begin{aligned}\sum_{c \in \mathcal{A}_m} p_m(c) h_{s,c}^2 &= (x_a - x_b)^\top \Sigma_m^\dagger \Sigma_m \Sigma_m^\dagger (x_a - x_b) \\ &= \|x_a - x_b\|_{\Sigma_m^\dagger}^2 \leqslant 4d,\end{aligned}$$

where the last inequality is Equation D.4. Moreover, Cauchy–Schwarz and Lemma B.1 give

$$|h_{s,c}| \leqslant \|x_a - x_b\|_{\Sigma_m^\dagger} \|z_{m,c}\|_{\Sigma_m^\dagger} \leqslant 2d.$$

Thus, the weighted-reward lemma applies with $V = 4d$ and $H = 2d$. For $x = \log(12T^2|\mathcal{A}_m|^2/\delta)$ and any fixed $n \leqslant T$, with failure probability at most $2e^{-x}$,

$$\left|\sum_{s=1}^{n} X_s\right| \leqslant \sqrt{64(L_\mu + M_\mu^2)dnx} + 8(R + M_\mu + \sqrt{L_\mu})dx$$
$$= \beta_{K,m}(n).$$

The failure probability for each fixed (a, b, n) is

$$2\exp\left(-\log\frac{12T^2|\mathcal{A}_m|^2}{\delta}\right) = \frac{\delta}{6T^2|\mathcal{A}_m|^2}.$$

Finally, a union bound over at most $|\mathcal{A}_m|^2 T$ choices of (a, b, n) gives a total failure probability of at most $\delta/(6T)$. □

Lemma D.4 (Concentration over the active difference span)**.** *Conditional on the start of phase m, with failure probability at most $\delta/(6T)$, simultaneously for every $n \leqslant T$ and every $a, b \in \mathcal{A}_m$,*

$$\left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} D_s\right| \leqslant \beta_{d,m}(n).$$

Proof of Lemma D.4. By Lemma B.1, we have $\mathrm{range}(\Sigma_m) = \mathrm{span}\{x_a - x_b : a, b \in \mathcal{A}_m\}$. Moreover, each centered feature $z_{m,a}$ lies in this subspace, and hence so does D_s. We therefore restrict the analysis to $\mathrm{range}(\Sigma_m)$ and define

$$U_s := \Sigma_m^{\dagger/2} D_s.$$

For any unit vector $q \in \mathrm{range}(\Sigma_m)$, $q^\top U_s$ is a scalar martingale difference since

$$\mathbb{E}[q^\top U_s \mid \mathcal{G}_s] = q^\top \Sigma_m^{\dagger/2} \mathbb{E}[D_s \mid \mathcal{G}_s] = 0.$$

For the weights $h_{s,c} = q^\top \Sigma_m^{\dagger/2} z_{m,c}$, we have

$$\sum_{c \in \mathcal{A}_m} p_m(c) h_{s,c}^2 = q^\top \Sigma_m^{\dagger/2} \Sigma_m \Sigma_m^{\dagger/2} q = 1,$$
$$|h_{s,c}| \leqslant \|q\|_2 \|z_{m,c}\|_{\Sigma_m^\dagger} \leqslant \sqrt{d}.$$

Hence Lemma F.5 applies with $V = 1$ and $H = \sqrt{d}$.

Let $\mathcal{N}$ be a $1/2$-net of the unit sphere in $\mathrm{range}(\Sigma_m)$ with $|\mathcal{N}| \leqslant 5^d$, and set

$$\lambda_d := d\log 5 + \log\frac{12T^2}{\delta}.$$

For any fixed $q \in \mathcal{N}$ and $n \leqslant T$, Lemma F.5 gives, with probability at least $1 - 2e^{-\lambda_d}$,

$$\left|q^\top \sum_{s=1}^{n} U_s\right| \leqslant \sqrt{16(L_\mu + M_\mu^2)n\lambda_d} + 4(R + M_\mu + \sqrt{L_\mu})\sqrt{d}\,\lambda_d. \qquad \text{(Lemma F.5)}$$

Since $2e^{-\lambda_d} = \frac{\delta}{6T^2 5^d}$, the failure probability for each fixed (q, n) is at most $\delta/(6T^2 5^d)$. Since $|\mathcal{N}| \leqslant 5^d$, a union bound over all $(q, n) \in \mathcal{N} \times [T]$ yields, with probability at least $1 - \delta/(6T)$, simultaneously for all $n \leqslant T$,

$$\max_{q \in \mathcal{N}} \left|q^\top \sum_{s=1}^{n} U_s\right| \leqslant \sqrt{16(L_\mu + M_\mu^2)n\lambda_d} + 4(R + M_\mu + \sqrt{L_\mu})\sqrt{d}\,\lambda_d.$$

Since $\mathcal{N}$ is a $1/2$-net of the unit sphere in $\mathrm{range}(\Sigma_m)$, the standard net argument gives

$$\left\|\sum_{s=1}^{n} U_s\right\|_2 \leqslant 2\max_{q \in \mathcal{N}} \left|q^\top \sum_{s=1}^{n} U_s\right| \qquad \text{(1/2-net inequality)}$$
$$\leqslant 2\sqrt{16(L_\mu + M_\mu^2)n\lambda_d} + 8(R + M_\mu + \sqrt{L_\mu})\sqrt{d}\,\lambda_d.$$

Finally, for any $a, b \in \mathcal{A}_m$, we derive that

$$\begin{aligned}
\left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} D_s\right| &= \left|\left(\Sigma_m^{\dagger/2}(x_a - x_b)\right)^\top \sum_{s=1}^{n} U_s\right| && \text{(definition of } U_s\text{)} \\
&\leqslant \|x_a - x_b\|_{\Sigma_m^\dagger} \left\|\sum_{s=1}^{n} U_s\right\|_2 && \text{(Cauchy–Schwarz)} \\
&\leqslant 2\sqrt{d}\left(2\sqrt{16(L_\mu + M_\mu^2)n\lambda_d} + 8(R + M_\mu + \sqrt{L_\mu})\sqrt{d}\,\lambda_d\right) \\
& && \text{(Equation D.4)} \\
&= \beta_{d,m}(n). && \text{(Equation D.6)}
\end{aligned}$$

This proves the lemma. □

Corollary D.1 (Uniform confidence for all phases)**.** *With probability at least $1 - \delta/3$, the following holds simultaneously for every phase m started by the algorithm that satisfies Equation D.1, every sample count $n \leqslant T$ in that phase, and every $a, b \in \mathcal{A}_m$:*

$$\left|(x_a - x_b)^\top \widehat{w}_{m,n} - \left(\sum_{s=1}^{n} \alpha_s\right)(x_a - x_b)^\top \theta^\star\right| \leqslant \beta_m(n) + \eta_m G_m(n), \tag{D.8}$$

where $\beta_m(n) := \min\{\beta_{K,m}(n), \beta_{d,m}(n)\}$.

Proof of Corollary D.1. For each phase, Lemmas D.3 and D.4 each fail with probability at most $\delta/(6T)$. Since at most T phases can be started, a union bound shows that both concentration bounds hold simultaneously over all phases with probability at least $1 - \delta/3$, whether or not they satisfy the localization invariant. On this event, consider any phase satisfying Equation D.1.

Using the definitions of D_s and $\widehat{w}_{m,n}$, we obtain

$$\begin{aligned}
\widehat{w}_{m,n} &= \Sigma_m^\dagger \sum_{s=1}^{n} z_{m,A_s} Y_s && \text{(definition of } \widehat{w}_{m,n}\text{, Equation 3)} \\
&= \Sigma_m^\dagger \sum_{s=1}^{n} D_s + \Sigma_m^\dagger \sum_{s=1}^{n} \mathbb{E}[z_{m,A_s} Y_s \mid \mathcal{G}_s] && \text{(definition of } D_s\text{)} \\
&= \Sigma_m^\dagger \sum_{s=1}^{n} D_s + \left(\sum_{s=1}^{n} \alpha_s\right) \Sigma_m^\dagger \Sigma_m \theta^\star + \Sigma_m^\dagger \sum_{s=1}^{n} r_s. && \text{(Lemma D.1)}
\end{aligned}$$

Since $\Sigma_m^\dagger \Sigma_m$ is the orthogonal projector onto $\mathrm{range}(\Sigma_m)$, Lemma B.1 implies that, for every $a, b \in \mathcal{A}_m$,

$$(x_a - x_b)^\top \Sigma_m^\dagger \Sigma_m \theta^\star = (x_a - x_b)^\top \theta^\star.$$

Therefore, we obtain

$$(x_a - x_b)^\top \widehat{w}_{m,n} - \left(\sum_{s=1}^{n} \alpha_s\right)(x_a - x_b)^\top \theta^\star = (x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} D_s + (x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} r_s.$$

Taking absolute values and applying the triangle inequality gives

$$\begin{aligned}
&\left|(x_a - x_b)^\top \widehat{w}_{m,n} - \left(\sum_{s=1}^{n} \alpha_s\right)(x_a - x_b)^\top \theta^\star\right| \\
&\quad\leqslant \left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} D_s\right| + \left|(x_a - x_b)^\top \Sigma_m^\dagger \sum_{s=1}^{n} r_s\right| && \text{(triangle inequality)} \\
&\quad\leqslant \beta_m(n) + \eta_m G_m(n). && \text{(Lemmas D.3, D.4, and D.2)}
\end{aligned}$$

This completes the proof. □

D.3 Observable confidence radius and elimination

To safely eliminate arms, we need to control the error in the estimated pairwise score differences

$$(x_a - x_b)^\top \widehat{w}_{m,n}.$$

Corollary D.1 shows that this error is bounded by

$$\beta_m(n) + \eta_m G_m(n),$$

where $G_m(n) = (\sum_{s=1}^{n} \alpha_s)\, W(\mathcal{A}_m)$ is the true pairwise score width. However, $G_m(n)$ is not observable because the slopes α_s depend on the unobserved common shifts. Thus, this bound cannot be used directly by the algorithm.

We therefore upper bound the same estimation error by the *observable* quantity $\widehat{q}_m(n)$, which depends only on quantities available to the learner. The following lemma makes this connection precise.

Lemma D.5 (Observable confidence radius)**.** *Let $\widehat{G}_m(n) = \max_{a,b\in\mathcal{A}_m}(x_a - x_b)^\top \widehat{w}_{m,n}$, and $G_m(n) = (\sum_{s=1}^{n} \alpha_s)\, W(\mathcal{A}_m)$. On the event in Corollary D.1, for every phase m and every $n \leqslant T$, we have*

$$|\widehat{G}_m(n) - G_m(n)| \leqslant \beta_m(n) + \eta_m G_m(n) \leqslant \widehat{q}_m(n).$$

The same radius $\widehat{q}_m(n)$ bounds the pairwise estimation error in Equation D.8.

Proof of Lemma D.5. Taking the maximum over ordered pairs in Equation D.8 and using $|\max_i u_i - \max_i v_i| \leqslant \max_i |u_i - v_i|$ gives

$$|\widehat{G}_m(n) - G_m(n)| \leqslant \beta_m(n) + \eta_m G_m(n).$$

Hence, we get

$$\begin{aligned} G_m(n) &\leqslant \widehat{G}_m(n) + \beta_m(n) + \eta_m G_m(n) \\ &\leqslant \frac{\widehat{G}_m(n) + \beta_m(n)}{1 - \eta_m}. && (\eta_m < 1) \end{aligned}$$

Therefore, we have

$$\begin{aligned} \beta_m(n) + \eta_m G_m(n) &\leqslant \beta_m(n) + \eta_m \frac{\widehat{G}_m(n) + \beta_m(n)}{1 - \eta_m} \\ &= \frac{\beta_m(n) + \eta_m \widehat{G}_m(n)}{1 - \eta_m} \\ &= \widehat{q}_m(n). && \text{(Equation 4)} \end{aligned}$$

This proves the lemma. □

Lemma D.6 (Accuracy at the stopping time)**.** *Suppose phase m stops after n samples according to Equation 5. On the event in Corollary D.1, we have*

$$\widehat{q}_m(n) \leqslant \frac{G_m(n)}{16}.$$

Proof of Lemma D.6. At the stopping time, we have

$$\beta_m(n) \leqslant \frac{\widehat{G}_m(n)}{64} \qquad \text{(Equation 5)}$$

and, by Equation D.1 and Lemma F.3, we also have

$$\eta_m \leqslant \frac{1}{64}.$$

Hence, we get

$$\begin{aligned} \widehat{q}_m(n) &= \frac{\beta_m(n) + \eta_m \widehat{G}_m(n)}{1 - \eta_m} && \text{(Equation 4)} \\ &\leqslant \frac{\widehat{G}_m(n)/64 + \widehat{G}_m(n)/64}{1 - 1/64} = \frac{2}{63}\widehat{G}_m(n). \end{aligned}$$

Moreover, we have

$$\begin{aligned} G_m(n) &\geqslant \widehat{G}_m(n) - \widehat{q}_m(n) && \text{(Lemma D.5)} \\ &\geqslant \frac{61}{63}\widehat{G}_m(n). \end{aligned}$$

Combining the two bounds gives

$$\widehat{q}_m(n) \leqslant \frac{2}{63}\widehat{G}_m(n) \leqslant \frac{2}{61}G_m(n) < \frac{G_m(n)}{16}.$$

This proves the lemma. □

Lemma D.7 (Safe elimination and width contraction)**.** *Suppose phase m stops and $\mathcal{A}_{m+1}$ is formed according to Equation 6. On the event in Corollary D.1, we have*

$$a^\star \in \mathcal{A}_{m+1}, \qquad W(\mathcal{A}_{m+1}) \leqslant \frac{W(\mathcal{A}_m)}{4} \leqslant \frac{\rho_m}{4}.$$

Proof of Lemma D.7. Let n be the stopping time and write $\widehat{q} = \widehat{q}_m(n)$. Since $\widehat{a}_m$ maximizes the estimated score, we have

$$(x_{\widehat{a}_m} - x_{a^\star})^\top \widehat{w}_{m,n} \geqslant 0.$$

Using Lemma D.5, we get

$$\begin{aligned} (x_{\widehat{a}_m} - x_{a^\star})^\top \widehat{w}_{m,n} &\leqslant \left(\sum_{s=1}^{n} \alpha_s\right)(x_{\widehat{a}_m} - x_{a^\star})^\top \theta^\star + \widehat{q} && \text{(Lemma D.5)} \\ &\leqslant \widehat{q}, && \text{(optimality of } a^\star \text{ and } \alpha_s > 0) \end{aligned}$$

since $(x_{\widehat{a}_m} - x_{a^\star})^\top \theta^\star \leqslant 0$. Thus, we have $(x_{\widehat{a}_m} - x_{a^\star})^\top \widehat{w}_{m,n} \leqslant 2\widehat{q}$, so $a^\star$ satisfies the retention condition in Equation 6 and therefore $a^\star \in \mathcal{A}_{m+1}$.

Now consider any $a \in \mathcal{A}_{m+1}$. We have

$$\begin{aligned} \left(\sum_{s=1}^{n} \alpha_s\right)(x_{a^\star} - x_a)^\top \theta^\star &\leqslant (x_{a^\star} - x_a)^\top \widehat{w}_{m,n} + \widehat{q} && \text{(Lemma D.5)} \\ &\leqslant (x_{\widehat{a}_m} - x_a)^\top \widehat{w}_{m,n} + \widehat{q} && \text{(definition of } \widehat{a}_m) \\ &\leqslant 3\widehat{q} && \text{(Equation 6)} \\ &\leqslant \frac{3}{16}G_m(n). && \text{(Lemma D.6)} \end{aligned}$$

Since $G_m(n) = (\sum_{s=1}^{n} \alpha_s)\, W(\mathcal{A}_m)$ and $\sum_{s=1}^{n} \alpha_s > 0$, it follows that

$$(x_{a^\star} - x_a)^\top \theta^\star \leqslant \frac{3}{16}W(\mathcal{A}_m) < \frac{1}{4}W(\mathcal{A}_m).$$

Therefore, since $a^\star \in \mathcal{A}_{m+1}$,

$$\begin{aligned} W(\mathcal{A}_{m+1}) &= \max_{a \in \mathcal{A}_{m+1}} (x_{a^\star} - x_a)^\top \theta^\star \\ &\leqslant \frac{1}{4}W(\mathcal{A}_m) \\ &\leqslant \frac{\rho_m}{4}. && \text{(Equation D.1)} \end{aligned}$$

This concludes the proof. □

D.4 Regret in a localized phase

Lemma D.8 (Scaled pairwise score width before stopping)**.** *On the event in Corollary D.1, if the stopping rule has not been satisfied after n samples, then*

$$G_m(n) \leqslant 67\beta_m(n).$$

If phase m ends after n_m samples, either because the stopping rule is satisfied or because no rounds remain, then

$$G_m(n_m) \leqslant 67\beta_m(n_m) + L_\mu \rho_m. \tag{D.9}$$

Proof of Lemma D.8. Before the stopping rule is satisfied, we have

$$\widehat{G}_m(n) < 64\beta_m(n). \qquad \text{(Equation 5)}$$

Moreover, Lemma D.5 gives

$$\widehat{G}_m(n) \geqslant (1-\eta_m)G_m(n) - \beta_m(n). \qquad \text{(Lemma D.5)}$$

Combining the two inequalities yields $(1-\eta_m)G_m(n) < 65\beta_m(n)$, which implies

$$\begin{aligned} G_m(n) &< \frac{65}{1-\eta_m}\beta_m(n) \\ &\leqslant \frac{65\cdot 64}{63}\beta_m(n) && (\eta_m \leqslant 1/64) \\ &< 67\beta_m(n). \end{aligned}$$

This proves the first claim.

