OPTIMAL REGRET FOR SINGLE INDEX BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the *single-index bandit* problem, where rewards depend on an unknown one-dimensional projection of high-dimensional contexts through an unknown reward function. This model extends linear and generalized linear bandits to a nonparametric setting, and is particularly relevant when the reward function is not known in advance. While optimal regret guarantees are known for monotone reward functions, the general non-monotone case remains poorly understood, with the best known bound being $\tilde{\mathcal{O}}(T^{3/4})$ (under standard boundedness, nonzero first-order Stein signal μ^*, and Lipschitz assumptions on the reward Kang et al. (2025)).

We close this gap by establishing the optimal regret for general single-index bandits. We propose a simple two-phase algorithm, namely, Zooming Single Index Bandit with Upper Confidence Bound (`ZoomSIB-UCB`), that first estimates the projection direction via a normalized Stein estimator. With an adaptive exploration and discretization, it then runs UCB over the resulting one-dimensional bins. This approach achieves a regret of $\tilde{\mathcal{O}}(T^{2/3})$, and improves significantly upon prior work without knowing μ^* or a lower bound on its magnitude. We also prove a matching minimax lower bound of $\tilde{\Omega}(T^{2/3})$, showing that the dependence on T is optimal up to logarithmic factors. Our upper and lower bounds together provide a sharp characterization of the regret in single-index bandits in terms of T. Empirical results further demonstrate the effectiveness and robustness of our approach.

1 INTRODUCTION

In a single index model (SIM) on data $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, we let y depend on the projection of x onto an unknown parameter $\theta^\star$ via an unknown function $f : \mathbb{R} \to \mathbb{R}$. Hence, learning in SIM is nonparametric since we need to estimate the parameter $\theta^\star$ as well as the unknown function f. As an extension of the Generalized Linear Model, the SIM is a versatile statistical framework that has been applied across longitudinal data analysis, quantile regression, and econometrics, and has more recently attracted substantial attention from the theoretical deep learning community as a tool for evaluating the ability of neural networks to learn low-dimensional representations, in contrast to kernel methods (Plan and Vershynin (2016), Brillinger (1982), Bruna and Hsu (2025)).

Learning with SIM remains an active area of research. Early foundational work by Ichimura (1993) and Härdle et al. (1993) establish semiparametric efficiency bounds and optimal estimation rates for the SIM in the classical low-dimensional regime. Hornik et al. (1989) and McCullagh and Nelder (1989) connect the SIM to neural networks and generalized linear models respectively. In the statistical learning theory literature, Plan and Vershynin (2016) and Goldstein et al. (2016) study SIMs through the lens of empirical process theory and Gaussian width. More recently, the connection between SIMs and gradient descent on two-layer neural networks has attracted significant attention (Barak et al. (2022), Abbe et al. (2022)).

Recently, there has been significant interest in understanding the single index model in the sequential framework. The multi-armed bandit is a canonical framework in modeling online (sequential) learning and is classically used for recommendation, clinical trials, targeted ads and hyperparameter tuning Lattimore and Szepesvári (2020), Bubeck and Cesa-Bianchi (2012). Traditionally, in order to incorporate contextual information, a variant of multi armed bandit framework, namely contextual linear bandit Chu et al. (2011), Abbasi-yadkori et al. (2011), Dani et al. (2008), Li et al. (2010) is studied where the reward is assumed to be a linear function of the context. However, in many applications like click through rate in recommendation systems the linear model fails to capture the

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.

reward. To address this, Filippi et al. (2010), Li et al. (2017), Zhang et al. (2025), Faury et al. (2020), Abeille et al. (2021) study the generalized linear bandit framework where the reward is assumed to depend on the context via a generalized linear model. Formally, the mean reward at time t is given by $f(x^T\theta^\star)$ where the reward function $f(.)$ is known in advance. Although such a model captures a lot of practical application, assumption about the knowledge of $f(\cdot)$ is often unrealistic and it has been shown in Ghosh et al. (2017), Bogunovic and Krause (2021) that misspecification can lead to linear regret. To address this issue, Kang et al. (2025) introduce single index bandit, where the reward function is not known in advance. When the reward function is monotone, the authors of Kang et al. (2025) show that exploiting the Stein estimator, an optimal regret scaling of $\mathcal{O}(\sqrt{T})$ is possible where T is the learning horizon. However, when the reward function is non-monotone, Kang et al. (2025) assumes bounded and Lipschitz structure on the reward function and use kernel based estimator to learn the (unknown) function. They use an explore then commit type algorithm, resulting in a regret of $\tilde{\mathcal{O}}(T^{3/4})$.

There has been a few recent works on single index learning in bandits. For example, in Arya et al. (2026), the authors assume that the reward function comes from a Reproducing Kernel Hilbert Space (RKHS) and exploiting the reproducing property as well as representer theorems, the authors show that in some cases $\mathcal{O}(\sqrt{T})$ is possible. Moreover, in Ma and Cai (2025) reward functions coming from the intersection of smoothness and monotone function class is studied. The paper also proves matching lower bounds for such smooth monotone functions.

Summarizing the above results, we see that for monotone single index bandits, the optimal regret is well understood in literature. However, for general non-monotone functions (with boundedness and Lipschitz structure), apart from the $\tilde{\mathcal{O}}(T^{3/4})$ regret in Kang et al. (2025), no other results are known to the best of our knowledge. This forms the central question of our work:

What is the optimal regret scaling for general single index bandit?

In this paper, we answer the above question. We keep the assumptions identical to that of Kang et al. (2025) (i.e., the reward function is bounded and Lipschitz, with nonzero first-order Stein signal) and obtain an improved regret bound of $\tilde{\mathcal{O}}(T^{2/3})$. Morevoer, we do not require the Gaussian design used in their general-link analysis. We leverage tools from the Lipschitz bandit literature Kleinberg (2004), Kleinberg et al. (2008), Agrawal (1995), Auer et al. (2007), Slivkins (2019) and propose an algorithm that uses adaptive discretization and an appropriately instantiated Upper Confidence Bound (UCB) algorithm. Moreover, we prove a minimax lower bound of $\tilde{\Omega}(T^{2/3})$ implying that no bandit algorithm can incur better regret in the single index setting. Combining these two results, we answer the aforementioned question that the correct regret scaling for (non-monotone) single index bandit is $\Theta(T^{2/3})$ (ignoring log factors). We now summarize our contributions.

1.1 Summary of contributions

Improved Regret: We propose a learning algorithm namely Zooming Single Index Bandit with Upper Confidence Bound (`ZoomSIB-UCB`) in Section 3. On a high level, the algorithm breaks the single index learning to a sequence of one-dimensional bandit problems. In the first phase of `ZoomSIB-UCB`, we use the Stein estimator Stein (1981) with proper normalization and leverage the Lipschitz property of the reward function to discretize the index space with appropriate resolution into N bins. In the second phase, we play UCB over N bins to choose the arm and observe the associated reward. It turns out that $N = \tilde{\mathcal{O}}(T^{1/3})$ bins are sufficient and as a result the regret of `ZoomSIB-UCB` is $\tilde{\mathcal{O}}(T^{2/3})$. This may be seen as a direct improvement over Kang et al. (2025).

No knowledge of μ^*: An explore-then-commit algorithm has to decide when to stop exploring. The right moment depends on $\mu^* = \mathbb{E}[f'(X^\top\theta^\star)]$, the average slope of the reward function f with feature vector X, which the learner does not know. GSTOR uses a prescribed exploration schedule Kang et al. (2025); we instead choose the exploration length from an observable error certificate. The idea is simple: the unnormalized Stein estimator $\hat{\theta}$ estimates $\mu^*\theta^\star$, so its norm $\|\hat{\theta}\|_1$ estimates μ^*, because the index vector is normalised so that $\|\theta^\star\|_1 = 1$. Dividing the estimation error by this norm gives a *certificate*, that is, a bound on the current error that the learner can compute. The algorithm compares an upper bound on exploration regret with a certified bound on discretization cost. Both bounds are

computable from the data and the supplied problem constants, so no separate choice of exploration length is required.

Matching Lower Bound: We prove a minimax lower bound showing that there exists a problem instance such that any learning algorithm suffers an expected regret of $\Omega(T^{2/3})$. Hence, our proposed algorithm is optimal in its dependence on T, when the number of arms (contexts) K scales as $K \asymp T^{1/3}\log T$ and Stein signal is bounded away from zero. We construct a one dimensional problem with bumps respecting the Lipschitz property. Even in this simple framework, we perturb the problem instance and using information theoretic framework, show a regret lower bound of $\tilde{\Omega}(T^{2/3})$. To the best of our knowledge, this is the first work to obtain a minimax lower bound for non-monotone single-index bandits.

Simulation: We demonstrate the empirical robustness of `ZoomSIB-UCB` through extensive simulations on complex synthetic geometries and high-dimensional real-world datasets (Section 6). On the fixed-signal synthetic instances, `ZoomSIB-UCB` achieves substantially lower regret than GSTOR Kang et al. (2025). The experiments also identify weak-signal settings where this finite-horizon advantage reverses.

1.2 Novelty and Technical Challenges

The improved regret bound and the matching lower bound require several new ideas. Our algorithm, `ZoomSIB-UCB`, combines discretization ideas from Lipschitz bandits Slivkins (2019); Kleinberg (2004) with sleeping bandits Kanade et al. (2009); Kleinberg et al. (2010). The first challenge is to decide how finely to discretize the projected index space. This depends on how accurately the unknown direction $\theta^\star$ has been estimated. The same issue also determines how long the algorithm should spend estimating $\theta^\star$. Their appropriate choices depend on estimation accuracy, which in turn depends on the unknown signal strength $\mu^\star$. We overcome this by using the estimator itself to obtain a data-dependent bound on the current estimation error. This bound determines both the bin width and when to stop the exploration phase, without requiring knowledge of $\mu^\star$ or a lower bound on it. A second challenge is that, at every round, only some of the bins contain available arms. Hence, the learner cannot treat all bins as available actions. We handle this using the sleeping-bandit framework Kleinberg et al. (2010), where the algorithm compares the available bins at each round. Finally, estimating $\theta^\star$ and learning the unknown reward function cannot be treated independently. The quality of the direction estimate determines the discretization used in the bandit phase. Moreover, because the exploration length is chosen from the data, its analysis requires an estimation guarantee that remains valid throughout the exploration phase. Combining the Stein estimator, the adaptive discretization rule, and sleeping-bandit UCB gives the final algorithm and regret bound.

For the lower bound, the main challenge is to construct hard instances that retain a nonzero Stein signal while remaining statistically difficult to distinguish. We use localized bumps together with a common smooth ramp and reduce the problem to a hard multi-armed bandit instance; a Bretagnolle–Huber argument then yields the $\Omega(T^{2/3})$ rate.

Notation. For integer r, $[r]$ denotes the set $\{1, \dots, r\}$. We write $a \asymp b$ if a and b are equal up to positive constant factors. The notation $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ hides polylogarithmic terms.

2 Setup and Preliminaries

We consider the sequential decision-making problem formulated as a Single-Index Bandit (SIB) Kang et al. (2025). Let T denote the time horizon. At each round $t \in [T]$, the learner observes an arm set $\mathcal{X}_t = \{x_{t,a} \in \mathbb{R}^d : a \in [K]\}$ consisting of K feature vectors. We assume these context vectors are drawn i.i.d. from a continuous distribution $\mathcal{D}$ with multivariate density $p(\cdot)$, and that each coordinate of $X \sim \mathcal{D}$ is κ-sub-Gaussian. The learner selects an arm $x_t \in \mathcal{X}_t$ and observes a stochastic reward:

$$y_t = f(x_t^\top \theta^\star) + \eta_t$$

where $\theta^\star \in \mathbb{R}^d$ is the unknown true parameter vector, and $f : \mathbb{R} \to \mathbb{R}$ is an unknown, continuously differentiable link function.

Assumption 1 (Noise). *The noise sequence $\{\eta_t\}_{t\geq 1}$ satisfies: η_t is independent of x_t (and of all arms $\mathcal{X}_t$ at round t), and is conditionally σ-sub-Gaussian: $\mathbb{E}[e^{\lambda\eta_t} \mid \mathcal{G}_t] \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$, where $\mathcal{G}_t := \sigma(\mathcal{X}_{1:t}, a_{1:t-1}, y_{1:t-1})$. The noise need not be i.i.d.*

Model Identifiability: One needs to address two issues regarding identifiability in the single index model- (a) scaling and (b) direction (sign). To ensure scaling, we adopt the standard normalization $||\theta^\star||_1 = 1$ (see Kang et al. (2025))[1]. Moreover, a quantity of interest for single index model is $\mu^* = \mathbb{E}[f'(X^T\theta^\star)]$ where $X \sim \mathcal{D}$. We assume only $\mu^* \neq 0$. The pairs $(\theta^\star, f)$ and $(-\theta^\star, f(-\cdot))$ give exactly the same rewards, and our algorithm never uses the sign of its estimate. So we take $\mu^* > 0$ without loss of generality. We assume no lower bound on μ^*, and the learner does not know μ^*.

Regret: Let $a_t^* := \arg\max_{a\in[K]} f(x_{t,a}^\top\theta^\star)$ denote the optimal arm at round t, and let $x_{t,*} := x_{t,a_t^*}$. The performance of the learner is measured by the regret over T rounds, defined as the cumulative difference between the expected reward of the optimal arm and the chosen arm:

$$R_T = \sum_{t=1}^{T} \left(f(x_{t,*}^\top\theta^\star) - f(x_t^\top\theta^\star)\right).$$

For non-monotone f, the optimal arm is not necessarily the one with the largest projection $x^\top\theta^\star$.

2.1 Stein's Identity and Parameter Estimation

Because the reward function $f(\cdot)$ is completely unknown, the usual generalized linear bandit estimators based on a specified link cannot be applied directly. To bypass this structural limitation, we leverage Stein's method Stein (1981).

Definition 2.1 (Score Function). *The score function $S^p : \mathbb{R}^d \to \mathbb{R}^d$ associated with the density $p(x)$ is defined as: $S^p(x) = -\nabla_x \log(p(x)) = -\frac{\nabla_x p(x)}{p(x)}$.*

For readability, we omit the superscript p and simply write $S(x)$ when the underlying distribution is clear from the context.