If the horizon is exhausted before the stopping condition is met, the first claim directly implies Equation D.9. Otherwise, suppose the phase stops after $n_m > 1$ samples. Applying the first claim at $n_m - 1$ gives

$$G_m(n_m - 1) \leqslant 67\beta_m(n_m - 1).$$

The contribution of the final sample is bounded by

$$\begin{aligned} G_m(n_m) - G_m(n_m-1) &= \alpha_{n_m} W(\mathcal{A}_m) && \text{(definition of } G_m) \\ &\leqslant L_\mu \rho_m, && (\alpha_{n_m} \leqslant L_\mu \text{ and Equation D.1}) \end{aligned}$$

and hence we get

$$\begin{aligned} G_m(n_m) &\leqslant 67\beta_m(n_m-1) + L_\mu\rho_m \\ &\leqslant 67\beta_m(n_m) + L_\mu\rho_m. && \text{(monotonicity of } \beta_m(n)) \end{aligned}$$

If $n_m = 1$, the same conclusion follows directly from $G_m(1) \leqslant L_\mu\rho_m$. □

Lemma D.9 (Conditional expected regret in one phase)**.** *For the first n samples of phase m, we have*

$$\sum_{s=1}^{n} \mathbb{E}\big[\mu(\nu_s + x_{a^\star}^\top\theta^\star) - \mu(\nu_s + x_{A_s}^\top\theta^\star) \,\big|\, \mathcal{G}_s\big] \leqslant e^{R\rho_m} G_m(n). \qquad \text{(D.10)}$$

Consequently, if the phase ends after n_m samples, then on the event in Corollary D.1, its conditional expected regret is at most

$$e^{R\rho_m}\left(67\beta_m(n_m) + L_\mu\rho_m\right).$$

Proof of Lemma D.9. For any $a \in \mathcal{A}_m$, the mean-value theorem gives some $\xi_{s,a}$ between $x_a^\top\theta^\star$ and $x_{a^\star}^\top\theta^\star$ such that

$$\begin{aligned} &\mu(\nu_s + x_{a^\star}^\top\theta^\star) - \mu(\nu_s + x_a^\top\theta^\star) \\ &\quad = \mu'(\nu_s + \xi_{s,a})(x_{a^\star} - x_a)^\top\theta^\star. && \text{(mean-value theorem)} \end{aligned}$$

Since both $\xi_{s,a}$ and $\bar{x}_m^\top\theta^\star$ lie between the minimum and maximum active-arm scores, we have

$$\left|\xi_{s,a} - \bar{x}_m^\top\theta^\star\right| \leqslant W(\mathcal{A}_m) \leqslant \rho_m. \qquad \text{(Equation D.1)}$$

Therefore,

$$\begin{aligned} \mu'(\nu_s + \xi_{s,a}) &\leqslant e^{R\rho_m}\mu'(\nu_s + \bar{x}_m^\top\theta^\star) && \text{(Lemma F.1)} \\ &= e^{R\rho_m}\alpha_s. && \text{(definition of } \alpha_s) \end{aligned}$$

Combining the above gives

$$\mu(\nu_s + x_{a^\star}^\top\theta^\star) - \mu(\nu_s + x_a^\top\theta^\star) \leqslant e^{R\rho_m}\alpha_s(x_{a^\star} - x_a)^\top\theta^\star.$$

Averaging over $A_s \sim p_m$, we obtain

$$\begin{aligned}\mathbb{E}\big[\mu(\nu_s + x_{a^\star}^\top\theta^\star) - \mu(\nu_s + x_{A_s}^\top\theta^\star)\,\big|\,\mathcal{G}_s\big] &\leqslant e^{R\rho_m}\alpha_s \sum_{a\in\mathcal{A}_m} p_m(a)(x_{a^\star} - x_a)^\top\theta^\star \qquad (A_s \sim p_m)\\ &= e^{R\rho_m}\alpha_s\left(x_{a^\star}^\top\theta^\star - \bar{x}_m^\top\theta^\star\right) \qquad \text{(definition of } \bar{x}_m)\\ &\leqslant e^{R\rho_m}\alpha_s W(\mathcal{A}_m).\end{aligned}$$

Summing over $s = 1, \dots, n$ gives

$$\begin{aligned}\sum_{s=1}^{n}\mathbb{E}\big[\mu(\nu_s + x_{a^\star}^\top\theta^\star) - \mu(\nu_s + x_{A_s}^\top\theta^\star)\,\big|\,\mathcal{G}_s\big] &\leqslant e^{R\rho_m}\left(\sum_{s=1}^{n}\alpha_s\right)W(\mathcal{A}_m)\\ &= e^{R\rho_m}G_m(n). \qquad \text{(definition of } G_m(n))\end{aligned}$$

This proves Equation D.10. The final claim follows by applying Lemma D.8 at $n = n_m$. □

E Proof of the Regret Upper Bound

In this section, we present the full proof of Theorem 1. We first aggregate the phase-wise confidence bounds (Lemma E.1) and use them to bound the conditional expected regret after initialization (Lemma E.2). We then convert this conditional expected regret bound into a bound on the realized pseudo-regret (Lemma E.3). Finally, we combine these bounds with the initial localization guarantee (Proposition C.1) to complete the proof.

Lemma E.1 (Sum of phase confidence radii)**.** *Let n_m be the number of samples collected in phase m. Then, we have*

$$\sum_m \beta_m(n_m) = \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\}\Big).$$

Proof of Lemma E.1. Let M denote the number of phases in which the algorithm collects samples. Since $\rho_{m+1} = \rho_m/4$, we have $\rho_m = \rho_1/4^{m-1}$. The algorithm continues sampling in phase m as long as $\rho_m > (L_\mu T)^{-1}$. Once $\rho_m \leqslant (L_\mu T)^{-1}$, it stops the phased sampling procedure and plays any surviving arm for the rest of the horizon.

Thus, if phase m is sampled, then $\rho_1/4^{m-1} > (L_\mu T)^{-1}$, or equivalently $4^{m-1} < \rho_1 L_\mu T$. Therefore,

$$M \leqslant 1 + \max\{0, \log_4(\rho_1 L_\mu T)\} = \mathcal{O}(\log T),$$

where we used $\rho_1 = \rho/4 = 1/(256R\sqrt{d})$ in Theorem 1.

Recall that $\beta_m(n) = \min\{\beta_{K,m}(n), \beta_{d,m}(n)\}$. If $\log K \leqslant d$, we use the pairwise bound $\beta_{K,m}$; otherwise, we use the span-wise bound $\beta_{d,m}$. By Equation D.5–Equation D.7, in either case, we have

$$\beta_m(n_m) = \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dn_m\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\}\Big),$$

where $\widetilde{\mathcal{O}}$ hides logarithmic factors in T, K, and $1/\delta$.

Summing the phase-wise bounds over $m = 1, \dots, M$ gives

$$\begin{aligned}\sum_{m=1}^{M}\beta_m(n_m) = \widetilde{\mathcal{O}}\Big(&\sqrt{L_\mu + M_\mu^2}\sqrt{d\min\{d, \log K\}}\sum_{m=1}^{M}\sqrt{n_m}\\ &+ (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\}\,M\Big).\end{aligned}$$

Since $\sum_{m=1}^{M} n_m \leqslant T$, Cauchy–Schwarz gives

$$\begin{aligned}\sum_{m=1}^{M}\sqrt{n_m} &\leqslant \sqrt{M\sum_{m=1}^{M} n_m} \qquad \text{(Cauchy–Schwarz)}\\ &\leqslant \sqrt{MT}\\ &= \widetilde{\mathcal{O}}(\sqrt{T}). \qquad (M = \mathcal{O}(\log T))\end{aligned}$$

Moreover, since $M = \mathcal{O}(\log T)$, the factor M multiplying the lower-order term is also absorbed into $\widetilde{\mathcal{O}}$. Therefore, we have

$$\sum_{m=1}^{M} \beta_m(n_m) = \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\}\Big),$$

which concludes the proof. □

Lemma E.2 (Conditional expected regret after initialization)**.** *Suppose the initialization stage satisfies the guarantee in Equation 2, i.e., $a^\star \in \mathcal{A}_1$ and $W(\mathcal{A}_1) \leqslant \rho_1 = \rho/4$. Let τ_{init} denote the first round after the initialization stage. Then, with probability at least $1 - \delta/3$, the conditional expected regret over all post-initialization rounds satisfies*

$$\sum_{t=\tau_{\text{init}}}^{T} \mathbb{E}\big[\mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_{A_t}^\top \theta^\star) \,\big|\, \mathcal{G}_t\big]$$
$$\leqslant \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\}\Big).$$

Proof of Lemma E.2. Conditional on the history at the start of the post-initialization procedure, the common stochastic event underlying Corollary D.1 holds with probability at least $1 - \delta/3$. We work on this event. By the initialization guarantee, $a^\star \in \mathcal{A}_1$ and $W(\mathcal{A}_1) \leqslant \rho_1$. Lemma D.7 then implies inductively that every phase reached by the algorithm satisfies

$$a^\star \in \mathcal{A}_m, \qquad W(\mathcal{A}_m) \leqslant \rho_m.$$

Therefore, Lemma D.9 applies whenever the algorithm collects samples in a phase.

Let M be the number of phases in which the algorithm collects samples, and let n_m be the number of samples collected in phase m. Thus, the sampled phases are $m = 1, \ldots, M$, and each satisfies $|\mathcal{A}_m| > 1$ and $\rho_m > (L_\mu T)^{-1}$. If the horizon is not exhausted after phase M, the algorithm stops either because $|\mathcal{A}_{M+1}| = 1$ or because $\rho_{M+1} \leqslant (L_\mu T)^{-1}$. In the former case, the remaining regret is zero; in the latter, the algorithm enters the terminal stage. Otherwise, the horizon is exhausted during phase M.

For each sampled phase $m = 1, \ldots, M$, we have

$$\sum_{s=1}^{n_m} \mathbb{E}\big[\mu(\nu_s + x_{a^\star}^\top \theta^\star) - \mu(\nu_s + x_{A_s}^\top \theta^\star) \,\big|\, \mathcal{G}_s\big] \leqslant e^{R\rho_m} G_m(n_m) \qquad \text{(Lemma D.9)}$$
$$\leqslant e^{R\rho_m}\left(67\beta_m(n_m) + L_\mu \rho_m\right). \qquad \text{(Lemma D.8)}$$

Since $\rho_{m+1} = \rho_m/4$, we have $\rho_m = \rho_1/4^{m-1}$. Hence,

$$\sum_{m=1}^{M} \rho_m = \rho_1 \sum_{m=0}^{M-1} 4^{-m} \leqslant \rho_1 \sum_{m=0}^{\infty} 4^{-m} = \frac{4}{3}\rho_1.$$

Moreover, for the choice of ρ used in Theorem 1 (i.e., $\rho = 1/(64R\sqrt{d})$ as specified in Equation F.1), we have $\rho_1 = \rho/4$ and hence $R\rho_1 = 1/(256\sqrt{d}) \leqslant 1/256$. Since $\rho_m \leqslant \rho_1$, it follows that, for every sampled phase m,

$$e^{R\rho_m} \leqslant e^{R\rho_1} \leqslant e^{1/256} = \mathcal{O}(1).$$

Summing the phase-wise regret bounds over $m = 1, \ldots, M$ and applying the bounds above gives

$$\sum_{m=1}^{M}\sum_{s=1}^{n_m} \mathbb{E}\big[\mu(\nu_s + x_{a^\star}^\top \theta^\star) - \mu(\nu_s + x_{A_s}^\top \theta^\star) \,\big|\, \mathcal{G}_s\big]$$
$$\leqslant 67e^{R\rho_1} \sum_{m=1}^{M} \beta_m(n_m) + L_\mu e^{R\rho_1} \sum_{m=1}^{M} \rho_m$$
$$= \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\}\Big) + \mathcal{O}(L_\mu\rho_1) \qquad \text{(Lemma E.1)}$$
$$= \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\} + 1\Big).$$

If the algorithm does not enter the terminal stage, the sampled phases cover all post-initialization rounds, and the above bound proves the claim.

It remains to consider the case in which $\rho_{M+1} \leqslant (L_\mu T)^{-1}$. For every surviving arm $a \in \mathcal{A}_{M+1}$, the mean-value theorem and $\mu' \leqslant L_\mu$ give

$$\begin{aligned}
\mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_a^\top \theta^\star) &\leqslant L_\mu (x_{a^\star} - x_a)^\top \theta^\star && (\mu' \leqslant L_\mu) \\
&\leqslant L_\mu W(\mathcal{A}_{M+1}) \\
&\leqslant L_\mu \rho_{M+1} \\
&\leqslant \frac{1}{T}. && (\rho_{M+1} \leqslant (L_\mu T)^{-1})
\end{aligned}$$

Since the terminal stage contains at most T rounds, its total conditional expected regret is at most 1. Combining this with the regret accumulated over the sampled phases proves the claim. □

Lemma E.3 (Conditional-to-realized pseudo-regret)**.** *Let τ_{init} denote the first round after initialization, and for $t = \tau_{\mathrm{init}}, \dots, T$, define*

$$\Delta_t := \mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_{A_t}^\top \theta^\star), \qquad \overline{\Delta}_t := \mathbb{E}[\Delta_t \mid \mathcal{G}_t].$$

Then, with probability at least $1 - \delta/3$, we have

$$\sum_{t=\tau_{\mathrm{init}}}^{T} \Delta_t \leqslant \sum_{t=\tau_{\mathrm{init}}}^{T} \overline{\Delta}_t + 2M_\mu \sqrt{2T \log \frac{6}{\delta}}.$$

Proof of Lemma E.3. Extend Δ_t and $\overline{\Delta}_t$ by zero before τ_{init}. Since completion of initialization is determined by the past, this gives a martingale-difference sequence over all T rounds. For every $t \geqslant \tau_{\mathrm{init}}$, we have

$$\begin{aligned}
\mathbb{E}[\Delta_t - \overline{\Delta}_t \mid \mathcal{G}_t] = \mathbb{E}[\Delta_t \mid \mathcal{G}_t] - \overline{\Delta}_t = 0, && \text{(definition of } \overline{\Delta}_t) \\
0 \leqslant \Delta_t \leqslant 2M_\mu, \\
0 \leqslant \overline{\Delta}_t \leqslant 2M_\mu, \\
|\Delta_t - \overline{\Delta}_t| \leqslant 2M_\mu.
\end{aligned}$$

Hence, by the two-sided Azuma–Hoeffding inequality,

$$\begin{aligned}
&\Pr\left(\left| \sum_{t=\tau_{\mathrm{init}}}^{T} (\Delta_t - \overline{\Delta}_t) \right| > 2M_\mu \sqrt{2T \log \frac{6}{\delta}} \right) \\
&\qquad \leqslant 2 \exp\left(- \frac{(2M_\mu)^2 \, 2T \log(6/\delta)}{2T(2M_\mu)^2} \right) && \text{(Azuma–Hoeffding)} \\
&\qquad = \frac{\delta}{3}.
\end{aligned}$$

Therefore, with probability at least $1 - \delta/3$, we obtain

$$\sum_{t=\tau_{\mathrm{init}}}^{T} \Delta_t = \sum_{t=\tau_{\mathrm{init}}}^{T} \overline{\Delta}_t + \sum_{t=\tau_{\mathrm{init}}}^{T} (\Delta_t - \overline{\Delta}_t) \leqslant \sum_{t=\tau_{\mathrm{init}}}^{T} \overline{\Delta}_t + 2M_\mu \sqrt{2T \log \frac{6}{\delta}},$$

which completes the proof. □

We now present the proof of Theorem 1.

Proof of Theorem 1. We use $\rho = 1/(64R\sqrt{d}) = \Theta((R\sqrt{d})^{-1})$ as specified in Equation F.1.

By Proposition C.1, with probability at least $1 - \delta/3$, whenever initialization is completed, we have

$$a^\star \in \mathcal{A}_1, \qquad W(\mathcal{A}_1) \leqslant \frac{\rho}{4} = \rho_1. \qquad \text{(Proposition C.1)}$$

Let τ_{init} denote the first round after initialization, or $T+1$ if the horizon is exhausted first. Since the instantaneous regret is at most $2M_\mu$, the cumulative regret incurred during the initialization stage is bounded as follows:

$$\begin{aligned}\textbf{Regret}(\tau_{\text{init}}-1) &\leqslant 2M_\mu N_{\text{init}}\left(\rho, \frac{\delta}{3}\right) \\ &= \widetilde{\mathcal{O}}\big(\kappa_{\text{init}}^2 K d\big)\,. \qquad \text{(Proposition C.1 and } \rho^{-2} = 4096R^2 d\text{)}\end{aligned}$$

Here M_μ, L_μ, R are fixed; if initialization is nonempty, $\kappa_{\text{init}} L_\mu \geqslant 1$, so the ceiling and linear Bernstein terms are absorbed by $\widetilde{\mathcal{O}}(\kappa_{\text{init}}^2 K d)$. If the horizon ends during initialization, the same bound already controls the total regret; otherwise, apply the following post-initialization bounds.

On the initialization event, Lemma E.2 gives, with additional failure probability at most $\delta/3$,

$$\begin{aligned}&\sum_{t=\tau_{\text{init}}}^{T} \mathbb{E}\big[\mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_{A_t}^\top \theta^\star) \,\big|\, \mathcal{G}_t\big] \\ &\quad\leqslant \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} + (R + M_\mu + \sqrt{L_\mu}) d \min\{d, \log K\} + 1\Big)\,. \\ &\qquad\qquad\qquad

Lemma F.2 (First-order Taylor remainder)**.** *For any c and $c+u$ in the feasible domain,*

$$\left|\mu(c+u)-\mu(c)-\mu'(c)u\right| \leqslant \mu'(c)|u|\big(e^{R|u|}-1\big).$$

Proof of Lemma F.2. By the fundamental theorem of calculus,

$$\mu(c+u)-\mu(c)-\mu'(c)u = \int_0^u \big(\mu'(c+v)-\mu'(c)\big)\,dv.$$

Lemma F.1 implies $|\mu'(c+v)-\mu'(c)| \leqslant \mu'(c)(e^{R|v|}-1)$. Bounding the integrand by its maximum over $|v| \leqslant |u|$ gives the result. □

For the concrete choice used in the proof of Theorem 1, set

$$\rho = \frac{1}{64R\sqrt{d}}, \qquad \rho_1 = \frac{\rho}{4}. \tag{F.1}$$

For phase m, define

$$\eta_m := 2\sqrt{d}\big(e^{R\rho_m}-1\big). \tag{F.2}$$

This is the relative nonlinear-error factor used below.

Lemma F.3 (The localization factor is small)**.** *If $\rho_m \leqslant \rho_1$, then $\eta_m \leqslant 1/64$.*

Proof of Lemma F.3. By Equation F.1, we have $R\rho_m \leqslant 1/(256\sqrt{d}) \leqslant 1$. Since $e^x - 1 \leqslant 2x$ for $x \in [0,1]$, we obtain $\eta_m \leqslant 4\sqrt{d}\,R\rho_m \leqslant \frac{1}{64}$. □

F.2 Martingale inequalities

Lemma F.4 (Canonical-family moment-generating function)**.** *Under Equation 1 and Assumptions 2–3, for every t, a and $|\lambda| \leqslant 1/(2R)$,*

$$\log \mathbb{E}\Big[e^{\lambda(Y_t-\mu(z_{t,a}))}\,\Big|\,\mathcal{G}_t, A_t=a\Big] \leqslant L_\mu\lambda^2.$$

Proof. The natural-parameter domain is $\mathbb{R}$, so the canonical identity and Taylor's formula with integral remainder give

$$\begin{aligned}\log \mathbb{E}\Big[e^{\lambda(Y_t-\mu(z_{t,a}))}\,\Big|\,\mathcal{G}_t, A_t=a\Big] &= m(z_{t,a}+\lambda)-m(z_{t,a})-\lambda\mu(z_{t,a})\\ &= \lambda^2\int_0^1 (1-u)\mu'(z_{t,a}+u\lambda)\,du.\end{aligned}$$

The global self-concordance assumption implies $\mu'(z_{t,a}+u\lambda) \leqslant e^{R|\lambda|}\mu'(z_{t,a})$. This follows by integrating the logarithmic derivative, as in Lemma F.1; the nondegenerate canonical family has $\mu'(z) > 0$ throughout its domain. Since $\mu'(z_{t,a}) \leqslant L_\mu$ and $R|\lambda| \leqslant 1/2$, the last display is at most $\frac{1}{2}e^{1/2}L_\mu\lambda^2 \leqslant L_\mu\lambda^2$. No separate reward-tail assumption is needed. □

Lemma F.5 (Conditional Bernstein inequality for weighted rewards)**.** *Suppose Equation 1 and Assumptions 1–3 hold. Let $h_{s,a}$ be $\mathcal{G}_s$-measurable scalar weights such that $\max_a |h_{s,a}| \leqslant H$ and $\sum_a \pi_s(a)h_{s,a}^2 \leqslant V$ for every s, where $H, V > 0$ are deterministic conditional on the start of the sampling procedure. Define $X_s := h_{s,A_s}Y_s - \mathbb{E}[h_{s,A_s}Y_s \mid \mathcal{G}_s]$. For every $n \geqslant 1$ and $x > 0$, conditional on that starting history,*

$$\Pr\left(\left|\sum_{s=1}^n X_s\right| \geqslant \sqrt{16(L_\mu+M_\mu^2)nVx} + 4(R+M_\mu+\sqrt{L_\mu})Hx\right) \leqslant 2e^{-x}.$$

The same conclusion holds for increments stopped at a sampling-termination time.

Proof. Fix s and write

$$U_s := h_{s,A_s}\big(Y_s - \mu(z_{s,A_s})\big), \qquad W_s := h_{s,A_s}\mu(z_{s,A_s}) - \mathbb{E}[h_{s,A_s}\mu(z_{s,A_s}) \mid \mathcal{G}_s].$$

Then $X_s = U_s + W_s$. We bound the conditional moment-generating function for $|\lambda| \leqslant [4(R + M_\mu + \sqrt{L_\mu})H]^{-1}$. First, $|2\lambda h_{s,a}| \leqslant 1/(2R)$, so Lemma F.4 gives

$$\begin{aligned}\mathbb{E}[e^{2\lambda U_s} \mid \mathcal{G}_s] &\leqslant \sum_a \pi_s(a) e^{4L_\mu \lambda^2 h_{s,a}^2} \\ &\leqslant 1 + 8L_\mu\lambda^2 \sum_a \pi_s(a) h_{s,a}^2 \leqslant e^{8L_\mu\lambda^2 V}.\end{aligned}$$

Here $4L_\mu\lambda^2 h_{s,a}^2 \leqslant 1/4$, and we used $e^u \leqslant 1 + 2u$ for $0 \leqslant u \leqslant 1$. For the action-randomization term, bounded conditional means imply

$$\mathbb{E}[W_s \mid \mathcal{G}_s] = 0, \qquad |W_s| \leqslant 2M_\mu H, \qquad \mathbb{E}[W_s^2 \mid \mathcal{G}_s] \leqslant M_\mu^2 V.$$

Since $|2\lambda W_s| \leqslant 1$, the inequality $e^u \leqslant 1 + u + u^2$ for $|u| \leqslant 1$ yields

$$\mathbb{E}[e^{2\lambda W_s} \mid \mathcal{G}_s] \leqslant 1 + 4\lambda^2 M_\mu^2 V \leqslant e^{4\lambda^2 M_\mu^2 V}.$$

Conditional Cauchy–Schwarz, which does not require independence of U_s and W_s, therefore gives

$$\begin{aligned}\log \mathbb{E}[e^{\lambda X_s} \mid \mathcal{G}_s] &\leqslant \tfrac{1}{2}\log \mathbb{E}[e^{2\lambda U_s} \mid \mathcal{G}_s] + \tfrac{1}{2}\log \mathbb{E}[e^{2\lambda W_s} \mid \mathcal{G}_s] \\ &\leqslant (4L_\mu + 2M_\mu^2)V\lambda^2 \leqslant 4(L_\mu + M_\mu^2)V\lambda^2.\end{aligned}$$

Iterated conditioning gives $\mathbb{E}\exp(\lambda\sum_{s=1}^n X_s) \leqslant \exp(4(L_\mu + M_\mu^2)nV\lambda^2)$ on the same interval of λ, conditional on the starting history. For the upper tail, apply Markov's inequality with

$$\lambda = \min\left\{\sqrt{\frac{x}{4(L_\mu + M_\mu^2)nV}},\ \frac{1}{4(R + M_\mu + \sqrt{L_\mu})H}\right\}.$$

If the first term is selected, the square-root threshold alone gives an exponent $-x$. If the second is selected, then $4(L_\mu + M_\mu^2)nV\lambda^2 \leqslant x$, while $\lambda\sqrt{16(L_\mu + M_\mu^2)nVx} \geqslant 4(L_\mu + M_\mu^2)nV\lambda^2$; the linear threshold consequently gives an exponent at most $-x$. Applying the same argument to $-X_s$ proves the two-sided bound. Finally, multiplying each increment by the predictable indicator that sampling is still active preserves its conditional moment-generating-function bound, proving the stopped version. □

Lemma F.6 (Martingale Hoeffding inequality)**.** *Let $(X_s, \mathcal{F}_s)$ be a martingale-difference sequence with $|X_s| \leqslant b$ almost surely. For every $n \geqslant 1$ and $\delta \in (0, 1)$,*

$$\Pr\left(\left|\sum_{s=1}^n X_s\right| \geqslant b\sqrt{2n\log(2/\delta)}\right) \leqslant \delta.$$

Proof. Conditional Hoeffding's lemma gives $\mathbb{E}[e^{\lambda X_s} \mid \mathcal{F}_{s-1}] \leqslant e^{\lambda^2 b^2/2}$. Iterated conditioning and Markov's inequality yield $\Pr(\sum_{s=1}^n X_s \geqslant u) \leqslant \exp(-u^2/(2nb^2))$. Apply the same bound to $-X_s$ and take $u = b\sqrt{2n\log(2/\delta)}$. □

G Proof of the Regret Lower Bound

We prove Theorem 2 using Bernoulli rewards with the logistic mean function $\sigma(z) = (1 + e^{-z})^{-1}$, a special case of Equation 1.

G.1 Hard instance construction.

We take $c_0 = 1$ and $C = 16$ in Theorem 2. Fix $d \geqslant 1$, $T \geqslant 16d^2$, $S \geqslant 1/8$, and $4 \leqslant \kappa_0 \leqslant \sqrt{T}/d$, and fix any adaptive randomized policy. Set $\gamma = 1/32$ and define

$$r := \gamma\frac{\kappa_0 d}{\sqrt{T}} = \frac{\kappa_0 d}{32\sqrt{T}}. \tag{G.1}$$

Then $0 < r \leqslant 1/32$.

Index the $K = 2^d$ arms by $a \in \{-1, +1\}^d$ and define

$$x_a = \frac{a}{\sqrt{d}}, \qquad \theta^\omega = \frac{r}{\sqrt{d}}\omega, \qquad \omega \in \{-1, +1\}^d.$$

Thus, $\|x_a\|_2 = 1$ and $\|\theta^\omega\|_2 = r \leqslant 1/32 \leqslant S$. Let $H(a, \omega) := \sum_{i=1}^d \mathbf{1}\{a_i \neq \omega_i\}$ denote the Hamming distance between a and ω. Then

$$x_a^\top \theta^\omega = \frac{r}{d} a^\top \omega = r\left(1 - \frac{2H(a,\omega)}{d}\right). \tag{G.2}$$

Consequently, all linear scores lie in $[-r, r]$, and $x_a^\top \theta^\omega = r$ if and only if $a = \omega$. Since σ is strictly increasing, the unique optimal arm under θ^ω is x_ω.

Let $B \geqslant 0$ be the unique solution to

$$\sigma'(B) = \frac{1}{\kappa_0}. \tag{G.3}$$

Such a solution exists because $\kappa_0 \geqslant 4$ and σ' decreases continuously from $1/4$ to zero on $[0, \infty)$.

For each $\omega \in \{-1, +1\}^d$ and each sign sequence $\xi = (\xi_1, \ldots, \xi_T) \in \{-1, +1\}^T$, define an instance by setting $\theta^\star = \theta^\omega$ and

$$\nu_t = B\xi_t - r, \qquad t \in [T].$$

The common-shift sequence is fixed before interaction and is not observed by the learner. Write a_t for the sign vector of the selected arm, so that $x_{A_t} = x_{a_t}$. At each round, conditional on the history and $a_t = a$, the reward in this fixed instance is Bernoulli with mean $\sigma(B\xi_t - r + x_a^\top \theta^\omega)$. This defines the instance family used below.

G.2 Proof of Theorem 2

Proof of Theorem 2. For the proof, draw ω uniformly from $\{-1, +1\}^d$ and draw independent Rademacher variables $\xi_1, \ldots, \xi_T$, independently of ω and the learner's randomization. We use $\mathbb{E}$ to denote expectation over ω, the signs $\xi_{1:T}$, the learner's randomization, and the reward observations.

Define the averaged mean function

$$q_B(u) := \frac{1}{2}\big(\sigma(B + u) + \sigma(-B + u)\big). \tag{G.4}$$

Conditional on $\mathcal{H}_{t-1}$, $a_t = a$, and ω, the fresh sign ξ_t still takes each value in $\{-1, +1\}$ with probability $1/2$. Therefore,

$$\begin{aligned}
&\Pr(Y_t = 1 \mid \mathcal{H}_{t-1}, a_t = a, \omega) \\
&\quad = \frac{1}{2}\sigma(B + x_a^\top \theta^\omega - r) + \frac{1}{2}\sigma(-B + x_a^\top \theta^\omega - r) && \text{(averaging over } \xi_t\text{)} \\
&\quad = q_B(x_a^\top \theta^\omega - r). && \text{(Equation G.4)}
\end{aligned}$$

Thus, after averaging over the hidden signs, the observable experiment is a stationary Bernoulli bandit with arm means $q_B(x_a^\top \theta^\omega - r)$. Let $\mathbb{P}_\omega$ denote the law of the full observable history $(a_1, Y_1, \ldots, a_T, Y_T)$ in this experiment, and let $\mathbb{E}_\omega$ denote the corresponding expectation.

By Equation G.2, $x_a^\top \theta^\omega - r = -2rH(a, \omega)/d \in [-2r, 0]$. We first bound q_B' on this interval. For the logistic mean function, Lemma F.1 applies with $R = 1$ and gives, for each $c \in \{-B, B\}$,

$$e^{-|u|}\sigma'(c) \leqslant \sigma'(c + u) \leqslant e^{|u|}\sigma'(c).$$

Since σ' is even, $\sigma'(-B) = \sigma'(B) = 1/\kappa_0$. Averaging the bounds at $c = B$ and $c = -B$ therefore yields

$$\frac{e^{-2r}}{\kappa_0} \leqslant q_B'(u) = \frac{1}{2}\big(\sigma'(B + u) + \sigma'(-B + u)\big) \leqslant \frac{e^{2r}}{\kappa_0}, \qquad u \in [-2r, 0]. \tag{G.5}$$

Moreover, logistic symmetry gives $q_B(0) = 1/2$, and $\sigma' \leqslant 1/4$ implies $q_B' \leqslant 1/4$. Hence, for $u \in [-2r, 0]$, $|q_B(u) - 1/2| \leqslant |u|/4 \leqslant r/2 \leqslant 1/64$. In particular, $q_B(u) \in [1/4, 3/4]$ throughout this interval.