Stein's Identity: Under the integrability and boundary conditions in Assumption 3, the generalized Stein identity states that $\mathbb{E}[f(X)S(X)] = \mathbb{E}[\nabla f(X)]$ Stein (1981). Applying this to our specific reward model yields:

$$\mathbb{E}[y_t S(x_t)] = \mathbb{E}[f(x_t^\top\theta^\star)S(x_t)] = \mathbb{E}[f'(x_t^\top\theta^\star)]\theta^\star := \mu^*\theta^\star$$

This identity demonstrates that the expected value of the reward, when weighted by the score function, perfectly aligns with the direction of the true parameter $\theta^\star$ (scaled by a constant μ^*) Stein (1981); Kang et al. (2025). This allows the learner to construct an estimator of $\mu^*\theta^\star$, and hence of the direction of $\theta^\star$, entirely without requiring the explicit form of $f(\cdot)$.

The learner must know the density p to evaluate S, as in Kang et al. (2025) (Remark E.1, Appendix E).

2.2 The Truncated Stein Estimator

Given n exploration samples $\{(x_i, y_i)\}_{i=1}^n$ (rounds in which the arm is chosen uniformly at random), the naive empirical approach would be to compute the sample average $\frac{1}{n}\sum_{i=1}^n y_i S(x_i)$. However, as shown in Kang et al. (2025), a truncation is typically required to control the tails. Let $\varphi_\tau : \mathbb{R}^d \to \mathbb{R}^d$ denote an element-wise truncation function

$$[\varphi_\tau(v)]_j = \text{sign}(v_j)\min(|v_j|, \tau), \quad \forall j \in [d]$$

where $\tau > 0$ is a carefully chosen truncation threshold. The unnormalized truncated Stein estimator is then defined as

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \varphi_\tau(y_i S(x_i)) \tag{1}$$

[1]Another option is ℓ_2 normalization. This normalization does not impose sparsity; it lets us control projections using a coordinatewise context bound.

Remark 2.2 (Truncation threshold). *The threshold τ balances the bias from discarding extreme values of $y_i S(x_i)$ against the variance reduction from truncation. We set $\tau = \tau_n := \sqrt{M(\sigma^2 + L_f^2)}\,\sqrt{n/\iota_1}$ with $\iota_1 := \log(16dT/\delta)$, where σ, L_f, M are as in Assumptions 2 and 3. This yields the optimal $\mathcal{O}(d/\sqrt{n})$ estimation rate (see Lemma B.1). Precisely, let $\hat{\theta}^{(n)}$ be the estimator* (1) *built from the first n exploration samples. Then, with probability at least $1 - \delta/8$, for all $n \le T_0$ at once,*

$$\|\hat{\theta}^{(n)} - \mu^*\theta^\star\|_1 \le \epsilon_n := C_\theta\, d\sqrt{\iota_1/n}, \qquad C_\theta := \tfrac{19}{5}\sqrt{M(\sigma^2 + L_f^2)}. \tag{2}$$

To resolve identifiability, we define the normalized estimator as $\hat{\theta}_0 = \hat{\theta}/\|\hat{\theta}\|_1$. We now establish the high-probability error bound for the normalized estimator $\hat{\theta}_0$.

Lemma 2.3 (Self-certifying normalized estimator). *Suppose $\|\hat{\theta}^{(n)} - \mu^*\theta^\star\|_1 \le \epsilon_n$, $\mu^* > 0$ and $\|\theta^\star\|_1 = 1$. Let $\hat{\theta}_0^{(n)} := \hat{\theta}^{(n)}/\|\hat{\theta}^{(n)}\|_1$ and define the* certificate

$$\hat{\gamma}_n := \min\Big\{2,\ \frac{2\epsilon_n}{\|\hat{\theta}^{(n)}\|_1}\Big\} \qquad (\hat{\gamma}_n := 2 \text{ and } \hat{\theta}_0^{(n)} := e_1 \text{ if } \hat{\theta}^{(n)} = 0). \tag{3}$$

Then (a) $\|\hat{\theta}_0^{(n)} - \theta^\star\|_1 \le \hat{\gamma}_n$; (b) $\hat{\gamma}_n$ is computed from data only and does not involve μ^*; and (c) if $\epsilon_n \le \mu^*/2$, then $\hat{\gamma}_n \le 4\epsilon_n/\mu^*$.

The idea is simple. Since $\|\mu^*\theta^\star\|_1 = \mu^*$, the observed norm $\|\hat{\theta}^{(n)}\|_1$ is itself an estimate of μ^*. So we can bound the error without knowing μ^*. If μ^* is small, $\hat{\gamma}_n$ stays large: the algorithm is never misled, it only learns slowly. The proof is in Appendix B.

Remark 2.4 (On Normalization and Norm Selection). *Lemma 2.3(c) demonstrates that the ℓ_1-normalization step does not degrade the fundamental convergence rate of the unnormalized estimator.*

3 Algorithm

We discuss our proposed algorithm Zooming Single Index Bandit with UCB (`ZoomSIB-UCB`).

Overview. `ZoomSIB-UCB` operates in two sequential phases. The high-level idea is to reduce the original high-dimensional contextual bandit problem to a sequence of one-dimensional bandit problems by exploiting the single-index structure. Since the reward depends on $x^\top\theta^\star$ alone, if we had access to $\theta^\star$ we could project every arm's context onto the real line and simply run a standard bandit algorithm on the resulting scalar values. Because $\theta^\star$ is unknown, we must first estimate it, and then carefully account for the estimation error when making decisions.

Phase 1 (Parameter Estimation). The algorithm dedicates the first T_0 rounds to estimating the unknown index direction $\theta^\star$. T_0 is not fixed in advance. It is chosen from the data by the stopping rule below.

During this phase, it pulls arms uniformly at random, collecting observations that are fed into the Stein-score estimator. The estimator produces $\hat{\theta}$ satisfying $\|\hat{\theta} - \mu^*\theta^\star\|_1 \le \epsilon$ with high probability. We then work with the ℓ_1-normalised version $\hat{\theta}_0 := \hat{\theta}/\|\hat{\theta}\|_1$, which estimates $\theta^\star$ directly. After each new sample we recompute $\hat{\theta}^{(n)}$ and its certificate $\hat{\gamma}_n$ (Lemma 2.3). We keep the estimate with the smallest certificate so far, and write $\bar{\gamma}_n := \min_{n' \le n} \hat{\gamma}_{n'}$ (with $\bar{\gamma}_0 := 2$).

The stopping rule: Let $L := 4\kappa\sqrt{\log(8dTK/\delta)}$. With high probability $\|x_{t,a}\|_\infty \le L$, so every index $x_{t,a}^\top\theta$ with $\|\theta\|_1 = 1$ lies in $[-L, L]$. Using $\hat{\theta}_0$ instead of $\theta^\star$ shifts an index by at most $L\bar{\gamma}_n$. So bins of width $2L\bar{\gamma}_n$ are the finest ones we can trust. The *certified width* is

$$g_n := \min\big\{2L,\ \max\{\Delta_{\min},\ 2L\bar{\gamma}_n\}\big\}, \qquad \Delta_{\min} := \max\Big\{T^{-1/3},\ \Big(\frac{\sigma\sqrt{2L\iota_2}}{L_{f'}\sqrt{T}}\Big)^{2/3}\Big\}, \tag{4}$$

where $\iota_2 := \log(16N_{\max}T/\delta)$ and $N_{\max} := \lceil 2LT^{1/3}\rceil + 1$. At width g_n the bias costs at most $2L_{f'}g_n$ per round. One exploration round costs at most $2L_f$. We keep exploring while

$$L_f\, n \ \le\ L_{f'}\, T\, g_n, \tag{5}$$

Algorithm 1 Zooming Single-Index Bandit with UCB (`ZoomSIB-UCB`)

Input: horizon T, number of arms K, failure probability δ, sub-Gaussian parameter κ of $\mathcal{D}$, noise parameter σ, reward bound L_f, Lipschitz bound $L_{f'}$, score moment bound M.
Initialize: $L = 4\kappa\sqrt{\log(8dTK/\delta)}$, $\iota_1 = \log(16dT/\delta)$, $\epsilon_n = C_\theta d\sqrt{\iota_1/n}$, $\Delta_{\min}$ and ι_2 as in (4); $n = 0$, $\bar{\gamma} = 2$, $\hat{\theta}_0 = e_1$, $g = 2L$.
Phase 1: (Certified exploration)
while $n < T$ **and** $L_f\, n \le L_{f'}\, T\, g$ **do**
  Pull arm a_t uniformly at random from $[K]$, observe y_t, and set $n \leftarrow n + 1$.
  Compute $\hat{\theta}^{(n)}$ by (1) from the n samples with $\tau_n = \sqrt{M(\sigma^2 + L_f^2)}\sqrt{n/\iota_1}$, and $\hat{\gamma}_n$ by (3).
  if $\hat{\gamma}_n < \bar{\gamma}$ **then**
    $\bar{\gamma} \leftarrow \hat{\gamma}_n$ and $\hat{\theta}_0 \leftarrow \hat{\theta}^{(n)}/\|\hat{\theta}^{(n)}\|_1$.
  end if
  $g \leftarrow \min\{2L, \max\{\Delta_{\min}, 2L\bar{\gamma}\}\}$.
end while
Set $T_0 = n$, $w = g$, $N = \lceil 2L/w \rceil$, $\Delta = 2L/N$. Partition $[-L, L]$ into N bins with centres $\tilde{z}_j = -L + (j - 1/2)\Delta$. Init $n_j = 0$, $S_j = 0$ for all $j \in [N]$.
Phase 2: (UCB over Bins)
for round $t = T_0 + 1, \ldots, T$ **do**
  Observe $x_{t,a}$ for all $a \in [K]$; compute $\hat{z}_{t,a} = x_{t,a}^\top \hat{\theta}_0$ and $b_{t,a} = \min\{N, \max\{1, \lceil (\hat{z}_{t,a} + L)/\Delta \rceil\}\}$.
  Let $\mathcal{B}_t = \{b_{t,a} : a \in [K]\}$. For $j \in \mathcal{B}_t$: $\mathrm{UCB}_j = +\infty$ if $n_j = 0$, else $\mathrm{UCB}_j = \frac{S_j}{n_j} + \sigma\sqrt{\frac{2\iota_2}{n_j}}$.
  Pull $a_t \in \arg\max_{a \in [K]} \mathrm{UCB}_{b_{t,a}}$, observe y_t, and update $n_{b_{t,a_t}} \leftarrow n_{b_{t,a_t}} + 1$, $S_{b_{t,a_t}} \leftarrow S_{b_{t,a_t}} + y_t$.
end for

that is, while the exploration cost so far is below the bias we would pay over the whole horizon at the current width. Both sides are known to the learner. T_0 is the first n at which (5) fails (or T). The comparison is between upper bounds on these costs. The floor $\Delta_{\min}$ balances the bin-learning term $\sigma\sqrt{(L/w)T\iota_2}$ against the approximation term $L_{f'}Tw$.

Phase 2 (UCB over Bins). Once $\hat{\theta}_0$ is computed, the algorithm discretizes the real line into $N = \lceil 2L/w \rceil \le 2LT^{1/3} + 1$ bins $B_1, \ldots, B_N$ of width $\Delta = 2L/N \le w$ covering $[-L, L]$, where $w := g_{T_0}$ is the final certified width. At each subsequent round t, the algorithm:

1. Projects each of the K presented arms onto the estimated index: $\hat{z}_{t,a} = x_{t,a}^\top \hat{\theta}_0$.
2. Determines which bin each arm falls into, forming the available bin set $\mathcal{B}_t \subseteq [N]$.
3. Runs UCB over the available bins—selecting the arm whose bin has the highest UCB index—and updates the sufficient statistics (pull count n_j and reward sum S_j) for that bin.

The certificate and Lipschitz assumption imply that all arms assigned to the same bin B_j have nearly the same mean reward, approximated by $\mu_j := f(\tilde{z}_j)$ where $\tilde{z}_j$ is the bin centre. If $w < 2L$, the bin radius and index-estimation error are each at most $w/2$. If $w = 2L$, there is one bin centered at zero and every true index lies within L of its centre on the context event. In either case, $|z_{t,a} - \tilde{z}_j| \le w$ and the reward error is at most $L_{f'}w$. When $T_0 < T$, the stopping rule bounds the resulting approximation cost by the exploration-cost bound.

Sleeping bins and zooming: At each round only a random subset $\mathcal{B}_t$ of bins contains an arm. So Phase 2 is a *sleeping* bandit, and the analysis compares with the best *available* bin $j_t^* = \arg\max_{j \in \mathcal{B}_t} \mu_j$ Kanade et al. (2009); Kleinberg et al. (2010). The grid is uniform, as in the Lipschitz-bandit trade-off $\Delta \approx T^{-1/3}$, but its width is set by the certificate. See Appendix E for details.

4 Regret Analysis

We bound the cumulative regret over the horizon T. Since the reward function f is not assumed to be monotonically increasing, the arm with the largest projection does not necessarily yield the highest reward. Let $a_t^* = \arg\max_{a \in [K]} f(x_{t,a}^\top \theta^\star)$ denote the optimal arm at round t, and define its true index value as $z_t^* = x_{t,a_t^*}^\top \theta^\star$.

Assumption 2 (Score Moment). *There exists a constant $M > 0$ such that $\mathbb{E}[S_j(X)^2] \le M$ for all $j \in [d]$, where $X \sim \mathcal{D}$.*

Remark 4.1. *Our Assumption 2 requires only a bounded second moment of the score function, which is same as Kang et al. (2025) for the Stein-based estimator.*

Assumption 3 (Context regularity and Lipschitz Continuity). *The context vectors are drawn i.i.d. from a continuous distribution $\mathcal{D}$ with a continuously differentiable density p satisfying $p(x) > 0$ for all $x \in \mathbb{R}^d$, and each coordinate of $X \sim \mathcal{D}$ is κ-sub-Gaussian. Furthermore, the unknown reward function $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, and there exist constants $L_f, L_{f'} > 0$ such that $|f(z)| \le L_f$ and $|f'(z)| \le L_{f'}$ for all z. In addition, $\mathbb{E}|f'(X^\top\theta^\star)| < \infty$, and $f(x^\top\theta^\star)p(x) \to 0$ as $\|x\| \to \infty$.*

The Stein moment and link conditions are closely related to those of Kang et al. (2025); GSTOR's general-link regret guarantee additionally uses Gaussian design.

Theorem 4.2 (Main Regret Bound). *Suppose Assumptions 1, 2, and 3 hold, $\|\theta^\star\|_1 = 1$, and $\mu^\star \neq 0$. Then, for any $\delta \in (0,1)$ with probability at least $1-\delta$, Algorithm 1 satisfies*[2]

$$R_T \le \widetilde{\mathcal{O}}\left[\left(\frac{d}{|\mu^\star|}T\right)^{2/3}\right].$$

Let us discuss the consequences of the above result.