Next, fix $\omega \in \{-1, +1\}^d$ and $i \in [d]$, and let $\omega^{(i)}$ be obtained by flipping the ith coordinate of ω. For every arm a, the parameter construction gives $|x_a^\top(\theta^\omega - \theta^{\omega^{(i)}})| = 2r/d$. Since both shifted scores lie in $[-2r, 0]$,

$$\begin{aligned}
&\left|q_B(x_a^\top \theta^\omega - r) - q_B(x_a^\top \theta^{\omega^{(i)}} - r)\right| \\
&\quad\leqslant \frac{e^{2r}}{\kappa_0}\left|x_a^\top(\theta^\omega - \theta^{\omega^{(i)}})\right| && \text{(Equation G.5 and the mean-value theorem)} \\
&\quad= \frac{2e^{2r}r}{\kappa_0 d}. && \text{(definitions of } x_a, \theta^\omega, \text{ and } \omega^{(i)})
\end{aligned}$$

Write $\mathrm{kl}(p, q)$ for the KL divergence between Bernoulli distributions with means p and q. For $p, q \in [1/4, 3/4]$,

$$\begin{aligned}
\mathrm{kl}(p, q) &\leqslant \frac{(p-q)^2}{q(1-q)} && (\log v \leqslant v - 1) \\
&\leqslant \frac{16}{3}(p-q)^2. && (q(1-q) \geqslant 3/16)
\end{aligned}$$

The learner uses the same action-selection rule under both parameters. Thus, the chain rule for KL divergence gives

$$\begin{aligned}
&D_{\mathrm{KL}}(\mathbb{P}_\omega \| \mathbb{P}_{\omega^{(i)}}) \\
&\quad= \sum_{t=1}^{T} \mathbb{E}_\omega\left[\mathrm{kl}\Big(q_B(x_{a_t}^\top \theta^\omega - r), q_B(x_{a_t}^\top \theta^{\omega^{(i)}} - r)\Big)\right] && \text{(chain rule for KL divergence)} \\
&\quad\leqslant \frac{16T}{3}\left(\frac{2e^{2r}r}{\kappa_0 d}\right)^2 \\
&\quad= \frac{64e^{4r}}{3}\frac{Tr^2}{\kappa_0^2 d^2} \\
&\quad= \frac{e^{4r}}{48} && \text{(Equation G.1)} \\
&\quad\leqslant \frac{e^{1/8}}{48} < \frac{1}{8}. && (r \leqslant 1/32)
\end{aligned}$$

By Pinsker's inequality, for every ω and i,

$$\mathrm{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega^{(i)}}) \leqslant \sqrt{\frac{1}{2}D_{\mathrm{KL}}(\mathbb{P}_\omega \| \mathbb{P}_{\omega^{(i)}})} \leqslant \frac{1}{4}. \tag{G.6}$$

We now bound the number of incorrectly selected coordinates. Fix $t \in [T]$, $i \in [d]$, and ω with $\omega_i = +1$. Since the selected actions are part of the observable history,

$$\begin{aligned}
&\mathbb{P}_\omega(a_{t,i} = -1) + \mathbb{P}_{\omega^{(i)}}(a_{t,i} = +1) \\
&\quad= 1 - \mathbb{P}_\omega(a_{t,i} = +1) + \mathbb{P}_{\omega^{(i)}}(a_{t,i} = +1) \\
&\quad\geqslant 1 - \mathrm{TV}(\mathbb{P}_\omega, \mathbb{P}_{\omega^{(i)}}) && \text{(definition of total variation)} \\
&\quad\geqslant \frac{3}{4}. && \text{(Equation G.6)}
\end{aligned}$$

Expanding the Hamming distance and pairing the parameters that differ only in coordinate i, we obtain

$$\begin{aligned}
&\frac{1}{2^d}\sum_{\omega\in\{-1,+1\}^d}\mathbb{E}_\omega[H(a_t,\omega)] \\
&\quad=\frac{1}{2^d}\sum_{i=1}^{d}\sum_{\omega\in\{-1,+1\}^d}\mathbb{P}_\omega(a_{t,i}\neq\omega_i) && \text{(definition of } H\text{)} \\
&\quad=\frac{1}{2^d}\sum_{i=1}^{d}\sum_{\omega:\,\omega_i=+1}[\mathbb{P}_\omega(a_{t,i}=-1)+\mathbb{P}_{\omega^{(i)}}(a_{t,i}=+1)] && \text{(pairing } \omega \text{ with } \omega^{(i)}\text{)} \\
&\quad\geqslant\frac{1}{2^d}\sum_{i=1}^{d}2^{d-1}\frac{3}{4} \\
&\quad=\frac{3d}{8}. && \text{(G.7)}
\end{aligned}$$

We next translate this uncertainty into regret. Under parameter θ^ω, the optimal arm is x_ω, and $x_\omega^\top\theta^\omega - r = 0$. For every arm a,

$$\begin{aligned}
&q_B(x_\omega^\top\theta^\omega - r) - q_B(x_a^\top\theta^\omega - r) \\
&\quad=\int_{x_a^\top\theta^\omega - r}^{x_\omega^\top\theta^\omega - r} q_B'(u)\,du \\
&\quad\geqslant\frac{e^{-2r}}{\kappa_0}\left(x_\omega^\top\theta^\omega - x_a^\top\theta^\omega\right) && \text{(Equation G.5)} \\
&\quad=\frac{2e^{-2r}r}{\kappa_0 d}H(a,\omega). && \text{(Equation G.2)}
\end{aligned}$$

Conditional on the past, the selected arm, and ω, averaging the original reward gap over the fresh sign ξ_t gives precisely the difference above. Averaging also over ω and summing over rounds therefore yields

$$\begin{aligned}
\mathbb{E}[\textbf{Regret}(T)] &= \frac{1}{2^d}\sum_{\omega\in\{-1,+1\}^d}\sum_{t=1}^{T}\mathbb{E}_\omega\left[q_B(x_\omega^\top\theta^\omega - r) - q_B(x_{a_t}^\top\theta^\omega - r)\right] \\
&\qquad\text{(tower property and Equation G.4)} \\
&\geqslant\frac{2e^{-2r}r}{\kappa_0 d}\sum_{t=1}^{T}\frac{1}{2^d}\sum_{\omega\in\{-1,+1\}^d}\mathbb{E}_\omega[H(a_t,\omega)] \\
&\geqslant\frac{2e^{-2r}r}{\kappa_0 d}\,T\,\frac{3d}{8} && \text{(Equation G.7)} \\
&=\frac{3e^{-2r}}{128}d\sqrt{T} && \text{(Equation G.1)} \\
&\geqslant\frac{3e^{-1/16}}{128}d\sqrt{T}. && (r\leqslant 1/32)
\end{aligned}$$

It remains to select a deterministic instance. The preceding expectation averages over finitely many choices of $(\omega, \xi_{1:T})$. Hence, there exist $\omega^\star\in\{-1,+1\}^d$ and $\xi^\star\in\{-1,+1\}^T$ such that

$$\mathbb{E}[\textbf{Regret}(T)\,|\,\omega=\omega^\star,\,\xi_{1:T}=\xi^\star]\geqslant\frac{3e^{-1/16}}{128}d\sqrt{T}.$$

Fix this realization and take $\theta^\star = \theta^{\omega^\star}$ and $\nu_t = B\xi_t^\star - r$. The common-shift sequence is now deterministic and fixed in advance. For this instance, the expectation is only over the learner's randomization and the reward observations.

Finally, we verify that the same instance has $\kappa^\star = \kappa_0$. Its unique optimal arm is $x_{a^\star} = x_{\omega^\star}$, with $x_{a^\star}^\top \theta^\star = r$. Thus, for every $t \in [T]$,

$$\begin{aligned}\sigma'(\nu_t + x_{a^\star}^\top \theta^\star) &= \sigma'(B\xi_t^\star - r + r) \\ &= \sigma'(B) && (\xi_t^\star \in \{-1, +1\} \text{ and the symmetry of } \sigma') \\ &= \frac{1}{\kappa_0}. && \text{(Equation G.3)}\end{aligned}$$

By the definition of $\kappa^\star$, we have

$$\kappa^\star = \left(\frac{1}{T}\sum_{t=1}^{T} \sigma'(\nu_t + x_{a^\star}^\top \theta^\star)\right)^{-1} = \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{\kappa_0}\right)^{-1} = \kappa_0.$$

The constructed deterministic instance therefore satisfies $\kappa^\star = \kappa_0$ and $\mathbb{E}[\textbf{Regret}(T)] \geqslant (3e^{-1/16}/128)d\sqrt{T} = \Omega(d\sqrt{T})$, with a universal implied constant. This concludes the proof of Theorem 2. □

G.3 Comparison with ordinary logistic bandits

Theorem 2 shows a qualitatively different role of local curvature from that in ordinary logistic bandits. We first recall the following global lower bound of Abeille et al. (2021), which provides a useful benchmark for this comparison.

Proposition G.1 (Global lower bound for ordinary logistic bandits)**.** *Let $\mathcal{X} = \mathbb{S}_d(0,1) := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. For every policy π and every $\kappa_0 \geqslant \frac{1}{\sigma'(1)}$ satisfying $T \geqslant d^2\kappa_0$, there exists a logistic-bandit instance $\theta^\star$ with $\kappa^\star = \kappa_0$ such that*

$$\mathbb{E}[\textbf{Regret}(T)] = \Omega\left(d\sqrt{\frac{T}{\kappa_0}}\right),$$

where $\kappa^\star := 1/\sigma'(x_{a^\star}^\top \theta^\star)$.

Proposition G.1 follows from Corollary 2 of Abeille et al. (2021) in the large-curvature regime $\kappa^\star \geqslant 1/\sigma'(1) \approx 5.09$. It gives an $\Omega(d\sqrt{T/\kappa^\star})$ lower-bound scale for ordinary logistic bandits at a fixed local-curvature level. In contrast, Theorem 2 shows that, under hidden common shifts, the regret can remain $\Omega(d\sqrt{T})$ even as $\kappa^\star$ grows to order $\sqrt{T}/d$. We next explain this difference through the information carried by a single reward observation.

Ordinary logistic bandit. Consider two nearby natural parameters $\bar{z} + u$ and $\bar{z} + v$, where $u, v \in [-2r, 0]$, and let $\Delta := v - u$ and $\alpha := \sigma'(\bar{z}) = 1/\kappa_0$. Define $P_u := \text{Bernoulli}(\sigma(\bar{z} + u))$ and $P_v := \text{Bernoulli}(\sigma(\bar{z} + v))$. For the logistic log-partition function $m(z) = \log(1 + e^z)$, the second-order Taylor theorem gives $D_{\mathrm{KL}}(P_u \| P_v) = \frac{1}{2}\sigma'(\zeta_{u,v})\Delta^2$ for some $\zeta_{u,v}$ between $\bar{z} + u$ and $\bar{z} + v$. Since $|\zeta_{u,v} - \bar{z}| \leqslant 2r$, Lemma F.1 yields

$$\frac{e^{-2r}}{2}\alpha\Delta^2 \leqslant D_{\mathrm{KL}}(P_u \| P_v) \leqslant \frac{e^{2r}}{2}\alpha\Delta^2. \tag{G.8}$$

Hence, $D_{\mathrm{KL}}(P_u \| P_v) = \Theta(\alpha\Delta^2)$.

For a d-dimensional hard family in which neighboring parameters differ in score by order r/d, the total KL over T rounds is therefore of order $T\alpha r^2/d^2$. Keeping neighboring instances indistinguishable requires $r = \Theta(d/\sqrt{\alpha T})$. Since the corresponding reward gap is of order αr, the cumulative regret scale becomes

$$T\alpha r = \Theta\left(d\sqrt{\alpha T}\right) = \Theta\left(d\sqrt{\frac{T}{\kappa_0}}\right). \tag{G.9}$$

This reproduces the information-theoretic scaling in Proposition G.1.

Semiparametric hidden-shift instance. We now consider the hard instance used in the proof of Theorem 2. Recall that each round contains a fresh unobserved sign $\xi_t \in \{-1, +1\}$. Conditional on ξ_t, a shifted score u produces a Bernoulli reward with mean $\sigma(B\xi_t + u)$. Since ξ_t is unobserved, the learner sees the sign-marginalized Bernoulli law

$$Q_u := \text{Bernoulli}(q_B(u)), \qquad q_B(u) = \frac{1}{2}\big(\sigma(B + u) + \sigma(-B + u)\big),$$

where $\alpha := \sigma'(B) = 1/\kappa_0$ and $u \in [-2r, 0]$. Thus, $D_{\mathrm{KL}}(Q_u \| Q_v)$ is exactly the one-step observable KL entering the chain-rule argument in the lower-bound proof.

By Equation G.5, $|q_B(v) - q_B(u)| = \Theta(\alpha|\Delta|)$ for $\Delta = v - u$. Moreover, $q_B(0) = 1/2$ and $q_B(u) \in [1/4, 3/4]$ on $[-2r, 0]$. Since $\mathrm{kl}(p, q) = \Theta((p-q)^2)$ when $p, q \in [1/4, 3/4]$, we obtain

$$2e^{-4r}\alpha^2\Delta^2 \leqslant D_{\mathrm{KL}}(Q_u \| Q_v) \leqslant \frac{16}{3}e^{4r}\alpha^2\Delta^2. \tag{G.10}$$

Hence, $D_{\mathrm{KL}}(Q_u \| Q_v) = \Theta(\alpha^2\Delta^2)$. Unlike in the ordinary logistic model, the observable reward mean remains close to $1/2$, so its variance stays of constant order even though the mean difference is only of order $\alpha\Delta$.

For the hypercube family in our proof, neighboring parameters differ in score by order r/d, so the total observable KL is now of order $T\alpha^2 r^2/d^2$. Keeping neighboring instances indistinguishable therefore allows $r = \Theta(d/(\alpha\sqrt{T})) = \Theta(\kappa_0 d/\sqrt{T})$, as in Equation G.1. Since the reward gap is still of order αr, the corresponding regret scale is

$$T\alpha r = \Theta(d\sqrt{T}). \tag{G.11}$$

The comparison explains why the favorable local-curvature dependence of ordinary logistic bandits does not carry over to the hidden-shift setting. In an ordinary logistic model, a small local slope reduces the difference between nearby reward means, but it also reduces the intrinsic Bernoulli variance at the same operating point. The resulting loss of statistical information is therefore only linear in the local slope. Under the hidden shifts used in our lower bound, the differences between observable arm means are attenuated by the same local slope, while the observable reward variance remains of constant order. Nearby instances therefore become harder to distinguish by an additional factor of the local slope. This exactly offsets the smaller reward gaps and leaves the semiparametric problem $\Omega(d\sqrt{T})$-hard even when the local-curvature parameter is large. Thus, local curvature alone no longer characterizes the statistical difficulty of the problem.

H WHY A DIRECT REDUCTION TO SEMIPARAMETRIC LINEAR BANDITS FAILS

A natural question is whether SGLBs can be reduced much more simply to semiparametric linear bandits, even if one is willing to pay a global curvature factor κ in the regret bound. Indeed, a reader might expect that, once

$$\frac{1}{\kappa} \leqslant \mu'(z) \leqslant L_\mu$$

is imposed over the relevant natural-parameter range, the nonlinear reward differences can be related to linear score differences and an existing semiparametric linear-bandit method, such as that of Kim et al. (2025), can then be applied directly. This seemingly natural reduction, however, does not go through. As we show below, even after paying the curvature dependence through κ, these bounds are sufficient to reduce the *regret objective* to a linear one, but they do not reduce the *observed feedback* to the semiparametric linear model required by the existing analysis.

Reducing the regret objective. For every round t, the mean-value theorem gives some ξ_t between $\nu_t + x_{A_t}^\top\theta^\star$ and $\nu_t + x_{a^\star}^\top\theta^\star$ such that

$$\begin{aligned}\mu(\nu_t + x_{a^\star}^\top\theta^\star) - \mu(\nu_t + x_{A_t}^\top\theta^\star) &= \mu'(\xi_t)(x_{a^\star} - x_{A_t})^\top\theta^\star \\ &\leqslant L_\mu(x_{a^\star} - x_{A_t})^\top\theta^\star.\end{aligned} \tag{H.1}$$

Therefore, we have

$$\mathbf{Regret}(T) \leqslant L_\mu \sum_{t=1}^{T}(x_{a^\star} - x_{A_t})^\top\theta^\star. \tag{H.2}$$

Thus, it would indeed suffice to control the latent linear-score regret on the right-hand side of Equation H.2. A natural question is whether one can simply apply a semiparametric linear-bandit method such as that of Kim et al. (2025).