Remark 4.3 (Improved Regret over Kang et al. (2025)). *Note that our result is a direct improvement over Kang et al. (2025), where the authors prove that for general (non-monotone) functions the regret is $\tilde{\mathcal{O}}(T^{3/4})$. On the other hand, we get an improved regret.*

Corollary 4.4 (Expected regret). *Under the same conditions as Theorem 4.2, and if $|f| \le L_f$ on $\mathbb{R}$ (so that $R_T \le 2L_fT$ always), setting $\delta = 1/T$, $\mathbb{E}[R_T] = \tilde{O}\big((1 + (d/|\mu^*|)^{2/3})T^{2/3}\big)$.*

No knowledge of μ^* is needed. The algorithm uses neither μ^* nor a lower bound on it. The certificate replaces μ^* by the observed $\|\hat{\theta}^{(n)}\|_1$, and the stopping rule picks T_0 from data. If μ^* is small, the algorithm explores longer and uses coarser bins, and the bound degrades smoothly.

4.1 Proof Sketch of Theorem 4.2

We first define a good event $\mathcal{E}$. On $\mathcal{E}$ (probability at least $1-\delta$), the Stein bound (2) holds for all n, and $\|x_{t,a}\|_\infty \le L$ for all t, a. Then every index lies in $[-L, L]$, so every arm gets a bin, and $\|\hat{\theta}_0 - \theta^\star\|_1 \le \bar{\gamma}_{T_0}$ by Lemma 2.3.

Step 1 (exploration is not too long). By Lemma 2.3(c) and $\bar{\gamma}_n \le 2$, we get $2L\bar{\gamma}_n \le A/\sqrt{n}$, so $g_n \le \Delta_{\min} + A/\sqrt{n}$. The test (5) passed at $n = T_0 - 1$, so $L_f n \le L_{f'} T(\Delta_{\min} + A/\sqrt{n})$. Solving gives $T_0 \le 2Q/L_f + (2L_{f'}A/L_f)^{2/3}T^{2/3} + 1$. Phase 1 costs at most $2L_fT_0$.

Step 2 (bias is paid for by exploration). Each arm is within w of its bin centre, so its mean is within $L_{f'}w$ of μ_j. If $T_0 < T$, the test failed at $n = T_0$, so $L_{f'}Tw < L_fT_0$; if $T_0 = T$, Phase 2 is empty. Hence the total bias in Phase 2 is below $4L_fT_0$.

Conditioned on $\mathcal{E}$, the regret is cleanly decomposed $R_T = R_T^{(1)} + R_T^{(2)}$ into Phase 1 and Phase 2. For Phase 2, we decompose the instantaneous regret into two parts- (a) discretization bias and (b) (core) bandit regret.

The discretization bias is handled by the Lipschitzness of f (Lemma C.5). The core bandit regret is handled by a sleeping-UCB analysis over the stochastic active bin sets (Lemma C.6). For N bins, the regret scales with $\sqrt{NT}$.Since $N \le 4L/\Delta_{\min}$, the choice of $\Delta_{\min}$ gives $\sigma\sqrt{NT\iota_2} \lesssim L_{f'}T\Delta_{\min} = Q$. Adding the three parts gives Theorem 4.2 (Appendix C).

[2]The suppressed constants depend on $L_f, L_{f'}, \sigma, M, \kappa$. Assume also $\Delta_{\min} \le 2L$.

5 MINIMAX LOWER BOUND

We establish the matching information-theoretic lower bound, demonstrating that no algorithm can achieve a regret scaling better than $\Omega(T^{2/3})$ in the worst case for the general single-index bandit.

Theorem 5.1 (Minimax Lower Bound)**.** *Let the time horizon T be sufficiently large. Under the standard regularity conditions (Assumptions 2,3), there exist an absolute constant $c > 0$ and $T_{\min} \in \mathbb{N}$ such that for all $T \geq T_{\min}$ and every policy π,*

$$\sup_{(\theta^\star, f, \mathcal{D})} \mathbb{E}\left[R_T(\pi)\right] \geq c\, T^{2/3},$$

where the supremum is taken over valid problem instances with ambient dimension $d = 1$, parameter $\theta^\star = +1$, context distribution $\mathcal{D} = N(0,1)$, $N = \lceil T^{1/3} \rceil$ and the number of available arms $K = \lceil 4N \log T / p_{\min} \rceil$, where ϕ is the standard Gaussian density and $p_{\min} = \phi(1)$ and $\mu^* \geq c_1/(2 \log T) > 0$.

The same lower bound also holds with a signal bounded away from zero: taking $\lambda = 1/4$ and $C_\rho \leq 4$ gives $\mu^\star \geq c_1/8$ for all sufficiently large T, with $L_f = 1$ and $L_{f'} = 3$.

Remark 5.2. *Combined with the $\tilde{\mathcal{O}}(T^{2/3})$ upper bound established in Theorem 4.2, this confirms that the minimax regret for the non-monotone single-index bandit is $\tilde{\Theta}(T^{2/3})$ in its dependence on T. This lower bound does not settle fixed K. The lower bound is for $d = 1$, so the optimal dependence on d remains open.*

5.1 OVERVIEW OF THE LOWER BOUND PROOF

The lower bound relies on a three-step reduction strategy: construct a family of nearly-identical problem instances, show that any algorithm must be confused between them for a long time, and conclude via information theory that this confusion forces $\Omega(T^{2/3})$ regret on at least one instance.

Step 1: Constructing a Localized Hard Family. We work in dimension $d = 1$ with $\theta^\star = 1$ and standard Gaussian contexts $\mathcal{D} = N(0,1)$. We partition the interval $[0,1]$ into $N = \lceil T^{1/3} \rceil$ equally spaced bins. For each Boolean vector $\beta \in \{-1,+1\}^N$, we define a link function $f_\beta(z) = \frac{1}{2} + \lambda\, \rho(z) + \varepsilon \sum_{j=1}^N \beta_j\, \psi_j(z)$ by embedding small, plateaued bump functions ψ_j of height $\varepsilon = 1/(2N)$ within each bin on top of a common smooth *ramp* ρ scaled by $\lambda = 1/\log T$ (Lemma D.1). Here ρ is a fixed, infinitely differentiable function that rises from 0 to 1 over $(-1,0)$ and is flat ($\rho \equiv 1$) on $[0,1]$; it therefore never interferes with the bumps, yet its slope on $(-1,0)$ guarantees a strictly positive expected Stein score $\mu^* \geq c_1 \lambda/2 > 0$ (Lemma D.1(d)), ensuring every instance is a *valid* single-index model with non-trivial signal rather than a degenerate one, as Theorem 4.2 requires ($\mu^* \neq 0$). This generates a family of 2^N functions that are globally identical except for $\pm\varepsilon$ perturbations localized inside the bins, strictly satisfying the boundedness and Lipschitz assumptions. Alternatively, fixing $\lambda = 1/4$ and $C_\rho \leq 4$ yields $\mu^\star \geq c_1/8$ without changing the availability or KL arguments; this constant-signal variant supports the matching claim.

Step 2: Reducing SIB to N-armed bandit problem (Section D.2). The primary technical hurdle in the single-index bandit is the *sleeping* structure: at each round, only a random subset of bins contains arms. We remove this complication by providing a massive menu of arms ($K = \lceil 4N \log T / p_{\min} \rceil$) so that, with high probability, every single localized inner bin contains at least one available arm at every round (Lemma D.2). Conditioned on this universal availability event $\mathcal{E}$, the sleeping structure vanishes, and the problem collapses into a pure, perpetually available N-armed stochastic bandit in which every pull of a downward-pointing bin incurs instantaneous regret at least ε (relative to the optimal reward $\frac{1}{2} + \lambda + \varepsilon$ attained on an active up-bump bin) (Lemma D.3).

Step 3: The Information-Theoretic Trap (Section D.3). To trap the algorithm, we isolate a *null* reference hypothesis $\beta^{(0)}$ (where every bump points downward, making every bin slightly sub-optimal) and N parallel *spike* hypotheses $\beta^{(j)}$ (identical to the null, except bin j is flipped to point upward, making it uniquely optimal). Since the ramp term $\lambda\rho$ is shared by all hypotheses and is flat over the bins, it cancels in every pairwise comparison, so null and spike differ only by the $\pm\varepsilon$ bump in bin j (a mean gap of 2ε). Because the bump height ε is extremely small, distinguishing the null from a spike requires massive exploration (Proposition D.4).

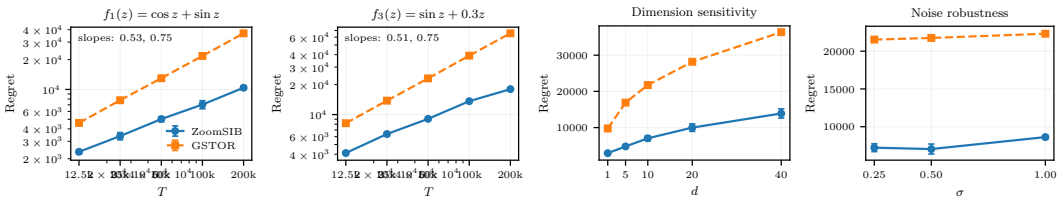

Figure 1: From left to right: horizon scaling for f_1 and f_3, dimension sensitivity, and noise robustness. Full results are in Appendix F.

The Bretagnolle-Huber inequality formalizes this into an inescapable dilemma for any algorithm: if the algorithm plays conservatively and explores bin j rarely, it cannot identify the optimal arm when placed in the spike hypothesis; conversely, if it explores aggressively to check for a spike, it suffers heavy regret pulling a downward-pointing bump if the true environment is the null hypothesis. Balancing this bias-variance tradeoff strictly dictates the choice $N = \Theta(T^{1/3})$, forcing a minimum expected regret of $\Omega(T^{2/3})$.

6 Experiments

Setup. Unless otherwise stated, $d = 10$, $K = 20$, $X \sim \mathcal{N}(0, I_d)$, $\theta^\star = e_1$, $\sigma = 0.5$, and we use 10 seeds. We test $f_1(z) = \cos z + \sin z$, $f_2(z) = ze^{-z^2}$, and $f_3(z) = \sin z + 0.3z$, comparing ZoomSIB-UCB with GSTOR. Appendix F.1 gives the full protocol. We evaluate calibrated practical implementations rather than literal theoretical constants; tuning budgets are not matched.

Horizon scaling. The first experiment asks how regret grows when the problem instance itself is held fixed. Only T changes. The fitted ZoomSIB slopes are 0.534, 0.544, and 0.509 for f_1, f_2, f_3, respectively. The corresponding GSTOR slopes are 0.747, 0.738, and 0.750. Figure 1 shows the results for f_1 and f_3. The complete curves are in Appendix F.2.

Table 1: Representative regret; $T = 200{,}000$ (synthetic) and $50{,}000$ (CoverType).

Setting / function	**ZoomSIB**	GSTOR	GSTOR / Zoom	T_0/T
$\cos z + \sin z$	**10,401**	36,775	3.54	0.041
ze^{-z^2}	**10,446**	20,312	1.94	0.074
$\sin z + 0.3z$	**18,017**	65,567	3.64	0.041
CoverType	**9,858**	24,369	2.47	0.104

Dimension sensitivity. We fix $\theta^\star = e_1$, so $X^\top \theta^\star \sim \mathcal{N}(0, 1)$ for every d. Thus, changing d changes the difficulty of estimating the direction, but not the one-dimensional reward problem. Regret increases with d, while ZoomSIB uses substantially fewer exploration rounds than GSTOR. Full results are in Appendix F.3.

Noise robustness. Increasing σ from 0.25 to 1 raises ZoomSIB's exploration fraction from 5.2% to 7.2% and increases its regret moderately. GSTOR uses the same exploration fraction across these noise levels. Full results are in Appendix F.6.

Additional experiments. The signal-strength experiment shows that ZoomSIB explores longer as $|\mu^\star|$ decreases, although $\mu^\star$ is not supplied to the algorithm. The selected exploration duration is also close to the best fixed choice chosen with hindsight. Finally, we test real context vectors from KDD99 and CoverType. On KDD99 the two methods have similar regret, while on CoverType ZoomSIB has about $2.47\times$ lower regret. These results are reported in Appendix F.4–F.7.

7 Conclusion and Open Problems

We establish $\tilde{O}(T^{2/3})$ regret for the stated non-monotone single-index class. For fixed dimension and fixed problem bounds, including a positive lower bound on the signal, the lower bound matches this horizon exponent when $K \asymp T^{1/3} \log T$. Several questions remain open: the optimal dependence on dimension and signal strength, and the minimax rate for fixed K. Our analysis relies on the Stein estimator and requires $\mu^* \neq 0$, so an alternative technique that removes this condition would be valuable. Other directions are single-index bandits beyond bounded Lipschitz structure and the implications of our results for the multi-index setting.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods, specifically writing the Python code for the experiments and for plotting the results in Section 6 and Appendix F. All experimental design, including which experiments to run and which questions to investigate, was conceived by the authors; generative AI was used only to write the code, not to design the experiments. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology or experiments, supporting qualitative or thematic data analysis, or interpreting results; translation and dataset cleaning/reformatting are not applicable to this work. Additionally, we used generative AI tools for editing and grammar-checking of the manuscript text. We have reviewed all AI-assisted work: LLM-generated code was verified manually for correctness by one of the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Annual Conference Computational Learning Theory*, 2022. URL https://api.semanticscholar.org/CorpusID:246904424.

Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3691–3699, 2021.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 263–274, 2008.

Rajeev Agrawal. Continuum-armed bandit problem. *SIAM Journal on Control and Optimization*, 33 (6):1926–1951, 1995.

Sakshi Arya and Hyebin Song. Semi-parametric batched global multi-armed bandits with covariates. *arXiv preprint arXiv:2503.00565*, 2025.

Sakshi Arya, Satarupa Bhattacharjee, and Bharath K Sriperumbudur. Kernel single-index bandits: Estimation, inference, and learning. *arXiv preprint arXiv:2603.18938*, 2026.

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

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT)*, pages 454–468, 2007.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Ilija Bogunovic and Andreas Krause. Misspecified gaussian process bandit optimization. *Advances in neural information processing systems*, 34:3004–3015, 2021.

David R Brillinger. A generalized linear model with Gaussian regressor variables. *A Festschrift for Erich L. Lehmann*, pages 97–114, 1982.

Joan Bruna and Daniel Hsu. Survey on algorithms for multi-index models. *Statistical Science*, 40(3): 378–391, 2025.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. $\mathcal{X}$-armed bandits. In *Journal of Machine Learning Research*, volume 12, pages 1655–1695, 2011a.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer, 2011b.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23, pages 41.1–41.14, 2012.

Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 190–198, 2012.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, pages 355–366, 2008.

Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 169–178, 2011.

Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3052–3062, 2020.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/c2626d850c80ea07e7511bbae4c76f4b-Paper.pdf.

Dylan J Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3199–3210, 2020.

Avishek Ghosh, Sayak Ray Chowdhury, and Aditya Gopalan. Misspecified linear bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Larry Goldstein, Stanislav Minsker, and Xiaohan Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Transactions on Information Theory*, 64:5513–5530, 2016. URL https://api.semanticscholar.org/CorpusID:49728615.

Wolfgang Härdle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178, 1993.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Hidehiko Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1-2):71–120, 1993.

Kwang-Sung Jun, Rebecca Willett, Stephen Wright, and Robert Nowak. Bilinear bandits with low-rank structure. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3163–3172, 2019.

Sham M Kakade, Adam Tauman Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pages 927–935, 2011.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *Annual Conference Computational Learning Theory*, 2009. URL `https://api.semanticscholar.org/CorpusID:7415296`.

Varun Kanade, H. Brendan McMahan, and Brent Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 272–279, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL `https://proceedings.mlr.press/v5/kanade09a.html`.

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. *arXiv preprint arXiv:2506.12751*, 2025.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 681–690, 2008.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2):245–272, 2010.

Robert D Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 17, pages 697–704, 2004.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, pages 817–824, 2008.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5662–5670, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670, 2010.

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

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Proceedings of the 31st Annual Conference on Learning Theory (COLT)*, pages 1463–1492, 2018.

Wanteng Ma and T Tony Cai. Nonparametric bandits with single-index rewards: Optimality and adaptivity. *arXiv preprint arXiv:2512.24669*, 2025.

Peter McCullagh and John A Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition, 1989.

Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pages 783–791, 2011.

Yaniv Plan and Roman Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Journal of Machine Learning Research*, volume 15, pages 2533–2568, 2014.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems (NeurIPS)*, 28:3420–3428, 2015.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 19–27, 2013.

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.

A Other Related Work

A.1 Linear Bandits, Contextual Bandits and Generalized Linear Bandits

Linear bandits extend the finite-armed bandit problem to structured continuous action spaces, with Auer (2002) establishing the foundational optimism-under-uncertainty template and Dani et al. (2008) providing the first $\tilde{O}(d\sqrt{T})$ regret bound via confidence ellipsoids, later sharpened to a tight $\tilde{O}(d\sqrt{T})$ by the canonical OFUL algorithm of Abbasi-yadkori et al. (2011) using regularized least-squares estimation. The matching $\Omega(d\sqrt{T})$ lower bound of Dani et al. (2008) confirms optimality. The adversarial setting is studied by Abernethy et al. (2008) and Bubeck et al. (2012), who establish near-optimal $\tilde{O}(\sqrt{dT})$ bounds, while subsequent work extended the framework to sparsity (Carpentier and Munos, 2012), low-rank structure (Jun et al., 2019), and misspecification (Lattimore et al., 2020). Contextual bandits condition rewards on an observed context vector, with LinUCB (Li et al., 2010) and its theoretical analysis (Chu et al., 2011) achieving $\tilde{O}(d\sqrt{T})$ regret in the linear case; the agnostic setting is addressed by Langford and Zhang (2008) via epoch-greedy exploration and by Dudík et al. (2011) via oracle reductions, with Foster and Rakhlin (2020) ultimately achieving the optimal $\tilde{O}(\sqrt{KT})$ rate through regression oracle reductions.

Generalized linear bandits (GLBs) pass the inner product $\langle \theta, a \rangle$ through a known link function $\mu(\cdot)$, with GLM-UCB Filippi et al. (2010) establishing $\tilde{O}(d\sqrt{T})$ regret under bounded link functions, Li et al. (2017) improving curvature dependence and Faury et al. (2020) achieving refined bounds for logistic bandits via self-concordance. Also Abeille et al. (2021) derive instance-dependent lower bounds that connect GLBs to the broader theory of structured bandits.

A.2 Single Index Model

Semiparametric efficiency bounds and optimal estimation rates for the SIM in the classical low-dimensional regime are established by Ichimura (1993) and Härdle et al. (1993). Hornik et al. (1989) and McCullagh and Nelder (1989) connect the SIM to neural networks and generalized linear models respectively, situating it within a broader modeling landscape. In the high-dimensional setting where $d \gg n$, Plan and Vershynin (2016) provides a seminal analysis showing that the index vector θ^* can be estimated from one-bit compressed measurements at the rate $O(s \log d/n)$ under

sparsity, leveraging the connection between the SIM and generalized linear measurements. Brillinger (1982) has earlier observed that ordinary least squares recovers θ^* up to a scalar in the SIM under Gaussian designs, a phenomenon later formalized and extended by Plan and Vershynin (2016) and Thrampoulidis et al. (2015). The work of Kalai and Sastry (2009) introduce the Isotron algorithm, the first computationally efficient method to jointly learn θ^* and f with provable guarantees under Gaussian covariates. Kakade et al. (2011) subsequently extend these results via the STAP algorithm, achieving polynomial sample and computational complexity for learning SIMs with monotone link functions.

Through the lens of empirical process theory and Gaussian width, the estimation error of convex programs for SIM recovery was characterized in terms of the geometry of the parameter set by Plan and Vershynin (2016) and Goldstein et al. (2016). More recently, the connection between SIMs and gradient descent on two-layer neural networks has attracted significant attention: Barak et al. (2022) and Abbe et al. (2022) studied the SIM as a prototypical model for understanding what gradient-based methods can and cannot learn, establishing computational-statistical gaps tied to the information exponent of the link function f. Bietti et al. (2022) and Damian et al. (2022) further showed that a single gradient step on the population loss suffices to recover the index direction up to a polynomial sample complexity, formalizing the so-called leap exponent as the key quantity governing the difficulty of the learning problem under staircase-structured link functions.

Closest to our setting among recent work is that of Arya and Song (2025), who study a *batched* global bandit with covariates under a single-index structure and propose the BIDS algorithm, combining sliced inverse regression with dynamic binning along the estimated index. Their formulation differs from ours on four counts. They operate in the batched setting with M pre-specified batch endpoints, at which alone the policy may be updated, whereas we are fully sequential. They observe a single covariate X_t per round and select among a fixed arm set with *arm-specific* links $g^{(k)}(x) = f^{(k)}(x^\top \beta_0)$, whereas we observe K i.i.d. arm-feature vectors under a *common* link f; indeed Arya and Song (2025) themselves note that the formulation of Kang et al. (2025), which we adopt, "differs substantially" from theirs. Their rates are driven by a margin parameter α, giving T^{β_M} with $\beta_M = \frac{1-\gamma}{1-\gamma^M}$ and $\gamma = \frac{1+\alpha}{3}$, and monotonicity of the link plays no role in their analysis; we impose no margin condition and study precisely the non-monotone case. Finally, their Assumption 3 requires bounded context support with the projected density bounded above and below, and their estimation rate is *assumed* as a black box (their Assumption 6), while we require only coordinatewise sub-Gaussian contexts and prove the corresponding rate for the truncated Stein estimator. Their optimality claim is also conditional, holding only for fixed K, known α, and a supplied pilot direction; without a pilot they obtain $\widetilde{\mathcal{O}}(\max\{K^{1/3}T^{2/3}, K^{1-\beta_M}T^{\beta_M}\})$. We note that the two sets of results agree where they should: in the no-margin limit $\alpha \to 0$ with $M \to \infty$ one has $\beta_M \to 2/3$, matching the $\widetilde{\Theta}(T^{2/3})$ rate established here.

A.3 Lipschitz Bandits

In Agrawal (1995), the authors establish the feasibility of consistent estimation in the one-dimensional setting, and Kleinberg (2004) derived nearly tight $\tilde{O}(T^{2/3})$ regret bounds, later tightened and connected to nonparametric regression by Auer et al. (2007). Kleinberg et al. (2008) extend the framework to general metric spaces via the zooming algorithm, achieving instance-dependent regret scaling with the zooming dimension rather than the ambient dimension, with a matching minimax lower bound of $\Omega(T^{(d+1)/(d+2)})$ for d-dimensional spaces confirming the tightness of this dependence. The $\mathcal{X}$-armed bandit framework of Bubeck et al. (2011a) introduce the HOO algorithm, which operates on a hierarchical partition of the action space and achieves regret depending on the near-optimality dimension, providing a unified treatment of Lipschitz and smooth bandits; Munos (2011) subsequently develop the parameter-free SOO and StoSOO algorithms, and Valko et al. (2013) extend this line of work to kernel-based methods exploiting RKHS structure. Slivkins (2014) generalize the framework to contextual bandits via a joint similarity metric over contexts and actions, Bubeck et al. (2011b) derive instance-dependent logarithmic regret bounds under a margin condition analogous to Tsybakov's noise condition, and Locatelli and Carpentier (2018) study adaptive algorithms achieving optimal rates across smoothness classes without knowledge of the Lipschitz constant.

B PROOFS

To establish the convergence of our normalized estimator, we first require a high-probability bound on the unnormalized truncated Stein estimator. Under our specific moment and regularity assumptions, this bound is achieved by carefully setting the truncation threshold τ. For completeness, we adapt the underlying concentration argument to our setting:

Assumption 1 only makes the noise conditionally sub-Gaussian, so the samples need not be i.i.d. Also, the number of exploration samples is random. We therefore need a bound that holds for all n at once. Its proof uses Freedman's inequality for martingales instead of an i.i.d. Bernstein bound.

Lemma B.1 (Stein estimation error, martingale version)**.** *Let Assumptions 1, 2 and 3 hold. Let* $v := \sqrt{M(\sigma^2 + L_f^2)}$ *and* $\iota_1 := \log(16dT/\delta)$*. For* $n \geq 1$*, let* $\hat{\theta}^{(n)}$ *be the estimator* (1) *built from the first* n *exploration samples with* $\tau_n := v\sqrt{n/\iota_1}$*. Then, with probability at least* $1 - \delta/8$*, for every* $n \leq T_0$*,*

$$\|\hat{\theta}^{(n)} - \mu^*\theta^\star\|_1 \leq \epsilon_n := C_\theta\, d\sqrt{\frac{\iota_1}{n}}, \qquad C_\theta := \frac{19}{5}\, v.$$

Proof. In Phase 1 every round explores, so the m-th sample comes from round m, and rounds $1, \ldots, T_0$ are the exploration rounds. Let $\mathcal{F}_m := \sigma(\mathcal{X}_{1:m}, a_{1:m}, y_{1:m})$. Include also the learner randomness $U_{1:m}$ in $\mathcal{F}_m$. Whether round m explores is decided from the first $m-1$ samples, so the event $\{m \leq T_0\}$ is $\mathcal{F}_{m-1}$-measurable. Fix $n \in [T]$ and $j \in [d]$, and write $x_m := x_{m,a_m}$ and $V_m := y_m S_j(x_m)$.

Step 1: conditional moments. On $\{m \leq T_0\}$, the arm a_m is uniform and independent of $\mathcal{X}_m$ and of the past, so $x_m \sim \mathcal{D}$ given $\mathcal{F}_{m-1}$. Also x_m is $\mathcal{G}_m$-measurable and $\mathbb{E}[\eta_m \mid \mathcal{G}_m] = 0$, $\mathbb{E}[\eta_m^2 \mid \mathcal{G}_m] \leq \sigma^2$. Hence, by Stein's identity and Assumption 3,

$$\mathbb{E}[V_m \mid \mathcal{F}_{m-1}] = \mu^*\theta_j^*, \qquad \mathbb{E}[V_m^2 \mid \mathcal{F}_{m-1}] \leq \mathbb{E}[f(X^\top\theta^\star)^2 S_j(X)^2] + \sigma^2\mathbb{E}[S_j(X)^2] \leq v^2.$$

The cross term vanishes because $\mathbb{E}[\eta_m \mid \mathcal{G}_m] = 0$.

Step 2: truncation bias. For any $\tau > 0$,

$$\big|\mathbb{E}[\varphi_\tau(V_m) \mid \mathcal{F}_{m-1}] - \mu^*\theta_j^*\big| \leq \mathbb{E}\big[|V_m|\mathbf{1}\{|V_m| > \tau\} \mid \mathcal{F}_{m-1}\big] \leq \frac{\mathbb{E}[V_m^2 \mid \mathcal{F}_{m-1}]}{\tau} \leq \frac{v^2}{\tau}.$$

Step 3: Freedman's inequality. Let $D_m := \mathbf{1}\{m \leq T_0\}\big(\varphi_{\tau_n}(V_m) - \mathbb{E}[\varphi_{\tau_n}(V_m) \mid \mathcal{F}_{m-1}]\big)$ for $m \leq n$. This is a martingale difference sequence with $|D_m| \leq 2\tau_n$ and $\mathbb{E}[D_m^2 \mid \mathcal{F}_{m-1}] \leq v^2$. Freedman's inequality gives, with probability at least $1 - 2\delta'$,

$$\Big|\frac{1}{n}\sum_{m\leq n} D_m\Big| \leq \sqrt{\frac{2v^2\log(1/\delta')}{n}} + \frac{4\tau_n\log(1/\delta')}{3n}.$$

On $\{n \leq T_0\}$ all indicators equal 1, so $\hat{\theta}_j^{(n)} - \frac{1}{n}\sum_{m\leq n}\mathbb{E}[\varphi_{\tau_n}(V_m) \mid \mathcal{F}_{m-1}] = \frac{1}{n}\sum_{m\leq n} D_m$.

Step 4: combine. Take $\delta' = \delta/(16dT)$, so $\log(1/\delta') = \iota_1$. Using Steps 2–3 and $\tau_n = v\sqrt{n/\iota_1}$,

$$|\hat{\theta}_j^{(n)} - \mu^*\theta_j^*| \leq \Big(1 + \sqrt{2} + \frac{4}{3}\Big)v\sqrt{\frac{\iota_1}{n}} \leq \frac{19}{5}\, v\sqrt{\frac{\iota_1}{n}},$$

since $1 + \sqrt{2} + 4/3 < 3.75$. Summing over $j \in [d]$ gives the ℓ_1 bound. A union bound over $j \in [d]$ and $n \in [T]$ costs $dT \cdot 2\delta' = \delta/8$. □

Proof of Lemma 2.3. Part (b) holds because ϵ_n is an explicit number and $\|\hat{\theta}^{(n)}\|_1$ is observed.