Centered estimating equation in the linear model. Fix a sampling distribution p, possibly chosen as a G-optimal design (Kim et al., 2025), and define

$$\bar{x} := \sum_a p(a) x_a, \qquad z_a := x_a - \bar{x}, \qquad \Sigma := \sum_a p(a) z_a z_a^\top.$$

In the additive semiparametric linear model, the common shift enters additively:

$$\mathbb{E}[Y_t \mid \mathcal{G}_t, A_t = a] = x_a^\top \theta^\star + \nu_t.$$

Since ν_t is action independent, feature centering removes it exactly:

$$\begin{aligned}
\mathbb{E}[z_{A_t} Y_t \mid \mathcal{G}_t] &= \sum_a p(a) z_a \left(x_a^\top \theta^\star + \nu_t \right) \\
&= \sum_a p(a) z_a \left(z_a^\top \theta^\star + \bar{x}^\top \theta^\star + \nu_t \right) \\
&= \sum_a p(a) z_a z_a^\top \theta^\star \qquad (\textstyle\sum_a p(a) z_a = 0) \\
&= \Sigma \theta^\star. \qquad \text{(H.3)}
\end{aligned}$$

This is the key centered estimating equation in the linear model. Under a fixed design, the population analogue of the centered regression estimator is

$$\frac{1}{n} \Sigma^\dagger \sum_{t=1}^{n} z_{A_t} Y_t,$$

whose conditional population mean is $\Sigma^\dagger \Sigma \theta^\star$. Since $\Sigma^\dagger \Sigma$ acts as the identity on the active difference span, centered regression recovers $\theta^\star$ on every relevant direction, up to stochastic estimation error. Thus, the same parameter that determines the latent linear regret in Equation H.2 is directly identified from the centered feedback.

Testing the same linear estimating equation under SGLB feedback. We now apply the same centering calculation to the SGLB observations. This step concerns a direct use of the centered linear estimating equation; it is distinct from fitting the nonlinear model by likelihood-based estimation.

Let $c_t := \nu_t + \bar{x}^\top \theta^\star$. Then, we have

$$\begin{aligned}
\mathbb{E}[z_{A_t} Y_t \mid \mathcal{G}_t] &= \sum_a p(a) z_a \mu(c_t + z_a^\top \theta^\star) \\
&= \sum_a p(a) z_a \left[\mu(c_t) + \alpha_{t,a} z_a^\top \theta^\star \right] \\
&= \sum_a p(a) \alpha_{t,a} z_a z_a^\top \theta^\star,
\end{aligned}$$

where $\alpha_{t,a} := \int_0^1 \mu'\left(c_t + u z_a^\top \theta^\star\right) \, du$. Thus, centering still removes the common level $\mu(c_t)$, but the nonlinear link leaves an arm-dependent local slope $\alpha_{t,a}$.

Define $H_t := \sum_a p(a) \alpha_{t,a} z_a z_a^\top$. The centered population moment is therefore

$$\mathbb{E}[z_{A_t} Y_t \mid \mathcal{G}_t] = H_t \theta^\star.$$

Under the additional global curvature condition, we have $\frac{1}{\kappa} \leqslant \alpha_{t,a} \leqslant L_\mu$, which implies

$$\frac{1}{\kappa} \Sigma \preceq H_t \preceq L_\mu \Sigma.$$

Hence, paying κ and L_μ does control the magnitude and conditioning of the centered population operator. However, it does not recover the linear estimating equation Equation H.3.

To see the obstruction directly, consider the centered linear residual that would appear in the semiparametric linear regression:

$$\begin{aligned}
\mathbb{E}\left[z_{A_t} \left(Y_t - z_{A_t}^\top \theta^\star \right) \middle| \mathcal{G}_t \right] &= \mathbb{E}[z_{A_t} Y_t \mid \mathcal{G}_t] - \mathbb{E}[z_{A_t} z_{A_t}^\top \mid \mathcal{G}_t] \theta^\star \\
&= (H_t - \Sigma) \theta^\star.
\end{aligned}$$

In the additive semiparametric linear model, this conditional mean is zero. For an SGLB, it is generally nonzero. Consequently, the residual sum contains a predictable drift in addition to martingale fluctuations.

The matrix inequalities above do not remove this drift. They only imply quadratic-form bounds of the form

$$\frac{1}{\kappa}u^\top \Sigma u \leqslant u^\top H_t u \leqslant L_\mu u^\top \Sigma u$$

for every vector u. They do not imply $H_t = \Sigma$, nor do they imply that $H_t\theta^\star$ differs from $\Sigma\theta^\star$ only by stochastic noise. In particular, if the same $H_t = H$ persists over many rounds, then the population average of the centered linear statistic converges to $\Sigma^\dagger H\theta^\star$, rather than to $\Sigma^\dagger\Sigma\theta^\star$. Thus, increasing the number of samples does not make this discrepancy vanish.

This is where a direct application of the centered semiparametric linear analysis breaks. The reduction in Equation H.1 successfully reduces the regret objective to a latent linear-score regret, but the centered SGLB feedback does not provide the mean-zero linear estimating equation used to learn the parameter governing that regret. Simply enlarging the linear confidence radius by factors depending on κ and L_μ does not remove this predictable bias.

Why standard GLM estimation is not a direct substitute. The preceding obstruction concerns the direct use of a linear centered estimator. For an ordinary GLB (Filippi et al., 2010), one would instead use an estimator adapted to the nonlinear link, such as a maximum-likelihood (MLE) estimator. Indeed, if the common shift ν_t were observed, the correctly specified negative log-likelihood contribution would be

$$\ell_t(\theta;\nu_t) = m(\nu_t + x_{A_t}^\top\theta) - Y_t(\nu_t + x_{A_t}^\top\theta).$$

Its score at the true parameter satisfies

$$\nabla_\theta\ell_t(\theta^\star;\nu_t) = x_{A_t}\left[\mu(\nu_t + x_{A_t}^\top\theta^\star) - Y_t\right], \qquad \mathbb{E}[\nabla_\theta\ell_t(\theta^\star;\nu_t) \mid \mathcal{G}_t, A_t] = 0.$$

Thus, if ν_t were known, the usual GLM estimating equation would remain correctly specified.

In our setting, however, ν_t is unobserved. The correctly specified likelihood above is therefore not available as a function of θ alone. If one simply omits the shift and uses the ordinary GLM loss

$$\ell_t^0(\theta) := m(x_{A_t}^\top\theta) - Y_t x_{A_t}^\top\theta,$$

then its score at $\theta^\star$ has conditional mean

$$\mathbb{E}[\nabla\ell_t^0(\theta^\star) \mid \mathcal{G}_t, A_t] = x_{A_t}\left[\mu(x_{A_t}^\top\theta^\star) - \mu(\nu_t + x_{A_t}^\top\theta^\star)\right],$$

which is generally nonzero. Hence, the usual mean-zero score argument for a correctly specified GLM is not recovered by simply replacing the linear estimator with an ordinary GLM likelihood.

Nor does feature centering by itself restore such a likelihood equation. For example, fitting a GLM using only the centered predictor $z_{A_t}^\top\theta$ would give the population score

$$\mathbb{E}\left[z_{A_t}\left(\mu(z_{A_t}^\top\theta^\star) - Y_t\right) \middle| \mathcal{G}_t\right] = \sum_a p(a)z_a\left[\mu(z_a^\top\theta^\star) - \mu(c_t + z_a^\top\theta^\star)\right],$$

which is generally nonzero. Thus, feature centering does not recover the mean-zero score structure of a correctly specified GLM, and simply replacing the linear estimator with an ordinary GLM likelihood does not yield a direct reduction.

Hence, the usual mean-zero score argument for a correctly specified GLM is not recovered by simply replacing the linear estimator with an ordinary GLM likelihood.

How localization resolves the obstruction. Our approach avoids estimating the shifts. When the active-arm score width is small, self-concordance implies that the arm-dependent local slopes are close to the common slope $\alpha_t := \mu'(c_t)$. Equivalently, the centered population moment admits the decomposition

$$H_t\theta^\star = \alpha_t\Sigma\theta^\star + \text{controlled remainder}.$$

This restores, up to a controlled nonlinear error, the common-scale structure needed to compare the latent linear scores. Lemmas D.1 and D.2 make this approximation precise. The adaptive stopping

rule then handles the unknown accumulated scale $\sum_t \alpha_t$ without requiring a global lower bound on μ' after initialization.

We emphasize that the discussion above concerns direct reductions to existing semiparametric linear-bandit or standard nuisance-free GLM analyses. It does not rule out separately designed κ-dependent algorithms for SGLBs.

I Improved Initial Localization

The initialization cost in Theorem 1 comes from bringing the full action set into the localized regime required by the main elimination procedure. In this section, we give two refinements of this step. Both localize the active set successively over a geometric schedule of score widths, using coarse comparisons when the active set is wide and increasing the comparison accuracy only as the set becomes localized.

The first refinement repeatedly applies the tournament and filtering procedure from Algorithm 1. This improves the initialization regret from $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 K d)$ to $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 K\sqrt{d})$. The second refinement additionally exploits the geometry of the known arm features. It groups arms with similar predicted scores and uses a polynomial number of informative comparisons to certify that the true score widths within these groups are small. This removes the dependence on K from the initialization regret, at the cost of a larger polynomial dependence on d and κ_{init}.

The two initializers have complementary regret bounds. Since these bounds depend only on known quantities such as $K, d, S, M_\mu, R, L_\mu, \kappa_{\mathrm{init}}, T$, and δ, the learner can choose the initializer with the smaller bound before interaction.

Theorem I.1 (Regret with improved initialization)**.** *Suppose Assumptions 1–4 hold, with the same known constant κ_{init} valid throughout either initialization procedure described below. Set $\rho = 1/(64R\sqrt{d})$. Then, with probability at least $1-\delta$, choosing the initializer with the smaller deterministic regret bound and subsequently running the localized elimination procedure of Algorithm 1 gives*

$$\textbf{Regret}(T) = \widetilde{\mathcal{O}}\Big(\min\Big\{\kappa_{\mathrm{init}}^2 K\sqrt{d},\ \kappa_{\mathrm{init}}^4 d^{9/2}\Big\} + \sqrt{dT\min\{d,\log K\}} + d\min\{d,\log K\}\Big)\,.$$

Discussion of Theorem I.1. The improvement concerns only the one-time cost of reaching the localized regime; the subsequent elimination procedure and its regret analysis are unchanged. The successive tournament initializer (Algorithm I.1) improves the initialization cost to $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 K\sqrt{d})$, while the geometric initializer removes the dependence on K entirely, giving $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^4 d^{9/2})$ with no hidden logarithmic dependence on K. As in our main algorithm, these improvements require neither a minimum arm gap nor additional assumptions on the arm geometry.

Throughout this section, both initializers are designed to return an active set satisfying $a^\star \in \mathcal{A}_1$ and $W(\mathcal{A}_1) \leqslant \rho/4$, as required by Equation 2. We use j to index initialization stages and reserve m for the subsequent elimination phases.

Proof of Theorem I.1. For the prescribed deterministic widths $W_j = 2S\,4^{-j}$ and number of stages P, compare before interaction the two deterministic bounds $\sum_{j=0}^{P-1} L_\mu W_j N_{\mathrm{tour}}(W_j)$ and $\sum_{j=0}^{P-1} L_\mu W_j N_{\mathrm{stage}}(W_j, \delta/(3P))$, using the exact budgets defined below and the same known κ_{init}. Choose the initializer with the smaller bound. If $K = 1$ or $2S \leqslant \rho/4$, both initializers require no observations. Otherwise, the subsequent proofs show that these bounds are respectively $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 K\sqrt{d})$ and $\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^4 d^{9/2})$ for fixed S, M_μ, L_μ, R. Only the chosen initializer is run, so its comparison event has failure probability at most $\delta/3$; no additional union bound over the two procedures is needed. If initialization is truncated, its partial regret is already bounded by the selected budget. Otherwise, its output satisfies Equation 2. Lemmas E.2 and E.3 contribute at most $2\delta/3$ additional failure probability and the unchanged post-initialization regret order. Adding the two regret contributions proves Theorem I.1. □

I.1 SUCCESSIVE TOURNAMENT INITIALIZATION

The first refinement uses the same tournament and filtering primitives as Step 1 of Algorithm 1, but changes how their comparison tolerances are scheduled. The original initializer uses the final target scale ρ from the beginning: it performs one tournament on the full action set with tolerance of order $\rho/\log K$, followed by filtering with tolerance of order ρ. In contrast, the *successive initializer* localizes the active set gradually, using a comparison tolerance matched to the current score width. The difference is summarized below:

	Original initializer	Successive initializer
width reduction	$2S \longrightarrow \rho/4$	$W_j \longrightarrow W_j/4$
tournament tolerance	$\dfrac{\rho}{8\lceil \log_2 K \rceil}$	$\dfrac{W_j}{8L_j}$
filtering tolerance	$\dfrac{\rho}{8}$	$\dfrac{W_j}{8}$
initialization regret	$\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 K d)$	$\widetilde{\mathcal{O}}(\kappa_{\mathrm{init}}^2 K\sqrt{d})$

where $L_j := \lceil \log_2 |\mathcal{B}_j| \rceil$.

We now describe the successive initializer. Set $\mathcal{B}_0 = [K]$ and $W_0 = 2S$. If $2S \leqslant \rho/4$ or $K = 1$, no initialization is needed. Otherwise, let

$$P := \left\lceil \log_4 \frac{8S}{\rho} \right\rceil, \qquad W_j := 2S\,4^{-j}, \quad j = 0, \dots, P.$$

The procedure maintains $a^\star \in \mathcal{B}_j$ and $W(\mathcal{B}_j) \leqslant W_j$, while reducing the prescribed width by a factor of four at each stage.

Fix stage j and suppose that $a^\star \in \mathcal{B}_j$ and $W(\mathcal{B}_j) \leqslant W_j$. Set $L_j := \lceil \log_2 |\mathcal{B}_j| \rceil$ and $\varepsilon_j := W_j/(8L_j)$. If $|\mathcal{B}_j| > 1$, we first run a *balanced tournament* on $\mathcal{B}_j$ to obtain a near-optimal reference arm. In each tournament round, the remaining arms are paired and compared using the pairwise test with tolerance ε_j. If the test selects one arm, that arm advances; if the comparison is unresolved, either arm may advance. When the number of remaining arms is odd, one arm advances without a comparison. The tournament continues until a single arm remains, denoted by $\widehat{a}_j$.

The tournament has at most L_j rounds. A confident comparison never advances an arm with a smaller linear score, while an unresolved comparison can decrease the score along a tournament path by at most ε_j. Therefore, Lemma C.2 gives

$$x_{a^\star}^\top \theta^\star - x_{\widehat{a}_j}^\top \theta^\star \leqslant L_j \varepsilon_j = \frac{W_j}{8}.$$

We then use $\widehat{a}_j$ as a reference to filter the current active set. For every $a \in \mathcal{B}_j \backslash \{\widehat{a}_j\}$, we compare $\widehat{a}_j$ and a using tolerance $W_j/8$. The arm a is removed only when the pairwise test selects $\widehat{a}_j$; otherwise, it is retained. Let $\mathcal{B}_{j+1}$ denote the surviving set. On the event that all pairwise tests are valid, Lemma C.3, with ρ there replaced by W_j, gives

$$a^\star \in \mathcal{B}_{j+1}, \qquad W(\mathcal{B}_{j+1}) \leqslant \frac{W_j}{4} = W_{j+1}.$$

We repeat the same procedure until either the active set becomes a singleton or the prescribed width reaches $\rho/4$.

The improvement comes from matching the comparison tolerance to the current score width. Up to logarithmic and problem-dependent factors, stage j requires order $1/W_j^2$ comparison samples, while every sampled arm incurs regret of order at most W_j. Its regret therefore scales as $1/W_j$. Summing over the geometric schedule is dominated by the final scale $W_j \asymp \rho$, giving a $1/\rho$ rather than $1/\rho^2$ dependence in the initialization regret. Since $\rho = \Theta(d^{-1/2})$, this improves the dimension dependence from d to $\sqrt{d}$.

Corollary I.1 (Regret with successive tournament initialization)**.** *Suppose Assumptions 1–4 hold, with Assumption 4 imposed throughout Algorithm I.1. Set $\rho = 1/(64R\sqrt{d})$. Then, with probability at*

Algorithm I.1 Successive tournament initial localization

Require: $\mathcal{X}, T, \delta, \rho, \kappa_{\text{init}}, S, M_\mu, L_\mu, R$
Set $\mathcal{B}_0 \leftarrow [K]$ and $W_0 \leftarrow 2S$.
if $W_0 \leqslant \rho/4$ or $|\mathcal{B}_0| = 1$ **then**
  return $\mathcal{B}_0$
end if
Set $P \leftarrow \lceil \log_4(8S/\rho) \rceil$ and $\delta_{\text{pair}} \leftarrow \delta/(6PK)$.
If the horizon is reached during a pairwise test, terminate without using the incomplete test.
for $j = 0, \dots, P-1$ **do**
  if $|\mathcal{B}_j| = 1$ **then**
    return $\mathcal{B}_j$
  end if
  Set $L_j \leftarrow \lceil \log_2 |\mathcal{B}_j| \rceil$ and $\varepsilon_j \leftarrow W_j/(8L_j)$.
  Tournament: Run a balanced tournament on $\mathcal{B}_j$ using tolerance ε_j and failure probability δ_{pair} per test; let $\widehat{a}_j$ be the winner.
  Filtering: Compare $\widehat{a}_j$ with every $a \in \mathcal{B}_j \backslash \{\widehat{a}_j\}$ using tolerance $W_j/8$ and failure probability δ_{pair} per test.
  Set $\mathcal{B}_{j+1} \leftarrow \{\widehat{a}_j\} \cup \{a \in \mathcal{B}_j \backslash \{\widehat{a}_j\} : \text{the test does not select } \widehat{a}_j \text{ over } a\}$.
  Set $W_{j+1} \leftarrow W_j/4$.
end for
return $\mathcal{B}_P$.

least $1-\delta$, the successive tournament initializer followed by the localized elimination procedure of Algorithm 1 satisfies

$$\textbf{Regret}(T) = \widetilde{\mathcal{O}}\Big(\kappa_{\text{init}}^2 K\sqrt{d} + \sqrt{dT\min\{d, \log K\}} + d\min\{d, \log K\}\Big).$$

Proof of Corollary I.1. If initialization is empty, its regret is zero. Otherwise, each stage uses fewer than $2K$ pairwise tests, and there are at most P stages. Since every test is assigned failure probability $\delta_{\text{pair}} = \delta/(6PK)$, a union bound shows that all initialization comparisons are simultaneously valid with probability at least $1 - \delta/3$. We work on this event.

At stage j, Lemma C.2 gives

$$(x_{a^\star} - x_{\widehat{a}_j})^\top \theta^\star \leqslant \frac{W_j}{8}.$$

Applying Lemma C.3, with ρ there replaced by W_j, then yields

$$a^\star \in \mathcal{B}_{j+1}, \qquad W(\mathcal{B}_{j+1}) \leqslant \frac{W_j}{4} = W_{j+1}.$$

Thus, by induction, every completed stage retains an optimal arm and reduces the active score width by a factor of four. If all P stages are completed, then $W_P \leqslant \rho/4$, so the output satisfies Equation 2. If the active set becomes a singleton earlier, its width is zero and the same guarantee holds.

We next bound the initialization regret. Let $\ell_K := \max\{1, \lceil \log_2 K \rceil\}$ and define

$$N_{\text{tour}}(W) := (K-1)N_{\text{pair}}\left(\frac{W}{8\ell_K}, \delta_{\text{pair}}\right) + (K-1)N_{\text{pair}}\left(\frac{W}{8}, \delta_{\text{pair}}\right). \tag{I.1}$$

Since $L_j \leqslant \ell_K$, this quantity deterministically upper bounds the number of observations used at any stage with prescribed width W.

If initialization contains at least one observation, Assumption 4 and bounded means imply $S \leqslant M_\mu \kappa_{\text{init}}$. Indeed, at any initialization round,

$$\frac{2S}{\kappa_{\text{init}}} \leqslant \int_{-S}^{S} \mu'(\nu_t + u)\, du = \mu(\nu_t + S) - \mu(\nu_t - S) \leqslant 2M_\mu.$$

Thus $W \leqslant 2S \leqslant 2M_\mu \kappa_{\text{init}}$. For every tolerance $\varepsilon \leqslant W$ used in the stage,

$$\frac{(R + M_\mu + \sqrt{L_\mu})\kappa_{\text{init}}}{\varepsilon} \leqslant \frac{2M_\mu(R + M_\mu + \sqrt{L_\mu})\kappa_{\text{init}}^2}{\varepsilon^2} \leqslant \frac{3(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2}{\varepsilon^2},$$

where $2M_\mu\sqrt{L_\mu} \leqslant M_\mu^2 + L_\mu$. Also $(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2/\varepsilon^2 \geqslant 1/4$, so the ceiling is absorbed as well. Using Equation C.1, we obtain

$$N_{\text{tour}}(W) = \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2 K}{W^2}\right).$$

Every arm sampled at stage j belongs to $\mathcal{B}_j$. Since $a^\star \in \mathcal{B}_j$ and $W(\mathcal{B}_j) \leqslant W_j$, the mean-value theorem and $\mu' \leqslant L_\mu$ imply that every initialization sample collected at stage j incurs regret at most $L_\mu W_j$:

$$\mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_{A_t}^\top \theta^\star) \leqslant L_\mu W_j.$$

Therefore, we have

$$\textbf{Regret}(\tau_{\text{init}} - 1) \leqslant \sum_{j=0}^{P-1} L_\mu W_j N_{\text{tour}}(W_j) = \widetilde{\mathcal{O}}\left((L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2 L_\mu K \sum_{j=0}^{P-1} W_j^{-1}\right).$$

By the definition of P, the final prescribed width satisfies $\rho/4 < W_{P-1} \leqslant \rho$. Reversing the geometric sum therefore gives

$$\sum_{j=0}^{P-1} W_j^{-1} = W_{P-1}^{-1} \sum_{k=0}^{P-1} 4^{-k} < \frac{16}{3\rho}.$$

Consequently, we derive

$$\begin{aligned}\textbf{Regret}(\tau_{\text{init}} - 1) &= \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2 L_\mu K}{\rho}\right) \\ &= \widetilde{\mathcal{O}}\left(R(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2 L_\mu K\sqrt{d}\right). \qquad (\rho^{-1} = 64R\sqrt{d})\end{aligned}$$

The same bound applies if the horizon truncates the final stage, since the observations collected before truncation are no more than the full stage budget.

If initialization is not completed before the horizon ends, the preceding bound already controls $\textbf{Regret}(T)$. Otherwise, the output satisfies Equation 2, and the post-initialization procedure is exactly that analyzed in Theorem 1. Applying Lemmas E.2 and E.3 with the remaining failure probability $2\delta/3$ and combining the two regret bounds gives

$$\textbf{Regret}(T) = \widetilde{\mathcal{O}}\begin{pmatrix} R(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2 L_\mu K\sqrt{d} \\ + \sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d, \log K\}} \\ + (R + M_\mu + \sqrt{L_\mu})d\min\{d, \log K\} + 1 \end{pmatrix}.$$

This concludes the proof of Corollary I.1. □

I.2 Geometric initialization

The second refinement removes the K-dependence of the initialization term by exploiting the arm geometry. At each stage, we maintain an ellipsoid containing $\theta^\star$ and group arms according to their predicted scores. When two arms in the same group can still have substantially different true scores, a randomized pairwise comparison produces a valid shallow cut of the ellipsoid. After a polynomial number of such cuts, every group has small true score width, and the tournament and filtering steps need to compare only one representative from each group.

For a stage with input active set $\mathcal{A}$ and width bound $W > 0$, we maintain an ellipsoid

$$\mathcal{E}(c, H) := \left\{\theta \in \mathbb{R}^d : (\theta - c)^\top H^{-1}(\theta - c) \leqslant 1\right\}, \qquad H \succ 0,$$

that contains $\theta^\star$ on the high-probability comparison event. For a vector h, write $\|h\|_H := \sqrt{h^\top H h}$.

Algorithm I.2 Geometric initial localization

Require: $\mathcal{X}, T, \delta, \rho, \kappa_{\text{init}}, S, M_\mu, R, L_\mu$
Set $\mathcal{B}_0 \leftarrow [K]$ and $W_0 \leftarrow 2S$.
if $W_0 \leqslant \rho/4$ or $|\mathcal{B}_0| = 1$ **then**
  return $\mathcal{B}_0$
end if
Set $P \leftarrow \lceil \log_4(8S/\rho) \rceil$.
for $j = 0, \dots, P-1$ **do**
  if $|\mathcal{B}_j| = 1$ **then**
    return $\mathcal{B}_j$
  end if
  Run Algorithm I.3 with $(\mathcal{A}, W, \delta_{\text{stage}}) \leftarrow (\mathcal{B}_j, W_j, \delta/(3P))$.
  if the horizon is exhausted **then**
    stop
  end if
  Let $\mathcal{B}_{j+1}$ be its output and set $W_{j+1} \leftarrow W_j/4$.
end for
return $\mathcal{B}_P$.

Define

$$
\begin{aligned}
w &:= \frac{W}{64 \max\{1, \kappa_{\text{init}} L_\mu\}}, & b &:= \frac{w}{33d}, \\
e &:= \frac{W}{128 \kappa_{\text{init}}}, & J_{\max} &:= \left\lceil \frac{2W}{b} \right\rceil + 2, \\
\ell &:= \max\{1, \lceil \log_2 J_{\max} \rceil\}, & N_{\text{cut}}(W) &:= \left\lceil 8d^2 \log \frac{4S}{b} \right\rceil + 1.
\end{aligned}
\tag{I.2}
$$

For stage failure probability δ_{stage}, set

$$
\delta_{\text{pair}} := \frac{\delta_{\text{stage}}}{N_{\text{cut}}(W) + 2J_{\max}}.
$$

The pairwise tests are those of Lemma C.1. For the final representative filtering step, define

$$
N_{\text{filter}}(e, \delta_{\text{pair}}) := \left\lceil \left(\frac{256(L_\mu + M_\mu^2)}{e^2} + \frac{16(R + M_\mu + \sqrt{L_\mu})}{e} \right) \log \frac{2}{\delta_{\text{pair}}} \right\rceil. \tag{I.3}
$$

If the horizon is reached during any statistical batch, the initialization procedure stops. An incomplete batch is never used to cut the ellipsoid or eliminate an arm. On the high-probability event analyzed below, a valid partition is always certified before the prescribed cut budget is exhausted.

For $h = x_a - x_{r_k}$, the retained halfspace is

Pairwise test result	Retained halfspace
a selected	$-h^\top \theta \leqslant b$
r_k selected	$h^\top \theta \leqslant b$
unresolved and $h^\top c \geqslant 0$	$h^\top \theta \leqslant 2b$
unresolved and $h^\top c < 0$	$-h^\top \theta \leqslant 2b$.

(I.4)

The bin width b is chosen smaller than the within-bin score-width threshold w to account for uncertainty in the predicted scores. The margin $L_\mu w + e$ in the final filtering step accounts for both within-bin score error and estimation error in the signed mean. Without these margins, filtering bins only through their representatives need not preserve an optimal arm.

I.3 Guarantee for geometric initialization

Corollary I.2 (Regret with geometric initial localization)**.** *Suppose Assumptions 1–4 hold, with Assumption 4 imposed throughout geometric initialization. Set $\rho = 1/(64R\sqrt{d})$. Then, with probability at least $1 - \delta$, Algorithm I.2, followed by the localized elimination phase of Algorithm 1, satisfies*

$$
\textbf{Regret}(T) = \widetilde{\mathcal{O}}\left(\kappa_{\text{init}}^4 d^{9/2} + \sqrt{dT \min\{d, \log K\}} + d \min\{d, \log K\} \right).
$$

Algorithm I.3 One stage of geometric localization

Require: Active set $\mathcal{A}$, width W, failure probability δ_{stage}
Set the parameters in Equation I.2 and $\delta_{\mathrm{pair}} \leftarrow \delta_{\mathrm{stage}}/(N_{\mathrm{cut}}(W) + 2J_{\max})$.
Initialize $c \leftarrow 0$ and $H \leftarrow 4S^2 I$.
If the horizon is reached during any statistical batch, stop without using the incomplete batch.
for $q = 0, \dots, N_{\mathrm{cut}}(W) - 1$ **do**
  if $\max_{a\in\mathcal{A}} x_a^\top c - \min_{a\in\mathcal{A}} x_a^\top c > 2W$ **then**
    Let h be the maximizing feature minus the minimizing feature and cut by $h^\top(\theta - c) \leqslant 0$.
  else
    Partition $\{x_a^\top c : a \in \mathcal{A}\}$ into bins $\mathcal{D}_1, \dots, \mathcal{D}_J$ of width b and choose $r_k \in \mathcal{D}_k$.
    if $\|x_a - x_{r_k}\|_H \leqslant 16db$ for every k and $a \in \mathcal{D}_k$ **then**
      Certify the partition and **break**.
    end if
    Choose a violating pair (a, r_k) and set $h \leftarrow x_a - x_{r_k}$.
    Run its pairwise test with tolerance b and select the cut in Equation I.4.
  end if
  Update (c, H) by Lemma I.3.
end for
if no partition was certified **then**
  Play any fixed arm in $\mathcal{A}$ for the remaining rounds and **stop**.
end if
if $J = 1$ **then**
  return $\mathcal{A}$
end if
Tournament: Run a balanced tournament on $r_1, \dots, r_J$ with tolerance $W/(64\ell)$ and failure probability δ_{pair} per test; let $\widehat{a}$ be the winner.
Let $\widehat{a} = r_{\widehat{k}}$ and set $\mathcal{A}' \leftarrow \mathcal{D}_{\widehat{k}}$.
for $k \in [J]\backslash\{\widehat{k}\}$ **do**
  Filtering: Compute $\overline{C}_k$ from $N_{\mathrm{filter}}(e, \delta_{\mathrm{pair}})$ fresh comparisons of $(\widehat{a}, r_k)$.
  if $\overline{C}_k \leqslant L_\mu w + e$ **then**
    Set $\mathcal{A}' \leftarrow \mathcal{A}' \cup \mathcal{D}_k$.
  end if
end for
return $\mathcal{A}'$.

The initialization term contains no dependence on K, including hidden logarithmic factors.

The geometric initializer replaces the arm-by-arm comparisons of the original procedure by a *polynomial number of informative comparisons* that certify small score widths within groups of arms. Its initialization regret bound is therefore independent of K. Computation still depends on the given arm list; for example, one geometric iteration can be implemented by scanning predicted scores and arm–representative distances in $\mathcal{O}(|\mathcal{A}|d^2)$ arithmetic operations.

The guarantee follows from the next two lemmas and proposition.

Lemma I.1 (Certification of within-bin score widths)**.** *Suppose $a^\star \in \mathcal{A}$, $W(\mathcal{A}) \leqslant W \leqslant 2S$, and $W > 0$. Use the parameters in Equation I.2. If every queried pairwise test satisfies Lemma C.1, then every ellipsoid produced by Algorithm I.3 contains $B(\theta^\star, b/2)$, and a partition is certified after fewer than $N_{\mathrm{cut}}(W)$ cuts. There are at most $J_{\max}$ nonempty bins, and every certified bin $\mathcal{D}_k$ satisfies*

$$W(\mathcal{D}_k) \leqslant w.$$

Lemma I.2 (One-stage localization)**.** *Suppose $a^\star \in \mathcal{A}$ and $W(\mathcal{A}) \leqslant W \leqslant 2S$, with $W > 0$. With probability at least $1 - \delta_{\mathrm{stage}}$, a completed call to Algorithm I.3 returns a set $\mathcal{A}'$ satisfying $a^\star \in \mathcal{A}'$ and $W(\mathcal{A}') \leqslant \frac{5W}{64} < \frac{W}{4}$. Moreover, the stage uses at most*

$$\widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^2 d^4}{W^2}\right)$$

observations and incurs regret at most

$$\widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^3 d^4}{W}\right).$$

The same bounds apply to a stage truncated by the horizon.