For (a), first note that $\|\hat{\theta}_0^{(n)} - \theta^\star\|_1 \leq \|\hat{\theta}_0^{(n)}\|_1 + \|\theta^\star\|_1 = 2$ always. This covers the case $\hat{\gamma}_n = 2$, including $\hat{\theta}^{(n)} = 0$. Now let $\hat{\theta}^{(n)} \neq 0$. Since $\mu^* > 0$ and $\|\theta^\star\|_1 = 1$, we have $\|\mu^*\theta^\star\|_1 = \mu^*$. The reverse triangle inequality gives

$$\big|\|\hat{\theta}^{(n)}\|_1 - \mu^*\big| \leq \|\hat{\theta}^{(n)} - \mu^*\theta^\star\|_1 \leq \epsilon_n. \tag{6}$$

Adding and subtracting $\mu^*\theta^\star$ inside the norm,

$$\|\hat{\theta}_0^{(n)} - \theta^\star\|_1 = \frac{\|\hat{\theta}^{(n)} - \|\hat{\theta}^{(n)}\|_1\theta^\star\|_1}{\|\hat{\theta}^{(n)}\|_1} \le \frac{\|\hat{\theta}^{(n)} - \mu^*\theta^\star\|_1 + |\mu^* - \|\hat{\theta}^{(n)}\|_1|\,\|\theta^\star\|_1}{\|\hat{\theta}^{(n)}\|_1} \le \frac{2\epsilon_n}{\|\hat{\theta}^{(n)}\|_1}.$$

Together with the bound 2, this proves (a).

For (c), if $\epsilon_n \le \mu^*/2$, then (6) gives $\|\hat{\theta}^{(n)}\|_1 \ge \mu^* - \epsilon_n \ge \mu^*/2 > 0$. Hence $2\epsilon_n/\|\hat{\theta}^{(n)}\|_1 \le 4\epsilon_n/\mu^*$. □

Let $\mathcal{E}_1$ be the event of Lemma B.1. On $\mathcal{E}_1$, Lemma 2.3(a) holds for every $n \le T_0$. So the estimate kept by the algorithm (the one with the smallest certificate) satisfies $\|\hat{\theta}_0 - \theta^\star\|_1 \le \bar{\gamma}_n$, where $\bar{\gamma}_n = \min_{n' \le n} \hat{\gamma}_{n'}$ is non-increasing in n.

Lemma B.2 (Envelope of the certificate)**.** *Let $A := 8LC_\theta d\sqrt{\iota_1}/\mu^*$. On $\mathcal{E}_1$, for every $1 \le n \le T_0$,*

$$2L\bar{\gamma}_n \le \min\left\{4L,\ \frac{A}{\sqrt{n}}\right\} \qquad \textit{and hence} \qquad g_n \le \Delta_{\min} + \frac{A}{\sqrt{n}}.$$

Proof. Since $\bar{\gamma}_n \le 2$, we have $2L\bar{\gamma}_n \le 4L$. Note that $8L\epsilon_n/\mu^* = A/\sqrt{n}$. If $\epsilon_n \le \mu^*/2$, Lemma 2.3(c) gives $2L\bar{\gamma}_n \le 2L\hat{\gamma}_n \le 8L\epsilon_n/\mu^* = A/\sqrt{n}$. If $\epsilon_n > \mu^*/2$, then $A/\sqrt{n} = 8L\epsilon_n/\mu^* > 4L \ge 2L\bar{\gamma}_n$. Finally, $g_n = \min\{2L, \max\{\Delta_{\min}, 2L\bar{\gamma}_n\}\} \le \max\{\Delta_{\min}, A/\sqrt{n}\} \le \Delta_{\min} + A/\sqrt{n}$. □

This lemma is the only place where μ^* appears. It is used in the analysis only; the algorithm never evaluates A.

Remark B.3 (The sign of μ^*)**.** *If $\mu^* < 0$, apply everything to the equivalent model $(-\theta^\star, f(-\cdot))$, whose Stein signal is $|\mu^*| > 0$. Then $\|\hat{\theta}_0^{(n)} + \theta^\star\|_1 \le \hat{\gamma}_n$, so $\hat{z}_{t,a} \approx -z_{t,a}$ and the bins are simply reflected: bin j has mean $f(-\tilde{z}_j)$. The function $z \mapsto f(-z)$ has the same L_f and $L_{f'}$, and the algorithm never uses the sign or the monotonicity of f. So all results hold with μ^* replaced by $|\mu^*|$.*

C Proof of Theorem 4.2

C.1 High-Probability Events

Before presenting the lemmas, we collect the probabilistic scaffolding on which the entire analysis rests. To ensure the final theorem holds with probability at least $1 - \delta$, we scale the failure probability of each individual event to $\delta/8$.

Define $L := 4\kappa\sqrt{\log(8dTK/\delta)}$, and let $\iota_1, \iota_2, N_{\max}, \Delta_{\min}$ and ϵ_n be as in Section 3. We assume T is large enough that $\Delta_{\min} \le 2L$ (so the finest grid has at least one bin).

Define the following three events:

- $\mathcal{E}_1 := \{\|\hat{\theta}^{(n)} - \mu^*\theta^\star\|_1 \le \epsilon_n\ \forall\, n \le T_0\}$, the event that the Stein estimates are accurate at every sample size. By Lemma B.1, $\mathbb{P}(\mathcal{E}_1) \ge 1 - \delta/8$.
- $\mathcal{E}_2 := \{\|x_{t,a}\|_\infty \le L,\ \forall t \in [T], a \in [K]\}$, the event that all context norms are controlled. By the sub-Gaussian tail and union bound over dTK coordinates, $\mathbb{P}(\mathcal{E}_2) \ge 1 - \delta/8$.
- $\mathcal{E}_4 := \left\{|M_j(n)| \le \sigma\sqrt{2\iota_2/n}\ \ \forall\, j \in [N_{\max}],\ n \in [T]\right\}$, the event that the noise averages in each bin concentrate, where $M_j(n) = \frac{1}{n}\sum_{m=1}^{n} \eta_{T_j(m)}$ is the average noise over the first n pulls of bin j. By the Azuma-Hoeffding inequality for the sub-Gaussian MDS (See Lemma C.3) and a union bound over $j \in [N_{\max}]$ and $n \in [T]$, $\mathbb{P}(\mathcal{E}_4) \ge 1 - \delta/8$.

Throughout the analysis, we work on the intersection event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, which satisfies $\mathbb{P}(\mathcal{E}) \ge 1 - 3\delta/8 \ge 1 - \delta$.

We now prove $\mathcal{E}_2$ and $\mathcal{E}_4$. ($\mathcal{E}_1$ follows directly from the estimation lemmas in the preceding section, namely Lemma B.1.)

Lemma C.1 (Context norm bound $\mathcal{E}_2$)**.** *Let constant L and event $\mathcal{E}_2$ be as defined above. Then $\mathbb{P}(\mathcal{E}_2) \geq 1 - \delta/8$.*

Proof. Fix a coordinate $j \in [d]$ and an arm index (t, a). Since $(x_{t,a})_j$ is κ-sub-Gaussian with zero mean (the moment-generating-function bound in Assumption 3 forces zero mean), the standard sub-Gaussian tail gives $\mathbb{P}(|(x_{t,a})_j| > u) \leq 2\exp(-\frac{u^2}{2\kappa^2})$. Setting $u = L = 4\kappa\sqrt{\log(8dTK/\delta)}$ yields

$$\mathbb{P}\big(|(x_{t,a})_j| > L\big) \leq 2\Big(\frac{\delta}{8dTK}\Big)^8 \leq \frac{\delta}{8dTK},$$

since $2q^8 \leq q$ for $q = \delta/(8dTK) \leq 1/8$. A union bound over all $d \cdot TK$ coordinate-arm pairs gives $\mathbb{P}(\mathcal{E}_2^c) \leq \delta/8$.

□

Index Range On $\mathcal{E}_2$, Hölder's inequality gives $|x_{t,a}^\top \theta| \leq \|x_{t,a}\|_\infty \|\theta\|_1 \leq L$ for every θ with $\|\theta\|_1 = 1$. This covers both $\theta^\star$ and $\hat{\theta}_0$, so the window W is not needed. (An independence assumption on the coordinates is not needed either: for $\|\theta\|_1 = 1$, convexity of $u \mapsto e^u$ gives $\mathbb{E}[e^{\lambda x^\top \theta}] \leq \sum_j |\theta_j| \, \mathbb{E}[e^{\lambda \operatorname{sgn}(\theta_j) x_j}] \leq e^{\kappa^2\lambda^2/2}$.)

Lemma C.2 (Every arm is binned)**.** *On $\mathcal{E}_1 \cap \mathcal{E}_2$, for all $t > T_0$ and $a \in [K]$: $|z_{t,a}| \leq L$, $|\hat{z}_{t,a}| \leq L$, and*

$$|\hat{z}_{t,a} - z_{t,a}| \leq L\,\bar{\gamma}_{T_0}.$$

In particular, every arm gets a bin in $[N]$ and $\mathcal{B}_t \neq \emptyset$.

Proof. Both $\theta^\star$ and $\hat{\theta}_0$ have unit ℓ_1 norm, so Hölder's inequality and $\mathcal{E}_2$ give $|z_{t,a}|, |\hat{z}_{t,a}| \leq L$. Next, $|\hat{z}_{t,a} - z_{t,a}| = |x_{t,a}^\top(\hat{\theta}_0 - \theta^\star)| \leq \|x_{t,a}\|_\infty \|\hat{\theta}_0 - \theta^\star\|_1 \leq L\bar{\gamma}_{T_0}$, by Lemma 2.3(a) applied to the kept estimate. □

For each bin $j \in [N]$, let $T_j(m)$ denote the round at which bin j is pulled for the m-th time, and let $a_j(m)$ be the arm played at that round. Let the observed reward as, $y_{T_j(m)} = f\big(z_{T_j(m),a_j(m)}\big) + \eta_{T_j(m)}$. The empirical mean of bin j after n pulls is $\bar{y}_j(n) = \frac{1}{n}\sum_{m=1}^n y_{T_j(m)}$. Decomposing gives,

$$\bar{y}_j(n) - \mu_j = \underbrace{\frac{1}{n}\sum_{m=1}^{n}\big[f(z_{T_j(m),a_j(m)}) - \mu_j\big]}_{\text{deterministic bias } B_j(n)} + \underbrace{\frac{1}{n}\sum_{m=1}^{n}\eta_{T_j(m)}}_{\text{noise average } M_j(n)}. \tag{7}$$

On $\mathcal{E}_1 \cap \mathcal{E}_2$, every arm assigned to bin j satisfies $|f(z_{t,a}) - \mu_j| \leq L_{f'} w$ (proved in Lemma C.5 below), so $|B_j(n)| \leq L_{f'} w =: s$ pathwise.

Lemma C.3 (Noise concentration)**.** *Conditioned on Phase 1, for each bin $j \in [N_{\max}]$:*

$$\mathbb{P}\left(\exists\, n \in [T] :\ |M_j(n)| > \sigma\sqrt{\frac{2\iota_2}{n}} \,\middle|\, \textit{Phase 1}\right) \leq \frac{\delta}{8N_{\max}}.$$

Consequently, a union bound over $j \in [N_{\max}]$ gives $\mathbb{P}(\mathcal{E}_4^c) \leq \delta/8$. Here $M_j(n)$ is considered only when bin j has been pulled at least n times by T; bins with no such pull count impose no condition.

Proof. The key observation is that $\hat{\theta}_0$ is computed entirely from Phase 1 data, so it is fixed throughout Phase 2. At round $t > T_0$, the arm a_t is selected based on past rewards, current contexts $\mathcal{X}_t$, and the frozen $\hat{\theta}_0$, making it $\mathcal{G}_t$-measurable (it depends on past rewards, current contexts, and the frozen $\hat{\theta}_0$). By Assumption 1, η_t is conditionally σ-sub-Gaussian given $\mathcal{G}_t$ and independent of x_t. Therefore, for the subsequence $T_j(1) < T_j(2) < \cdots$ at which bin j is pulled, each $\eta_{T_j(m)}$ satisfies $\mathbb{E}\big[e^{\lambda\eta_{T_j(m)}} \mid \mathcal{G}_{T_j(m)}\big] \leq e^{\lambda^2\sigma^2/2}$, which shows that $\{\eta_{T_j(m)}\}_{m\geq 1}$ is a martingale difference sequence (MDS) with σ-sub-Gaussian increments. Applying the Azuma-Hoeffding bound and a union bound over $n \in [T]$ gives the result. For fixed (j, n), the two-sided bound is $2e^{-\iota_2} = \delta/(8N_{\max}T)$. Since

T_0 is a stopping time, conditioning on the data up to T_0 keeps this martingale structure. More explicitly, condition on $\mathcal{F}_{T_0}$ and stop the nonnegative exponential supermartingale

$$\exp\left\{ u \sum_{t=T_0+1}^{v} \mathbf{1}\{b_{t,a_t} = j\}\eta_t - \frac{u^2\sigma^2}{2} \sum_{t=T_0+1}^{v} \mathbf{1}\{b_{t,a_t} = j\} \right\}$$

at the nth pull of bin j, or at T if it has fewer pulls. Chernoff's bound controls the event that the nth pull occurs and its noise average exceeds the stated threshold. Selection is measurable before the reward is observed, by the enlarged filtration in Assumption 1. This justifies the conditional bound without a Markov assumption. □

Proof Roadmap: Conditioned on $\mathcal{E}$, the regret is cleanly decomposed $R_T = R_T^{(1)} + R_T^{(2)}$ into Phase 1 and Phase 2. Lemma C.4 bounds the length T_0 of Phase 1 and shows that the stopping rule pays for the bias. For Phase 2, we decompose the instantaneous regret into a discretisation bias—handled via the Lipschitzness of f (Lemma C.5)—and the core bandit regret, which we bound using a sleeping-UCB analysis with bounded within-bin bias over the stochastic active bin sets (Lemma C.6).

C.2 Phase 1 Regret

Lemma C.4 (Length of Phase 1). *Let $Q := L_{f'} T \Delta_{\min}$ and A as in Lemma B.2. On $\mathcal{E}_1$,*

$$T_0 \le \bar{n} := \frac{2Q}{L_f} + \Big(\frac{2L_{f'}A}{L_f}\Big)^{2/3} T^{2/3} + 1.$$

Moreover, if $T_0 < T$, then $L_{f'} T w < L_f T_0$, where $w = g_{T_0}$. On $\mathcal{E}_2$, the regret of Phase 1 satisfies $R_T^{(1)} \le 2L_f T_0$.

Proof. If $T_0 \le 1$ there is nothing to prove. Otherwise let $n := T_0 - 1 \ge 1$. The test (5) passed at n, so by Lemma B.2,

$$L_f n \le L_{f'} T g_n \le L_{f'} T \Big(\Delta_{\min} + \frac{A}{\sqrt{n}}\Big) = Q + \frac{L_{f'} T A}{\sqrt{n}}.$$

One of the two terms on the right is at least $L_f n/2$. In the first case $n \le 2Q/L_f$. In the second case $n^{3/2} \le 2L_{f'} T A / L_f$, i.e., $n \le (2L_{f'}A/L_f)^{2/3} T^{2/3}$. Either way $T_0 = n + 1 \le \bar{n}$.

If $T_0 < T$, the loop stopped because the test failed at $n = T_0$, i.e., $L_f T_0 > L_{f'} T g_{T_0} = L_{f'} T w$.

Finally, on $\mathcal{E}_2$ every index lies in $[-L, L]$, where $|f| \le L_f$. So each round costs at most $2L_f$. □

C.3 Phase 2 Regret Decomposition

For $t > T_0$, let $b_t^* = b_{t,a_t^*}$ denote the bin assigned to the optimal available arm, and $b_t = b_{t,a_t}$ denote the bin of the arm actually played. We decompose the instantaneous regret as:

$$f(z_t^*) - f(z_{t,a_t}) = \underbrace{\big[f(z_t^*) - \mu_{b_t^*}\big]}_{\text{(i)}} + \underbrace{\big[\mu_{b_t^*} - \mu_{b_t}\big]}_{\text{(ii)}} + \underbrace{\big[\mu_{b_t} - f(z_{t,a_t})\big]}_{\text{(iii)}} \tag{8}$$

where $\mu_j := f(\tilde{z}_j)$ is the true reward evaluated at the bin centre $\tilde{z}_j$. Terms (i) and (iii) represent the discretisation bias (the error from approximating a continuous arm's reward by its bin centre). Term (ii) is bounded above by the regret against the best available bin: $\mu_{b_t^*} - \mu_{b_t} \le \mu_{j_t^*} - \mu_{j_t}$. Term (ii) itself need not be nonnegative.

Lemma C.5 (Discretisation Bias). *On $\mathcal{E}_1 \cap \mathcal{E}_2$, for every $t > T_0$ and every arm a assigned to bin $b_{t,a} = j$, we have $|z_{t,a} - \tilde{z}_j| \le w$ and $|f(z_{t,a}) - \mu_j| \le L_{f'} w =: s$. Consequently, if $T_0 < T$,*

$$\sum_{t=T_0+1}^{T} \big[\textit{(i)} + \textit{(iii)}\big] \le 2L_{f'} T w < 2L_f T_0.$$

Proof. *Case* $w = 2L$. Then $N = 1$ and $\tilde{z}_1 = 0$. By Lemma C.2, $|z_{t,a} - \tilde{z}_1| = |z_{t,a}| \le L \le w$.

Case $w < 2L$. Then the cap in (4) is not active, so $w = \max\{\Delta_{\min}, 2L\bar{\gamma}_{T_0}\} \ge 2L\bar{\gamma}_{T_0}$. By Lemma C.2, $|z_{t,a} - \hat{z}_{t,a}| \le L\bar{\gamma}_{T_0} \le w/2$. By the definition of the bin, $|\hat{z}_{t,a} - \tilde{z}_j| \le \Delta/2 \le w/2$, since $\Delta = 2L/N \le w$. The triangle inequality gives $|z_{t,a} - \tilde{z}_j| \le w$.

In both cases $z_{t,a}, \tilde{z}_j \in [-L, L]$, where $|f'| \le L_{f'}$. So $|f(z_{t,a}) - f(\tilde{z}_j)| \le L_{f'}w$. Applying this to $a = a_t^*$ and $a = a_t$ and summing over at most T rounds gives $2L_{f'}Tw$. The last inequality is Lemma C.4. □

Lemma C.6 (Sleeping-UCB Regret)**.** *Consider N bins with means $\mu_1, \dots, \mu_N$, stochastic availability $\mathcal{B}_t \subseteq [N]$, and per-pull deterministic bias $s = L_{f'}w$. Let UCB select $j_t = \arg\max_{j \in \mathcal{B}_t}(\bar{y}_j + U_j(n_j))$ where $U_j(n) = \sigma\sqrt{2\iota_2/n}$ (and $U_j(0) = +\infty$). Define the best available bin as $j_t^* = \arg\max_{j \in \mathcal{B}_t} \mu_j$. On $\mathcal{E}$, we have:*

$$\sum_{t=T_0+1}^{T} \left[\mu_{j_t^*} - \mu_{j_t}\right] \le 2L_f N + 4\sigma\sqrt{2NT\iota_2} + 2sT$$

Proof. On $\mathcal{E}$, the empirical bin means satisfy $|\bar{y}_j(n) - \mu_j| \le s + U_j(n)$ simultaneously for all $j \in [N]$ and all pull counts, where $s = L_{f'}w$ is the deterministic within-bin bias (bounded pathwise on $\mathcal{E}_1 \cap \mathcal{E}_2$ via Lemma C.5) and $U_j(n) = \sigma\sqrt{2\iota_2/n}$ is the stochastic confidence radius (bounded on $\mathcal{E}_4$ via Lemma C.3).

First pulls. Consider the rounds at which the played bin has $n_{j_t} = 0$. There are at most N of them, and each costs at most $\mu_{j_t^*} - \mu_{j_t} \le 2L_f$. At every other round, every available bin has been pulled at least once; otherwise its index would be $+\infty$ and it would be played. So the argument below applies to them.

Fix round t with $n_{j_t} \ge 1$. Since $j_t^* \in \mathcal{B}_t$, the UCB index of the optimal bin satisfies $\bar{y}_{j_t^*} + U_{j_t^*}(n_{j_t^*}) \ge \mu_{j_t^*} - s$. By the UCB selection rule, $\bar{y}_{j_t} + U_{j_t}(n_{j_t}) \ge \bar{y}_{j_t^*} + U_{j_t^*}(n_{j_t^*}) \ge \mu_{j_t^*} - s$. Combining this with the upper confidence bound on j_t gives:

$$\mu_{j_t^*} - \mu_{j_t} \le 2U_{j_t}(n_{j_t}) + 2s$$

Summing over t and bounding the confidence width sum by grouping rounds per arm yields:

$$\sum_{t > T_0 :\, n_{j_t} \ge 1} U_{j_t}(n_{j_t}) \le \sum_{j=1}^{N} \sum_{n=1}^{N_j(T)} \sigma\sqrt{\frac{2\iota_2}{n}} \le \sum_{j=1}^{N} 2\sigma\sqrt{2N_j(T)\,\iota_2}$$

Applying Cauchy-Schwarz with $\sum_j N_j(T) = T - T_0 \le T$:

$$\sum_{j=1}^{N} \sqrt{N_j(T)} \le \sqrt{N \sum_{j=1}^{N} N_j(T)} \le \sqrt{NT}$$

Therefore, $\sum_{t:\, n_{j_t} \ge 1} U_{j_t}(n_{j_t}) \le 2\sigma\sqrt{2NT\iota_2}$. Adding the first-pull cost $2L_f N$ and the deterministic $2sT$ gives $\sum_t [\mu_{j_t^*} - \mu_{j_t}] \le 2L_f N + 4\sigma\sqrt{2NT\iota_2} + 2sT$.

□

Theorem 4.2 (Restated). *Suppose Assumptions 1, 2 and 3 hold, $\|\theta^\star\|_1 = 1$, $\mu^* \ne 0$, and $\Delta_{\min} \le 2L$. With probability at least $1 - \delta$, Algorithm 1 satisfies $R_T \le 10P + 20Q + 8L_f + 4L_f LT^{1/3}$, where $P = L_f^{1/3} L_{f'}^{2/3} A^{2/3} T^{2/3}$, $Q = L_{f'} T \Delta_{\min}$ and $A = 8LC_\theta d\sqrt{\iota_1}/|\mu^*|$.*

Proof of Theorem 4.2. By Remark B.3 we may take $\mu^* > 0$. We work on $\mathcal{E}$, which has probability at least $1 - 3\delta/8$.

Phase 1. By Lemma C.4, $R_T^{(1)} \le 2L_f T_0$. If $T_0 = T$, then $R_T \le 2L_f T_0$ and we go directly to the last step.

Phase 2. Let $T_0 < T$. By the decomposition above, Lemma C.5 bounds the sum of (i) and (iii) by $2L_{f'}Tw < 2L_fT_0$. For (ii), $b_t^* \in \mathcal{B}_t$ gives $\mu_{b_t^*} \le \mu_{j_t^*}$, so (ii) $\le \mu_{j_t^*} - \mu_{j_t}$. By Lemma C.6 and $2sT = 2L_{f'}wT < 2L_fT_0$,

$$\sum_{t>T_0} \text{(ii)} \le 2L_fN + 4\sigma\sqrt{2NT\iota_2} + 2L_fT_0.$$

Adding the three parts,

$$R_T \le 6L_fT_0 + 2L_fN + 4\sigma\sqrt{2NT\iota_2}. \tag{9}$$

UCB term. Since $w \ge \Delta_{\min}$ and $2L/\Delta_{\min} \ge 1$, we have $N = \lceil 2L/w \rceil \le 2L/\Delta_{\min} + 1 \le 4L/\Delta_{\min}$, and also $N \le 2LT^{1/3} + 1$. By the definition of $\Delta_{\min}$ in (4), $\sigma\sqrt{2L\iota_2} \le L_{f'}\sqrt{T}\,\Delta_{\min}^{3/2}$. Hence

$$4\sigma\sqrt{2NT\iota_2} \le 4\sigma\sqrt{\frac{8L\iota_2 T}{\Delta_{\min}}} = 8\sigma\sqrt{2L\iota_2}\sqrt{\frac{T}{\Delta_{\min}}} \le 8L_{f'}T\Delta_{\min} = 8Q,$$

and $2L_fN \le 4L_fLT^{1/3} + 2L_f$.

Last step. By Lemma C.4,

$$6L_fT_0 \le 6L_f\bar{n} = 12Q + 6\cdot 2^{2/3}L_f^{1/3}L_{f'}^{2/3}A^{2/3}T^{2/3} + 6L_f = 12Q + 6\cdot 2^{2/3}P + 6L_f.$$

Plugging into (9) and using $6\cdot 2^{2/3} < 10$ gives $R_T \le 10P + 20Q + 8L_f + 4L_fLT^{1/3}$.

Rate. $Q = L_{f'}T\Delta_{\min} = \max\{L_{f'}T^{2/3},\ L_{f'}^{1/3}(\sigma\sqrt{2L\iota_2})^{2/3}T^{2/3}\} = \tilde{\mathcal{O}}(T^{2/3})$, with no polynomial dependence on d and no dependence on μ^*; L and ι_2 retain logarithmic dependence on d.. Since $A = 8LC_\theta d\sqrt{\iota_1}/\mu^*$, we have $P = \tilde{\mathcal{O}}((d/\mu^*)^{2/3}T^{2/3})$. Hence $R_T = \tilde{\mathcal{O}}((1+(d/\mu^*)^{2/3})T^{2/3})$. □

Remark C.7 (Unknown Lipschitz constant)**.** *Suppose the algorithm uses a guess $\hat{L}$ in place of $L_{f'}$ in* (4) *and* (5)*. The proof goes through with $L_{f'}$ replaced by $\hat{L}$ in P and Q, except that the bias bound becomes $2L_{f'}Tw < 2(L_{f'}/\hat{L})L_fT_0$. So the regret is multiplied by at most $\max\{1, L_{f'}/\hat{L}\}$, and the $T^{2/3}$ rate is kept. In contrast, σ, M and L_f must be valid upper bounds, since the certificate is correct only if ϵ_n is.*

Remark C.8 (Where μ^* matters)**.** *The threshold $n_0 = 4C_\theta^2 d^2\iota_1/\mu^{*2}$ is sufficient for $\epsilon_n \le \mu^*/2$ and the bound in Lemma 2.3(c). The certificate can be informative earlier; this threshold is not a necessary exploration duration. This burn-in does not appear in the bound: $n_0 \le T$ implies $n_0 \lesssim (d/\mu^*)^{2/3}T^{2/3}$ up to constants and logarithmic factors, so it is dominated by P whenever the bound is non-trivial.*

D Proof of Theorem 5.1

D.1 The Hard Instance

Context distribution and dimension. We work in dimension $d = 1$ with $\theta^* = 1$ (so $\|\theta^*\|_1 = 1$), and set the context distribution to $\mathcal{D} = N(0,1)$. Let ϕ and Φ denote the standard Gaussian density and CDF respectively. $\phi \in C^\infty$, $\phi(x) > 0$ for all $x \in \mathbb{R}$, coordinates are 1-sub-Gaussian, and $\mathbb{E}[S(X)^2] = \mathbb{E}[X^2] = 1 = M$. Define $p_{\min} := \min_{z\in[0,1]}\phi(z) = \phi(1) = \frac{1}{\sqrt{2\pi}}e^{-1/2}$ and $p_{\max} := \max_{z\in[0,1]}\phi(z) = \phi(0) = \frac{1}{\sqrt{2\pi}}$.

Noise. The noise is $\eta_t \overset{iid}{\sim} N(0,1)$, which satisfies Assumption 1 with $\sigma = 1$.

Parameter choices. Fix the following parameters throughout the proof:

$$N = \lceil T^{1/3} \rceil, \qquad K = \left\lceil \frac{4N\log T}{p_{\min}} \right\rceil, \qquad \varepsilon = \frac{1}{2N}, \qquad \lambda = \frac{1}{\log T}. \tag{10}$$

Bins and inner halves. Partition $[0,1]$ into N disjoint bins $B_j = [\frac{j-1}{N}, \frac{j}{N})$ of width $w = 1/N$, with centres $c_j = (j - 1/2)/N$. Define the inner half-bin:

$$B_j^\circ = \left[c_j - \tfrac{w}{4}, c_j + \tfrac{w}{4}\right] \subset B_j. \tag{11}$$

These inner halves have width $w/2 = 1/(2N)$ and are strictly pairwise disjoint. (In this section $w = 1/N$ is the bin width of the hard instance, not the width of Algorithm 1.)