Proposition I.1 (Total cost of geometric initial localization)**.** *Under the assumptions of Corollary I.2, with probability at least $1-\delta/3$, Algorithm I.2, whenever completed before the horizon, returns $\mathcal{A}_1$ satisfying Equation 2. Its number of observations is at most*

$$\widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^4 L_\mu^2 d^4}{\rho^2}\right).$$

Let τ_{init} denote the first round after initialization, with $\tau_{\text{init}} = T+1$ if initialization is not completed. On the same event,

$$\textbf{Regret}(\tau_{\text{init}}-1) = \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^4 L_\mu^3 d^4}{\rho}\right).$$

The hidden logarithmic factors depend on $S, M_\mu, R, d, \kappa_{\text{init}}, L_\mu, \rho^{-1}, \delta^{-1}$, but not on K.

Proof of Corollary I.2. By Proposition I.1, with probability at least $1-\delta/3$, we have

$$\begin{aligned}\textbf{Regret}(\tau_{\text{init}}-1) &= \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^4 L_\mu^3 d^4}{\rho}\right)\\ &= \widetilde{\mathcal{O}}\Big(R(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^4 L_\mu^3 d^{9/2}\Big). \qquad (\rho^{-1} = 64R\sqrt{d})\end{aligned}$$

If initialization is not completed by round T, this already bounds the total regret.

Otherwise, on the same event, $a^\star \in \mathcal{A}_1$ and $W(\mathcal{A}_1) \leqslant \rho/4 = \rho_1$. These are exactly the conditions required by Lemma E.2. Since the subsequent localized elimination phase is unchanged, Lemma E.2 bounds its conditional expected regret with additional failure probability at most $\delta/3$, and Lemma E.3 converts this to realized pseudo-regret with another failure probability at most $\delta/3$. Thus, on the intersection of these events, we obtain

$$\begin{aligned}&\sum_{t=\tau_{\text{init}}}^{T}\left[\mu(\nu_t + x_{a^\star}^\top\theta^\star) - \mu(\nu_t + x_{A_t}^\top\theta^\star)\right]\\ &= \widetilde{\mathcal{O}}\Big(\sqrt{L_\mu + M_\mu^2}\sqrt{dT\min\{d,\log K\}} + (R + M_\mu + \sqrt{L_\mu})d\min\{d,\log K\} + 1\Big).\end{aligned}$$

A union bound gives probability at least $1-\delta/3-\delta/3-\delta/3 = 1-\delta$. Adding the initialization and post-initialization regret proves the corollary. □

I.4 Proofs of the geometric initialization guarantees

We first prove the ellipsoid update used by the geometric procedure, followed by the within-bin certification, the one-stage localization guarantee, and the total initialization cost.

Lemma I.3 (Ellipsoid update for a shallow cut)**.** *Let $\mathcal{E}(c,H)$ be an ellipsoid, let $h \neq 0$, and set $s := \sqrt{h^\top H h}$. For $0 \leqslant \alpha \leqslant 1/(8d)$, there exists an ellipsoid containing*

$$\mathcal{E}(c,H) \cap \left\{\theta : h^\top(\theta - c) \leqslant \alpha s\right\}$$

whose volume is at most $\exp(-1/(8d))\,\mathrm{vol}(\mathcal{E}(c,H))$. For $d \geqslant 2$, one such update is

$$\begin{aligned}c' &= c - \frac{1-d\alpha}{d+1}\frac{Hh}{s},\\ H' &= \frac{d^2(1-\alpha^2)}{d^2-1}\left(H - \frac{2(1-d\alpha)}{(d+1)(1-\alpha)}\frac{Hhh^\top H}{s^2}\right).\end{aligned} \tag{I.5}$$

For $d = 1$, the retained interval itself may be used.

Proof of Lemma I.3. An invertible affine transformation followed by a rotation reduces the claim to the unit ball cut by $u_1 \leqslant \alpha$. For $d \geqslant 2$, define

$$\beta = \frac{1-d\alpha}{d+1}, \qquad p = \frac{d(1+\alpha)}{d+1}, \qquad q = \frac{d\sqrt{1-\alpha^2}}{\sqrt{d^2-1}}.$$

In these coordinates, the ellipsoid in Equation I.5 has center $-\beta e_1$, semiaxis p in the first coordinate, and semiaxis q in every orthogonal direction.

If $\|u\|_2 \leqslant 1$ and $u_1 = t \leqslant \alpha$, then $t \in [-1, \alpha]$ and its squared norm in the new ellipsoid is

$$F(t) := \frac{(t+\beta)^2}{p^2} + \frac{1-t^2}{q^2}.$$

Since $\alpha \leqslant 1/(8d) \leqslant 1/d$, we have $q^2 \geqslant p^2$. Hence, F is convex, while direct substitution gives $F(-1) = F(\alpha) = 1$. Therefore, $F(t) \leqslant 1$ throughout $[-1, \alpha]$, proving containment.

The logarithm of the volume ratio is

$$f(\alpha) = \log \frac{d(1+\alpha)}{d+1} + \frac{d-1}{2} \log \frac{d^2(1-\alpha^2)}{d^2-1}.$$

Using $\log(1-x) \leqslant -x$ and $\log(1+x) \leqslant x$,

$$\begin{aligned} f(0) &= \log\left(1 - \frac{1}{d+1}\right) + \frac{d-1}{2} \log\left(1 + \frac{1}{d^2-1}\right) \\ &\leqslant -\frac{1}{d+1} + \frac{d-1}{2(d^2-1)} \\ &= -\frac{1}{2(d+1)}. \end{aligned}$$

Moreover, we have

$$f'(\alpha) = \frac{1-d\alpha}{1-\alpha^2} \leqslant 1,$$

where the last inequality follows from $\alpha^2 \leqslant d\alpha$. Thus, we get

$$f(\alpha) \leqslant -\frac{1}{2(d+1)} + \frac{1}{8d} \leqslant -\frac{1}{8d}.$$

Affine transformations preserve volume ratios, proving the claim for $d \geqslant 2$. For $d = 1$, the retained interval has relative length $(1+\alpha)/2 \leqslant 9/16 < \exp(-1/8)$. □

Proof of Lemma I.1. We work on the event that every queried pairwise test satisfies Lemma C.1.

1. Preservation of a neighborhood of $\theta^\star$. Initially, $\mathcal{E}(0, 4S^2 I) = B(0, 2S)$. Since $\|\theta^\star\|_2 \leqslant S$ and $b/2 \leqslant S$,

$$B(\theta^\star, b/2) \subseteq B(0, 2S).$$

First consider a cut made because the predicted active score width exceeds $2W$. Let h be the maximizing feature minus the minimizing feature. Then $h^\top c > 2W$, $\|h\|_2 \leqslant 2$, and $h^\top \theta^\star \leqslant W(\mathcal{A}) \leqslant W$. Thus, every $\theta \in B(\theta^\star, b/2)$ satisfies

$$\begin{aligned} h^\top \theta &\leqslant h^\top \theta^\star + \|h\|_2 \|\theta - \theta^\star\|_2 \\ &\leqslant W + b < 2W < h^\top c. \end{aligned}$$

Hence the central cut $h^\top(\theta - c) \leqslant 0$ preserves the entire ball.

Now consider a statistical cut from a pair (a, r_k) in one bin and write $h = x_a - x_{r_k}$. Since the two predicted scores lie in an interval of width b, $|h^\top c| \leqslant b$. If the pairwise test selects a, then $h^\top \theta^\star > 0$, so every point in $B(\theta^\star, b/2)$ satisfies $h^\top \theta \geqslant -b$. If the test selects r_k, the symmetric argument gives $h^\top \theta \leqslant b$. If the comparison is unresolved, then $|h^\top \theta^\star| \leqslant b$, and every point in the ball satisfies $|h^\top \theta| \leqslant 2b$. Thus every halfspace in Equation I.4 preserves $B(\theta^\star, b/2)$.

Write the retained inequality as $\widetilde{h}^\top \theta \leqslant z$, where $\widetilde{h} \in \{h, -h\}$. In every case, we have

$$0 \leqslant z - \widetilde{h}^\top c \leqslant 2b.$$

The pair is queried only when $\|h\|_H > 16db$, so

$$0 \leqslant \frac{z - \widetilde{h}^\top c}{\|\widetilde{h}\|_H} < \frac{1}{8d}.$$

Lemma I.3 therefore applies to every statistical cut, as well as to every central cut. Its containment property proves the ball invariant inductively.

2. Termination of the geometric step. Each cut reduces the ellipsoid volume by a factor of at most $\exp(-1/(8d))$. After n cuts, the ball invariant gives

$$1 \leqslant \frac{\mathrm{vol}(\mathcal{E}_n)}{\mathrm{vol}(B(\theta^\star, b/2))} \leqslant \left(\frac{4S}{b}\right)^d \exp\left(-\frac{n}{8d}\right).$$

Consequently, $n \leqslant 8d^2 \log(4S/b) < N_{\mathrm{cut}}(W)$. Whenever the current partition is not certified, the algorithm can make another valid cut. Hence the cut budget cannot be reached without first certifying a partition.

3. True widths of certified bins. A partition is formed only when the predicted active score width is at most $2W$. It therefore contains at most $\lceil 2W/b \rceil + 2 = J_{\max}$ nonempty bins. At certification, for every $a, a' \in \mathcal{D}_k$, we have

$$\|x_a - x_{a'}\|_H \leqslant \|x_a - x_{r_k}\|_H + \|x_{a'} - x_{r_k}\|_H \leqslant 32db.$$

Since $\theta^\star \in \mathcal{E}(c, H)$, Cauchy–Schwarz in the ellipsoid norm gives

$$\begin{aligned} |(x_a - x_{a'})^\top \theta^\star| &\leqslant |(x_a - x_{a'})^\top c| + |(x_a - x_{a'})^\top (\theta^\star - c)| \\ &\leqslant b + \|x_a - x_{a'}\|_H \\ &\leqslant (32d + 1)b \\ &\leqslant 33db = w. \end{aligned}$$

Taking the maximum over pairs proves $W(\mathcal{D}_k) \leqslant w$. □

Proof of Lemma I.2. 1. Simultaneous validity of statistical batches. For any pair (i, j) fixed within a batch, the same calculation as in Lemma C.1 gives

$$m_s := \mathbb{E}[C_s \mid \mathcal{G}_s] = \alpha_{s,ij}(x_i - x_j)^\top \theta^\star, \qquad \frac{1}{\kappa_{\mathrm{init}}} \leqslant \alpha_{s,ij} \leqslant L_\mu. \tag{I.6}$$

For a filtering batch of length N, let $\overline{m} := N^{-1} \sum_{s=1}^N m_s$. The common shift may vary adaptively within the batch, so $\overline{m}$ is the average predictable signed mean rather than a fixed mean. By Lemma F.5 with $H = 2, V = 4$ and Equation I.3,

$$|\overline{C} - \overline{m}| \leqslant e$$

with failure probability at most δ_{pair}. The pairwise tests satisfy Lemma C.1 with the same failure probability. Both statements hold conditional on the history before each adaptively selected batch.

The geometric step uses at most $N_{\mathrm{cut}}(W)$ pairwise tests, the representative tournament uses at most $J_{\max} - 1$ tests, and filtering uses at most $J_{\max} - 1$ batches. Thus, a union bound gives simultaneous validity with probability at least

$$1 - \big(N_{\mathrm{cut}}(W) + 2J_{\max}\big)\delta_{\mathrm{pair}} = 1 - \delta_{\mathrm{stage}}.$$

In particular, no union bound over all K arms or all possible pairs is required. We work on this event below.

2. Preservation of an optimal arm. Lemma I.1 gives a certified partition with $W(\mathcal{D}_k) \leqslant w$ for every bin. If $J = 1$, the algorithm returns the input set, whose width is at most $w < W/4$. Assume $J \geqslant 2$.

By Lemma C.2, the representative tournament gives

$$\max_{k \in [J]} x_{r_k}^\top \theta^\star - x_{\widehat{a}}^\top \theta^\star \leqslant \frac{W}{64}.$$

Let $\mathcal{D}_{k^\star}$ be the bin containing $a^\star$. Its representative lies within score distance w of $a^\star$. Since $\widehat{a}$ is itself an arm, $x_{\widehat{a}}^\top \theta^\star \leqslant x_{a^\star}^\top \theta^\star$. Hence,

$$(x_{\widehat{a}} - x_{r_{k^\star}})^\top \theta^\star \leqslant w.$$

If $\mathcal{D}_{k^\star}$ is the champion's bin, it is retained by construction. Otherwise, Equation I.6 implies that the average predictable signed mean in its filtering batch is at most $L_\mu w$. The same conclusion holds

when the score difference is negative. Therefore, $\overline{C}_{k^\star} \leqslant L_\mu w + e$, so the optimal bin is retained. Thus, $a^\star \in \mathcal{A}'$.

3. Width of the surviving set. For every retained non-champion bin,

$$\overline{m}_k \leqslant \overline{C}_k + e \leqslant L_\mu w + 2e.$$

If $(x_{\hat{a}} - x_{r_k})^\top \theta^\star > 0$, the lower derivative bound in Equation I.6 gives

$$(x_{\hat{a}} - x_{r_k})^\top \theta^\star \leqslant \kappa_{\text{init}}(L_\mu w + 2e).$$

The same upper bound is automatic when the gap is nonpositive and for the champion representative itself.

The best representative is within w of $a^\star$, while the tournament winner is within $W/64$ of the best representative. Hence,

$$(x_{a^\star} - x_{\hat{a}})^\top \theta^\star \leqslant w + \frac{W}{64}.$$

For any retained arm $a \in \mathcal{D}_k$, $(x_{r_k} - x_a)^\top \theta^\star \leqslant w$. Combining the three bounds gives

$$\begin{aligned}(x_{a^\star} - x_a)^\top \theta^\star &\leqslant \frac{W}{64} + (\kappa_{\text{init}} L_\mu + 2)w + 2\kappa_{\text{init}} e \\ &= \frac{W}{64}\left(2 + \frac{\kappa_{\text{init}} L_\mu + 2}{\max\{1, \kappa_{\text{init}} L_\mu\}}\right) \\ &\leqslant \frac{5W}{64}.\end{aligned}$$

Since $a^\star$ is retained, this is exactly an upper bound on $W(\mathcal{A}')$.

4. Number of observations and stage regret. A deterministic upper bound on the observation budget of one stage is

$$\begin{aligned}N_{\text{stage}}(W, \delta_{\text{stage}}) := N_{\text{cut}}(W) N_{\text{pair}}(b, \delta_{\text{pair}}) \\ + (J_{\max} - 1) N_{\text{pair}}\left(\frac{W}{64\ell}, \delta_{\text{pair}}\right) \\ + (J_{\max} - 1) N_{\text{filter}}(e, \delta_{\text{pair}}),\end{aligned}$$

where $\delta_{\text{pair}} = \delta_{\text{stage}}/(N_{\text{cut}}(W) + 2J_{\max})$.

If the stage collects no observations, both claims are trivial. Otherwise, fix any initialization round t. By Assumption 1, $|\mu(z)| \leqslant M_\mu$ over the feasible natural-parameter range. Integrating the initialization curvature condition over $[-S, S]$ gives

$$\frac{2S}{\kappa_{\text{init}}} \leqslant \int_{-S}^{S} \mu'(\nu_t + u)\, du = \mu(\nu_t + S) - \mu(\nu_t - S) \leqslant 2M_\mu.$$

Thus, $W \leqslant 2S \leqslant 2M_\mu \kappa_{\text{init}}$. Moreover, $1/\kappa_{\text{init}} \leqslant \mu' \leqslant L_\mu$ implies $\kappa_{\text{init}} L_\mu \geqslant 1$.

Consequently,

$$b = \frac{W}{2112 d \kappa_{\text{init}} L_\mu}, \qquad J_{\max} = \mathcal{O}(d \kappa_{\text{init}} L_\mu),$$

while $N_{\text{cut}}(W) = \mathcal{O}(d^2 \log(4S/b))$ and $\ell = \mathcal{O}(\log J_{\max})$. As in the successive-tournament proof, for every pairwise tolerance at most W, the linear Bernstein term and ceiling are absorbed into $\mathcal{O}((L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2/\varepsilon^2)$. For filtering, $e = W/(128\kappa_{\text{init}}) \leqslant M_\mu/64$, so Equation I.3 likewise gives $N_{\text{filter}}(e, \delta_{\text{pair}}) = \mathcal{O}((L_\mu + M_\mu^2 + RM_\mu)e^{-2}\log(2/\delta_{\text{pair}}))$. Using the pairwise-test budget,

$$\begin{aligned}N_{\text{cut}}(W) N_{\text{pair}}(b, \delta_{\text{pair}}) &= \widetilde{\mathcal{O}}\left(d^2 \frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^2}{b^2}\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\text{init}}^4 L_\mu^2 d^4}{W^2}\right).\end{aligned}$$

The representative tournament and filtering together require $\widetilde{\mathcal{O}}\Big(\frac{(L_\mu+M_\mu^2+RM_\mu)\kappa_{\mathrm{init}}^2 J_{\max}}{W^2}\Big)$ observations, which is no larger than the preceding term up to logarithmic factors. The ceilings and linear Bernstein terms are absorbed by the preceding mean-scale bounds. Therefore, we have

$$N_{\mathrm{stage}}(W,\delta_{\mathrm{stage}}) = \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^2 d^4}{W^2}\right).$$

Finally, every sampled arm belongs to the input set $\mathcal{A}$. Since $a^\star \in \mathcal{A}$ and $W(\mathcal{A}) \leqslant W$, the mean-value theorem implies that every sampled round incurs regret at most $L_\mu W$:

$$0 \leqslant \mu(\nu_t + x_{a^\star}^\top \theta^\star) - \mu(\nu_t + x_{A_t}^\top \theta^\star) \leqslant L_\mu W.$$

Multiplying by the observation budget gives

$$\widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^3 d^4}{W}\right)$$

regret for the stage. The same bounds apply if the stage is truncated by the horizon. □

Proof of Proposition I.1. If $2S \leqslant \rho/4$ or there is only one arm, initialization is empty and its output already satisfies Equation 2. Otherwise, let

$$P = \left\lceil \log_4 \frac{8S}{\rho} \right\rceil, \qquad W_j = 2S\,4^{-j}, \quad j = 0,\dots,P-1.$$

Assign failure probability $\delta/(3P)$ to each stage. The conditional guarantee of Lemma I.2 and a union bound imply that every completed stage is valid with probability at least $1-\delta/3$. We work on this event.

Starting from $a^\star \in \mathcal{B}_0$ and $W(\mathcal{B}_0) \leqslant 2S = W_0$, induction gives

$$a^\star \in \mathcal{B}_j, \qquad W(\mathcal{B}_j) \leqslant W_j$$

at the beginning of every stage reached by the algorithm. If all P stages are completed, then $W_P \leqslant \rho/4$. If initialization stops earlier because only one arm remains, that arm is optimal and the remaining width is zero. Thus, whenever initialization finishes, its output satisfies Equation 2.

If initialization contains no observations, both cost claims are trivial. Otherwise, Lemma I.2 implies that stage j uses at most $\widetilde{\mathcal{O}}\Big(\frac{(L_\mu+M_\mu^2+RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^2 d^4}{W_j^2}\Big)$ observations and incurs regret at most $\widetilde{\mathcal{O}}\Big(\frac{(L_\mu+M_\mu^2+RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^3 d^4}{W_j}\Big)$. The same bounds apply to a final stage truncated by the horizon.

By the definition of P, the final prescribed width satisfies $\rho/4 < W_{P-1} \leqslant \rho$. Since $W_j = 2S\,4^{-j}$,

$$\sum_{j=0}^{P-1} W_j^{-2} = W_{P-1}^{-2} \sum_{k=0}^{P-1} 16^{-k} < \frac{256}{15\rho^2},$$
$$\sum_{j=0}^{P-1} W_j^{-1} = W_{P-1}^{-1} \sum_{k=0}^{P-1} 4^{-k} < \frac{16}{3\rho}.$$

Therefore, summing over the stages gives at most $\widetilde{\mathcal{O}}\Big(\frac{(L_\mu+M_\mu^2+RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^2 d^4}{\rho^2}\Big)$ initialization observations and

$$\mathbf{Regret}(\tau_{\mathrm{init}}-1) = \widetilde{\mathcal{O}}\left(\frac{(L_\mu + M_\mu^2 + RM_\mu)\kappa_{\mathrm{init}}^4 L_\mu^3 d^4}{\rho}\right).$$

Finally, P, $J_{\max}$, $N_{\mathrm{cut}}(W_j)$, and all confidence allocations in the geometric procedure are independent of K. Their logarithmic factors can be bounded using the final prescribed width and $P = \mathcal{O}(1+\log(S/\rho))$. This proves the claimed absence of K from the initialization bound, including its hidden logarithmic factors. □

J Experiment Details and Additional Results

In this section, we provide additional details on the experimental setup in Section 6 and report further experiments examining the effect of curvature.

J.1 Experimental setup

We consider a fixed set of $K = 20$ arms with features $x_a \in \mathbb{R}^d$, where $d = 5$ and $\|x_a\|_2 \leqslant 1$. Unless stated otherwise, we set $\|\theta^\star\|_2 = S = 3$, $B = 5$, and $T = 30{,}000$. At each round, the common shift $\nu_t \in [-B, B]$ is determined before the learner samples an arm A_t, after which it observes a Bernoulli reward with mean $\mu(x_{A_t}^\top \theta^\star + \nu_t)$. We evaluate both the logistic link $\mu(z) = (1 + e^{-z})^{-1}$ and the probit link $\mu(z) = \Phi(z)$, with the latter included as a robustness test beyond our canonical-family assumptions. All reported curves are averaged over 20 independent seeds.

Common-shift environments. We evaluate the algorithms under three different nuisance processes.

1. Fixed-period Beta environment. We generate a nuisance trajectory that changes smoothly over time and repeats every $P = 5{,}000$ rounds. Specifically, we first construct a triangular wave that moves from 0 to 1 and back to 0 within each period, and then map this wave through the quantile function of a $\mathrm{Beta}(2, 10)$ distribution. Let $F_{2,10}^{-1}$ denote this quantile function. We draw $\phi \sim \mathrm{Unif}(0, 1)$ to randomly choose the starting point within the cycle and define

$$q_t = 1 - \left| 2 \left\{ \frac{t}{P} + \phi \right\} - 1 \right|, \qquad \nu_t = B \left(2F_{2,10}^{-1}(q_t) - 1 \right),$$

where $\{\cdot\}$ denotes the fractional part. Because the random starting point makes q_t marginally uniform on $[0, 1]$, ν_t has the same marginal distribution as $B(2U - 1)$ for $U \sim \mathrm{Beta}(2, 10)$. At the same time, the values of ν_t are strongly correlated over time because the same trajectory repeats every $5{,}000$ rounds. This nuisance process is generated independently of the learner.

2. Variable-period Beta environment. We use the same Beta-quantile construction, but allow the duration of each cycle to change. Each new cycle length is independently sampled uniformly from $\{4{,}000, 4{,}001, \ldots, 6{,}000\}$. Thus, ν_t follows the same type of smooth Beta-shaped trajectory, but the timing of its cycles is irregular rather than exactly periodic. We initialize the process by sampling the first cycle in proportion to its length and choosing a uniformly random starting point within that cycle. This ensures that the marginal distribution of ν_t remains the same transformed $\mathrm{Beta}(2, 10)$ distribution as in the fixed-period environment. Hence, the two environments have the same marginal nuisance distribution but different temporal dependence.

3. Adversarial environment. We finally consider a policy-dependent nuisance process. At round t, the learner first announces its action distribution $\pi_t \in \Delta(K)$. After observing π_t, $\theta^\star$, and the arm features, but before the current action and reward are sampled, the adversary chooses a common shift $\nu_t \in [-B, B]$ such that

$$\sum_{a=1}^{K} \pi_t(a)\, \mu(x_a^\top \theta^\star + \nu_t) = \frac{1}{2}. \tag{J.1}$$

In other words, the adversary keeps the learner's marginal reward mean at $1/2$ regardless of how its action distribution changes. This makes the environment particularly challenging: assigning more probability to better arms does not increase the aggregate reward level, and the marginal Bernoulli variance remains at its maximum value of $1/4$. At the same time, the conditional reward means of individual arms can still differ, so the problem retains informative action-dependent structure. Unlike the two Beta environments, the nuisance here therefore adapts directly to the learner's current policy.

A feasible shift always exists in our experiments. Since $\|x_a\|_2 \leqslant 1$ and $\|\theta^\star\|_2 = S \leqslant B$, we have $x_a^\top \theta^\star \in [-S, S]$. Hence, at $\nu_t = -B$ all natural parameters are nonpositive, whereas at $\nu_t = B$ they are nonnegative. Because both the logistic and probit links are continuous, increasing, and satisfy $\mu(0) = 1/2$, the left-hand side of Equation J.1 crosses $1/2$ for some $\nu_t \in [-B, B]$. The same adversarial rule is applied to every algorithm.

Hyperparameter tuning. For a controlled comparison, we tune a single multiplicative exploration coefficient

$$c \in \{0.05, 0.1, 0.5, 1\}$$

separately for each algorithm and environment. The coefficient scales the algorithm-specific quantity that directly controls exploration, such as a confidence radius, Thompson-sampling scale, or phase sample budget. Table J.1 summarizes how c is applied. All other algorithm-specific parameters and regularization constants are kept fixed. For each algorithm–environment pair, we report the value of c

Algorithm	**Quantity scaled by** c
`GLM-UCB` (Filippi et al., 2010)	Exploration coefficient in the UCB bonus: $\rho_t = c\,\rho_{t,0}$.
`GLOC` (Jun et al., 2017)	Confidence-set radius: $r_t = c\,r_{t,0}$; equivalently, its squared radius is multiplied by c^2.
`RS-GLinCB` (Sawarni et al., 2024)	The switching/elimination confidence coefficient and the main-stage UCB coefficient are both multiplied by c.
`GLB-OMD` (Zhang et al., 2025)	Confidence radius: $\beta_t^{(c)} = c\,\beta_t$.
`OFU-ECOLog` (Faury et al., 2022)	The confidence radii used in the warm-up and subsequent optimistic stages are multiplied by c; corresponding squared-radius quantities are therefore multiplied by c^2.
`OFULog+` (Lee et al., 2024)	Likelihood confidence-set radius: $L_t(\theta) - L_t(\widehat{\theta}_t) \leqslant c^2\beta_{t,0}^2$.
`BOSE` (Krishnamurthy et al., 2018)	Confidence multiplier in arm elimination: $\gamma = c\,\gamma_0$.
`Semiparametric TS` (Kim & Paik, 2019)	Thompson-sampling scale: $v = c\,v_0$, so that the sampling covariance becomes $c^2 v_0^2 B_t^{-1}$.
`SBE` (Kim et al., 2025)	Phase sample budget: $n_\ell = \lceil c\,n_{\ell,0} \rceil$.
`LE-SGLB` (Ours)	Confidence radius: $\beta_m^{(c)}(n) = c\,\beta_m(n)$

Table J.1: Exploration coefficient used for hyperparameter tuning.

yielding the lowest final mean cumulative regret.

J.2 Additional results: varying curvature

We further examine the effect of curvature by varying the parameter norm. While the main experiments use $\|\theta^\star\|_2 = S = 3$, here we consider $S \in \{3, 4, 5\}$ and keep $B = 5$, $d = 5$, $K = 20$, $T = 30{,}000$, and all other settings unchanged.

For the logistic link, the resulting global inverse-curvature values over the feasible natural-parameter range are approximately 2.98×10^3, 8.11×10^3, and 2.20×10^4 for $S = 3, 4, 5$, respectively. For the probit link, the corresponding values are approximately 1.98×10^{14}, 9.73×10^{17}, and 1.30×10^{22}. Thus, increasing S moves the problem into progressively lower-curvature regimes.

Ordinary GLBs (Abeille et al., 2021; Faury et al., 2022; Zhang et al., 2025) can exploit favorable local curvature to improve the leading regret term by a factor of $1/\sqrt{\kappa^\star}$. In contrast, Theorem 2 shows that this benefit need not carry over to SGLBs.

Figure J.1 shows that `LE-SGLB` remains stable as S increases from 3 to 5, despite the corresponding growth in both the global and local inverse-curvature parameters. In ordinary logistic or generalized linear bandits, larger $\kappa^\star$ can improve the leading regret through favorable local curvature (Abeille et al., 2021; Zhang et al., 2025). Here, however, this improvement is not evident for the baseline methods, and several of them incur even larger regret as both S and the corresponding $\kappa^\star$ increase. This suggests that the unobserved common shifts prevent standard GLB methods from exploiting the same curvature advantage. The observation is consistent with Theorem 2, which shows that a larger $\kappa^\star$ need not make the semiparametric problem statistically easier.

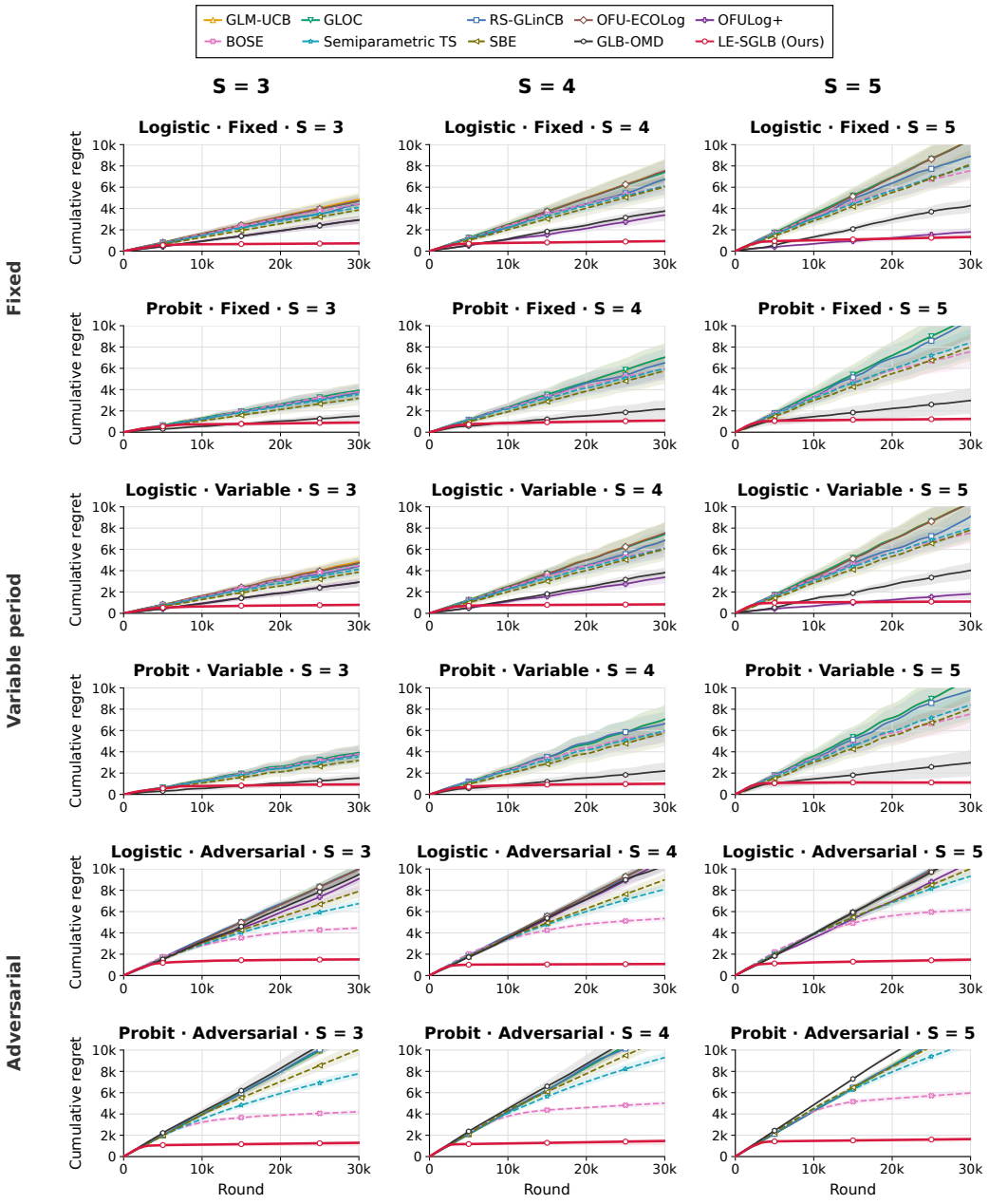

Figure J.1: Cumulative regret for logistic and probit SGLBs with $S \in \{3, 4, 5\}$, with all other experimental settings held fixed. Columns correspond to different values of S, while rows show the fixed-period, variable-period, and adversarial environments for the two link functions. Curves show means over 20 independent seeds, and shaded regions indicate one standard error.