Bump functions. Keep the same bins, inner halves and amplitude. Let $q(u) = 3u^2 - 2u^3$ for $0 \le u \le 1$, and define

$$\psi_j(z) = \begin{cases} 1, & |z - c_j| \le w/4, \\ q\big(\frac{4}{w}(w/2 - |z - c_j|)\big), & w/4 < |z - c_j| < w/2, \\ 0, & |z - c_j| \ge w/2. \end{cases}$$

Since $q(0) = q'(0) = q'(1) = 0$ and $q(1) = 1$, these bumps are continuously differentiable, equal one on B_j°, and vanish outside B_j. They satisfy $0 \le \psi_j \le 1$ and $\|\psi_j'\|_\infty \le 6N$. Only the transitions have changed.

Ramp function. Let $\rho : \mathbb{R} \to [0, 1]$ be a fixed C^∞ function satisfying $\rho(z) = 0$ for $z \le -1$, ρ is strictly increasing on $(-1, 0)$, and $\rho(z) = 1$ for $z \ge 0$. Let $C_\rho = \|\rho'\|_\infty$. Note that ρ is flat on $[0, 1]$, so it does not interfere with the bumps.

Hypothesis class. For each $\beta \in \{-1, +1\}^N$, define:

$$f_\beta(z) = \tfrac{1}{2} + \lambda\rho(z) + \varepsilon \sum_{j=1}^{N} \beta_j \psi_j(z), \qquad \forall z \in \mathbb{R}. \tag{12}$$

Lemma D.1 (Validity of the Instance)**.** *For every $\beta \in \{-1, +1\}^N$ and all sufficiently large T, the function f_β satisfies:*

- *(a) $\|f_\beta\|_\infty \le \frac{1}{2} + \lambda + \varepsilon \le 1 \le L_f$; hence also $\mathbb{E}[f_\beta(X)^2 S(X)^2] \le \mathbb{E}[X^2] = 1 \le L_f^2 M$, and the boundary condition of Assumption 3 holds;*
- *(b) $\|f_\beta'\|_\infty \le \max\{\lambda C_\rho, 3\} \le 3 \le L_{f'}$ for sufficiently large T, everywhere. The ramp and bump derivatives have disjoint supports, and at most one bump derivative is nonzero at any point;*
- *(c) $f_\beta \equiv \frac{1}{2} + \lambda + \varepsilon\beta_j$ on B_j°, since $\rho \equiv 1$ on $[0, 1] \supset B_j^\circ$;*
- *(d) $\mu^* := \mathbb{E}[f_\beta'(X)] = \lambda c_1 + O(T^{-1/3}) > 0$, where $c_1 = \int_{-1}^{0} \rho'(x)\phi(x)\,dx > 0$ is a positive constant independent of β and T. In particular, $\mu^* \ge \lambda c_1/2 \ge c_1/(2\log T)$, so $\mu^* \ne 0$ as Theorem 4.2 requires.*

Proof of part (d). We compute $\mu^* = \lambda\mathbb{E}[\rho'(X)] + \varepsilon \sum_j \beta_j \mathbb{E}[\psi_j'(X)]$. For the first term: ρ' is supported on $[-1, 0]$ with $\rho' > 0$ there, so $\mathbb{E}[\rho'(X)] = \int_{-1}^{0} \rho'(x)\phi(x)\,dx =: c_1 > 0$. Since $\phi(x) \ge \phi(1)$ on $[-1, 0]$ and $\int_{-1}^{0} \rho'(x)\,dx = 1$, we have $c_1 \ge \phi(1) > 0$.

For the second term: integration by parts gives $\int \psi_j'(x)\phi(x)\,dx = \int \psi_j(x)\, x\, \phi(x)\,dx$ (boundary terms vanish since ψ_j has compact support). Each integral is $O(w \cdot p_{\max}) = O(1/N)$. Summed over N terms with $\varepsilon = 1/(2N)$: total contribution is $O(1/N) = O(T^{-1/3})$.

For T large enough, $O(T^{-1/3}) < \lambda c_1/2$, giving $\mu^* \ge \lambda c_1/2 > 0$. □

Smoothness clarification. With the continuously differentiable bumps above, f_β already satisfies Assumption 3. No mollification is needed. The exact inner plateaus and disjoint supports are preserved, so the availability, per-round regret and KL arguments below apply unchanged.

Remark (constant signal). The preceding proof also applies with $\lambda = 1/4$ and a fixed ramp satisfying $C_\rho \le 4$. In this case, Lemma D.1(a)–(c) holds with $L_f = 1$ and $L_{f'} = 3$, while the proof of part (d) gives $\mu^\star = c_1/4 + O(T^{-1/3}) \ge c_1/8$ for all sufficiently large T. The availability, regret, and KL arguments are unchanged: the ramp is constant on the bins and is shared by every hypothesis.

D.2 The Universal Availability Event

To force the algorithm to face an N-armed bandit, the random menu of arms must cover every inner half-bin. Let $I_{t,j} = \mathbf{1}_{\{\exists a \in [K]: x_{t,a} \in B_j^\circ\}}$ represent the indicator that an arm lands in the inner half of bin j at round t. Define the universal availability event:

$$\mathcal{E} = \{I_{t,j} = 1 \text{ for all } (t,j) \in [T] \times [N]\}. \tag{13}$$

Lemma D.2 (Availability Probability)**.** *Under $\mathcal{D} = N(0,1)$ and $K = \lceil 4N \log T / p_{\min} \rceil$, we have $\mathbb{P}(\mathcal{E}^c) \le 2T^{-2/3}$.*

Proof. Since $\phi(x) \ge p_{\min}$ for all $x \in [0,1]$, and $|B_j^\circ| = 1/(2N)$, the probability that a single context falls in B_j° is at least $p_{\min}/(2N)$. Therefore:

$$\mathbb{P}(I_{t,j} = 0) \le \left(1 - \frac{p_{min}}{2N}\right)^K \le \exp\left(-\frac{K p_{\min}}{2N}\right) \le \exp(-2\log T) = T^{-2}.$$

Applying a union bound over $T \times N$ pairs gives $\mathbb{P}(\mathcal{E}^c) \le TN \cdot T^{-2} = N/T$. Since $N = \lceil T^{1/3} \rceil \le 2T^{1/3}$, we conclude $\mathbb{P}(\mathcal{E}^c) \le 2T^{-2/3}$. □

Lemma D.3 (Per-Round Regret)**.** *Let j_t denote the bin of the arm chosen at round t (set $j_t := 0$ if the chosen arm lies outside $[0,1]$). Under any hypothesis $\beta \ne -\mathbf{1}$, conditional on $\mathcal{E}$, the instantaneous regret satisfies $r_t(\beta) \ge \varepsilon \cdot \mathbf{1}_{\{\beta_{j_t} = -1\}}$, and $r_t(\beta) \ge \varepsilon$ when $j_t = 0$.*

Proof. On $\mathcal{E}$, every inner half-bin B_j° contains at least one arm. Because $\beta \ne -\mathbf{1}$, there exists an optimal bin $j^\star$ where $\beta_{j^\star} = +1$. By $\mathcal{E}$, there exists an arm $a^\star$ with $x_{t,a^\star} \in B_{j^\star}^\circ$, yielding $f_\beta(x_{t,a^\star}) = \frac{1}{2} + \lambda + \varepsilon$. Any arm outside $[0,1]$ achieves reward at most $\frac{1}{2} + \lambda < \frac{1}{2} + \lambda + \varepsilon$ (since $\rho \le 1$ and no bumps are active outside $[0,1]$). If the algorithm selects an arm in a sub-optimal bin where $\beta_{j_t} = -1$, the reward is at most $\frac{1}{2} + \lambda$, ensuring regret at least ε. □

D.3 KL Computation and the Bretagnolle-Huber Contradiction

Proposition D.4 (KL Divergence)**.** *For any algorithm π, hypothesis $\beta \in \{-1,+1\}^N$, and bin $j \in [N]$, let $\beta^{\oplus j}$ denote β with the j-th component flipped. Let $N_j^\pi(\beta)$ denote the expected number of pulls from bin j. Then:*

$$\mathrm{KL}\left(P_\beta^\pi \,\middle\|\, P_{\beta^{\oplus j}}^\pi\right) \le 2\varepsilon^2 N_j^\pi(\beta). \tag{14}$$

Proof. The contexts $\{x_{t,a}\}$ are drawn i.i.d. from $N(0,1)$ independently of β, and the noise is $\eta_t \sim N(0,1)$. By the chain rule for KL divergence, the divergence accumulates solely from the reward distributions of the selected arms. Because f_β and $f_{\beta^{\oplus j}}$ differ only inside B_j, and the Gaussian KL divergence for means differing by $\Delta\mu$ with unit variance is $\frac{1}{2}(\Delta\mu)^2$, with $|\Delta\mu| = 2\varepsilon|\psi_j(x)| \le 2\varepsilon$, summing over the $N_j^\pi(\beta)$ rounds where the chosen arm falls in bin j yields the result. □

Proof of Theorem 5.1. Let $\beta^{(0)} = -\mathbf{1}$ be the *null* hypothesis, and $\beta^{(j)} = -\mathbf{1} + 2e_j$ be the *spike* hypothesis where bin j is uniquely optimal. Define the event $A_j = \{\#\{t : j_t = j\} \ge T/2\}$.

Under $\beta^{(j)}$, on the event $\mathcal{E} \cap A_j^c$, the algorithm plays sub-optimal bins (bins other than j, or arms outside $[0,1]$) for more than $T/2$ rounds. By Lemma D.3:

$$\mathbb{E}_{\beta^{(j)}}[R_T] \ge \frac{\varepsilon T}{2} \mathbb{P}_{\beta^{(j)}}(\mathcal{E} \cap A_j^c) \ge \frac{\varepsilon T}{2}\left(\mathbb{P}_{\beta^{(j)}}(A_j^c) - 2T^{-2/3}\right). \tag{15}$$

By Proposition D.4, $\mathrm{KL}(P_{\beta^{(0)}} \| P_{\beta^{(j)}}) \le 2\varepsilon^2 N_j^\pi(\beta^{(0)})$. Since $\sum_j N_j^\pi(\beta^{(0)}) \le T$, Markov's inequality ensures there exists $J^\star \subseteq [N]$ with $|J^\star| \ge N/2$ such that for all $j \in J^\star$, $N_j^\pi(\beta^{(0)}) \le 2T/N$. For any $j \in J^\star$:

$$\mathrm{KL}\left(P_{\beta^{(0)}} \,\middle\|\, P_{\beta^{(j)}}\right) \le 2\varepsilon^2 \left(\frac{2T}{N}\right) = \frac{T}{N^3} \le 1,$$

where we substituted $\varepsilon = 1/(2N)$ and $N = \lceil T^{1/3} \rceil$.

The Bretagnolle-Huber inequality gives:

$$\mathbb{P}_{\beta^{(0)}}(A_j) + \mathbb{P}_{\beta^{(j)}}(A_j^c) \geq \frac{1}{2} \exp\left(-\mathrm{KL}\left(P_{\beta^{(0)}} \,\middle\|\, P_{\beta^{(j)}}\right)\right) \geq \frac{1}{2e}. \tag{16}$$

We now conduct a case analysis of $J^\star$:

- **Case A:** Suppose there exists $j \in J^\star$ with $\mathbb{P}_{\beta^{(j)}}(A_j^c) \geq \frac{1}{4e}$. For T large enough that $2T^{-2/3} \leq \frac{1}{8e}$, substituting into (15):

$$\mathbb{E}_{\beta^{(j)}}[R_T] \geq \frac{\varepsilon T}{2} \cdot \frac{1}{8e} = \frac{T}{32eN} \geq \frac{T}{32e(2T^{1/3})} = \frac{T^{2/3}}{64e}.$$

- **Case B:** Suppose for all $j \in J^\star$, $\mathbb{P}_{\beta^{(j)}}(A_j^c) < \frac{1}{4e}$. By (16), $\mathbb{P}_{\beta^{(0)}}(A_j) > \frac{1}{4e}$ for all $j \in J^\star$. Summing:

$$\frac{N}{8e} \leq \sum_{j \in J^\star} \mathbb{P}_{\beta^{(0)}}(A_j) \leq 2,$$

  which is impossible for $N \geq 16e$, i.e., $T \geq (16e)^3$.

(The bound $\sum_j \mathbb{P}_{\beta^{(0)}}(A_j) \leq 2$ holds because at most two bins can each receive at least $T/2$ of the T pulls.) Case A must hold. Taking the supremum over the hypothesis class concludes the proof with $c = 1/(64e)$. □

E Deferred Discussion from the Main Text

This section records further details on the assumptions and the reduction to bins.

Remark E.1 (Regularity conditions for Stein's identity)**.** *The identity $\mathbb{E}[f(X^\top\theta^*)S(X)] = \mathbb{E}[f'(X^\top\theta^*)]\theta^*$ follows from the Generalized Stein's Lemma Stein (1981), and requires the boundary term $f(x^\top\theta^*)p(x)$ to vanish as $\|x\| \to \infty$. We state this condition in Assumption 3. For example, bounded f satisfies this boundary condition whenever $p(x) \to 0$; the boundary condition itself is explicitly assumed. Additionally, the algorithm requires evaluation of the score function $S(x) = -\nabla \log p(x)$, and hence the density p must be known to the learner. This is consistent with the framework of Kang et al. (2025), where the same requirement holds implicitly. Using an estimated score in applications introduces a separate approximation error. Such an implementation is not covered automatically by the known-score theorem.*

Remark E.2. *The sub-Gaussian coordinate condition is standard in contextual bandits Abbasi-yadkori et al. (2011); Dani et al. (2008) and is adopted by (Kang et al., 2025, Assumption 2.3 and Appendix H.2). Gaussian and bounded distributions satisfy it. Sub-exponential distributions do not, but they only change L to $\mathcal{O}(\log(dTK/\delta))$. The stated theorem assumes $|f| \leq L_f$ on all of $\mathbb{R}$. A bound only on $[-L, L]$ does not imply the score-weighted moment bound used in the Stein certificate or the bound on regret on failure events. The unbounded-link experiment is therefore a test outside the stated assumptions, rather than an instance covered by this theorem.*

Connection to zooming algorithms: Standard zooming algorithms for Lipschitz bandits maintain an adaptive cover of the arm space: bins are created and refined on the fly, with finer resolution near high-reward regions. `ZoomSIB-UCB` instead uses a uniform partition of the projected index line at a resolution chosen from the data. The "zooming" here is therefore not adaptive in the classical sense—it is a one-shot reduction from a d-dimensional arm space to a uniform grid on $\mathbb{R}$, made possible by the single-index structure. The choice of $\Delta = T^{-1/3}$ directly resolves the fundamental bias-variance trade-off for discretizing Lipschitz functions. Grouping continuous arms into bins of width Δ introduces a per-round approximation bias of $\mathcal{O}(\Delta)$, accumulating to $\mathcal{O}(T\Delta)$ regret over the horizon. Conversely, making the bins finer increases the total number of bins $N \approx 1/\Delta$. The statistical cost to explore N arms using UCB is $\mathcal{O}(\sqrt{NT}) = \mathcal{O}(\sqrt{T/\Delta})$. Minimising the total regret $\mathcal{O}(T\Delta + \sqrt{T/\Delta})$ by balancing these two terms yields the optimal width $\Delta = T^{-1/3}$. At

this resolution, the number of bins is $N = \mathcal{O}(T^{1/3})$, ensuring both the discretization bias and the exploration regret scale at the optimal $\mathcal{O}(T^{2/3})$ rate. In ZoomSIB-UCB this width (more precisely $\Delta_{\min}$) is only a floor. Bins finer than the estimation error cannot be trusted, so the actual width is $w = g_{T_0} \geq \Delta_{\min}$.

Connection to sleeping experts and bandits: A distinctive and non-trivial feature of the algorithm is that at each round t only a random subset $\mathcal{B}_t$ of bins is actually available: a bin j is present in $\mathcal{B}_t$ only if at least one of the K freshly drawn arms happens to fall inside B_j. Since the arms $x_{t,a}$ are drawn i.i.d. from a continuous distribution each round, $\mathcal{B}_t$ is random and changes with t. This is precisely the stochastic availability model of sleeping bandits Kanade et al. (2009); Kleinberg et al. (2010). A standard UCB analysis that treats all N bins as perpetually available would be incorrect as a bin might go many rounds without any arm landing in it, and the UCB indices for absent bins should simply not be considered. The sleeping-UCB analysis correctly accounts for this by comparing the algorithm's reward only to the best available bin $j_t^* = \arg\max_{j \in \mathcal{B}_t} \mu_j$ at each round, rather than to the globally best bin.

Remark E.3 (Sample splitting)**.** *A crucial structural property of ZoomSIB-UCB is that $\hat{\theta}_0$ is computed entirely from Phase 1 data and is* fixed *throughout Phase 2. Therefore, the bin assignment $b_{t,a}$ is a deterministic function of $x_{t,a}$ and $\hat{\theta}_0$. Conditioned on Phase 1, the bin index of the selected arm is $\mathcal{G}_t$-measurable: it depends on the past and on the contexts $\mathcal{X}_t$, but not on y_t. Here T_0 is a stopping time, so "conditioned on Phase 1" means conditioned on the data up to round T_0. This makes the noise sequence within each bin a martingale difference sequence (MDS), which is essential for the regret analysis later.*

Modular reduction. The reduction applies to policies whose regret guarantees tolerate bounded, possibly history-dependent observation bias.

Proposition E.4. *Let $s = L_{f'} w$. Suppose $\mathcal{A}$ guarantees regret $\mathrm{Reg}_{\mathcal{A}}(N, T; s)$ against the best available reference mean under observations $y_t = \mu_{j_t} + b_t + \eta_t$, with adapted availability, pre-reward measurable b_t, $|b_t| \leq s$, and the stated conditional noise bound. Run Phase 1 as before and apply $\mathcal{A}$ to its bins. On the intersection of the approximation and policy-guarantee events,*

$$R_T \leq R_T^{(1)} + \mathrm{Reg}_{\mathcal{A}}(N, T; s) + 2L_{f'} T w.$$

Here $R_T^{(1)} \leq 2L_f T_0$, and $2L_{f'} T w < 2L_f T_0$ when $T_0 < T$. For sleeping UCB, Lemma C.6 supplies $\mathrm{Reg}_{\mathcal{A}}(N, T; s) \leq 2L_f N + 4\sigma\sqrt{2NT\iota_2} + 2sT$.

Proof. On $\mathcal{E}_1 \cap \mathcal{E}_2$, decompose the instantaneous regret at any round $t > T_0$:

$$f(x_{t,*}^\top \theta^*) - f(x_t^\top \theta^*) = \underbrace{\big[f(z_{t,*}) - \mu_{b_{t,*}}\big]}_{\text{(i)}} + \underbrace{\big[\mu_{b_{t,*}} - \mu_{b_t}\big]}_{\text{(ii)}} + \underbrace{\big[\mu_{b_t} - f(z_t)\big]}_{\text{(iii)}}.$$

By Lemma C.5, terms (i) and (iii) are each bounded by $L_{f'} w$ pathwise on $\mathcal{E}_1 \cap \mathcal{E}_2$. Summing over $t > T_0$:

$$\sum_{t=T_0+1}^{T} \big[\text{(i)} + \text{(iii)}\big] \leq 2L_{f'} w \cdot T.$$

For term (ii): the best available bin satisfies $\mu_{j_t^*} \geq \mu_{b_{t,*}}$ by definition, so $\mu_{b_{t,*}} - \mu_{b_t} \leq \mu_{j_t^*} - \mu_{j_t}$. With $s = L_{f'} w$, the observations satisfy $y_t = \mu_{j_t} + b_t + \eta_t$ with $|b_t| \leq s$ on the approximation event. The bias-robust guarantee assumed in Proposition E.4 therefore bounds the sum of term (ii) by $\mathrm{Reg}_{\mathcal{A}}(N, T; s)$ on the policy's guarantee event. This is a pathwise use of intersecting good events, not an assertion that conditioning on the full good event preserves centered noise.

Adding the Phase 1 cost $R_T^{(1)}$:

$$R_T \leq R_T^{(1)} + \mathrm{Reg}_{\mathcal{A}}(N, T; s) + 2L_{f'} T w.$$

Here $R_T^{(1)} \leq 2L_f T_0$ on $\mathcal{E}_2$, and $2L_{f'} T w < 2L_f T_0$ by Lemma C.4 when $T_0 < T$. Instantiating $\mathcal{A}$ with sleeping UCB (Lemma C.6) recovers Theorem 4.2, so the modular form does not change the bound. □

F Detailed Experiments

This appendix evaluates the empirical behavior of ZoomSIB-UCB along the dimensions that are most closely tied to the theory: growth with the horizon, sensitivity to the ambient dimension, adaptation to the unknown signal strength $|\mu^\star|$, the quality of the data-driven stopping rule, robustness to reward noise, and behavior on real context vectors. The purpose of these experiments is not to re-prove the worst-case regret theorem empirically, but to test whether the mechanisms used by the algorithm—direction estimation, certified discretization, and adaptive exploration—behave as intended at finite horizons.

Questions addressed by the experiments. The experiments are organized around the following questions. (i) On a fixed instance, how does the regret of ZoomSIB-UCB grow with T, and how does this compare with GSTOR, the $\theta^\star$-oracle discretization benchmark, and random exploration? (ii) How strongly do regret and exploration depend on the ambient dimension when the one-dimensional index distribution is held fixed? (iii) Does the stopping rule adapt to the unknown signal strength $|\mu^\star|$, exploring longer when the Stein signal becomes weaker without using $\mu^\star$ as an input? (iv) Is the exploration length selected by the adaptive rule close to the exploration length one would choose with hindsight? (v) How does the method react to increasing reward noise? (vi) Do the same qualitative conclusions persist when the contexts are taken from high-dimensional real datasets rather than generated from an exact Gaussian model?

F.1 Experimental protocol and implementations

Unless otherwise stated, we use $d = 10$, $K = 20$, $X \sim \mathcal{N}(0, I_d)$, $\theta^\star = e_1$, and Gaussian reward noise with standard deviation $\sigma = 0.5$. We consider three smooth links,

$$f_1(z) = \cos(z) + \sin(z), \qquad f_2(z) = ze^{-z^2}, \qquad f_3(z) = \sin(z) + 0.3z.$$

Note, the link $f_3(z) = \sin z + 0.3z$ is unbounded; its experiments therefore test finite-sample behavior outside the theorem's bounded-link assumption. All reported final experiments use 10 independent seeds (200–209). For a fixed seed, competing algorithms are evaluated with common random numbers, so they see the same context menus and reward-noise stream. Throughout this appendix, we use *regret* for the cumulative gap computed from the noise-free conditional means. The reward noise therefore affects learning but not the evaluation metric. Curves and tables report the mean regret, and error bars are one standard error of the mean.

ZoomSIB implementation. The literal constants in the high-probability analysis are intentionally conservative. At the horizons that can be simulated, directly substituting those constants often makes the analysed algorithm explore for essentially the entire horizon. We therefore use a calibrated practical instantiation that preserves the two-phase structure of Algorithm 1 but replaces worst-case constants by observable empirical quantities. The empirical stopping rule estimates an exploration cost c_{exp} and the loss caused by the current discretization $c_{\mathrm{bias}}(w_n)$ from the Phase-1 sample, and continues exploration while the estimated total exploration cost remains below the estimated discretization cost. No value or lower bound for $\mu^\star$ is supplied to the algorithm.

In preliminary calibration runs with $T = 50{,}000$, $K = 10$, $d \in \{5, 20\}$, and ten seeds per dimension, the empirical-Bernstein multiplier 0.15 retained directional-certificate coverage at every recorded estimator update. The smaller tested multiplier 0.05 did not. We therefore use 0.15 in the practical implementation. The Gaussian ℓ_2 width rule is a practical choice for the Gaussian experiments; the theorem uses the stated ℓ_∞–ℓ_1 bound.

At each stopping check, we resample synthetic arm menus from the Phase-1 projected samples. Fine and current-width bin estimates on these menus approximate, respectively, the best available reward and the reward obtained using the current grid. Their differences estimate the cost of uniform exploration and the loss from the current discretization; we stop when the estimated accumulated exploration cost exceeds the estimated remaining discretization loss.

Baselines. We compare against three references where appropriate. *Oracle-θ* is given the true direction $\theta^\star$ and therefore isolates the price of estimating the single-index direction; it still uses the same one-dimensional binning/UCB stage. *GSTOR* uses the two independent exploration blocks

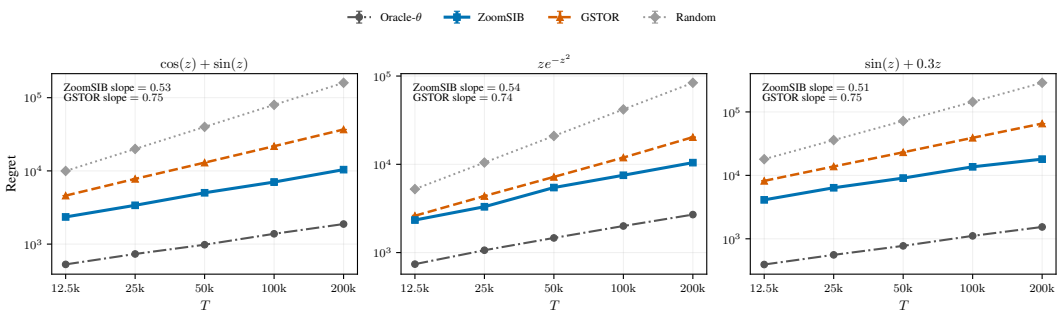

Figure 2: Fixed-instance horizon scaling on log–log axes. Only T changes across points; the fitted slopes are printed inside each panel.

Table 2: Horizon scaling. The left block reports regret at $T = 200{,}000$; the right block reports the empirical log–log slope fitted over the largest four horizons.

	Regret at $T = 200{,}000$				Empirical slope	
Link	Oracle-θ	**ZoomSIB**	GSTOR	Random	**ZoomSIB**	GSTOR
$\cos z + \sin z$	$1{,}879 \pm 26$	$10{,}401 \pm 480$	$36{,}775 \pm 101$	$159{,}769 \pm 108$	**0.534**	0.747
ze^{-z^2}	$2{,}694 \pm 25$	$10{,}446 \pm 449$	$20{,}312 \pm 140$	$83{,}877 \pm 37$	**0.544**	0.738
$\sin z + 0.3z$	$1{,}537 \pm 13$	$18{,}017 \pm 222$	$65{,}567 \pm 139$	$288{,}067 \pm 98$	**0.509**	0.750
Reference exponent	–	–	–	–	$2/3$	$3/4$

prescribed by the general-link procedure of Kang et al.: each block has length $T_1 = \lceil d^{3/8} T^{3/4} \rceil$ (clipped when necessary), the first block estimates the direction with a truncated Stein estimator, the second fits a one-dimensional uniform-kernel Nadaraya–Watson estimator with bandwidth $h = T_1^{-1/3}$, and the algorithm then commits greedily. *Random* selects an arm uniformly. We do not use ESTOR as a primary baseline here because it is designed for monotone links, whereas the chosen links are not assumed to satisfy monotonicity.

F.2 Fixed-instance horizon scaling

The first experiment asks how regret grows when the problem instance itself is held fixed. This distinction is important: changing a lower-bound bump construction with T changes both the horizon and the instance and therefore does not provide a clean empirical estimate of a T-exponent. Here only T varies, over

$$T \in \{12{,}500, 25{,}000, 50{,}000, 100{,}000, 200{,}000\},$$

with $d = 10$, $K = 20$, and $\sigma = 0.5$ fixed.

Figure 2 shows a consistent separation between the methods. Fitting a log–log line to the largest four horizons gives ZoomSIB slopes 0.534, 0.544, and 0.509 for f_1, f_2, f_3, respectively. The corresponding GSTOR slopes are 0.747, 0.738, and 0.750, while random exploration is essentially linear. These fixed-instance slopes should not be read as estimates of the minimax exponent: the three smooth test functions can be easier than the worst-case Lipschitz class. They are nevertheless consistent with the fact that the $\widetilde{O}(T^{2/3})$ theorem is an upper bound, and they show that the empirical growth of ZoomSIB is markedly slower than the $T^{3/4}$ behavior of GSTOR on these instances.

At $T = 200{,}000$, GSTOR incurs approximately $3.54\times$, $1.94\times$, and $3.64\times$ the regret of ZoomSIB on f_1, f_2, f_3, respectively. The oracle gap remains non-negligible, especially on the multimodal f_3, indicating that a substantial part of the remaining regret is still the price of learning the projection direction rather than the one-dimensional bandit problem alone.

F.3 Sensitivity to the ambient dimension

To isolate dimension from changes in the reward landscape, we fix $\theta^\star = e_1$ and $X \sim \mathcal{N}(0, I_d)$. Then $X^\top \theta^\star \sim \mathcal{N}(0, 1)$ for every d, so varying d changes the difficulty of estimating the direction but not

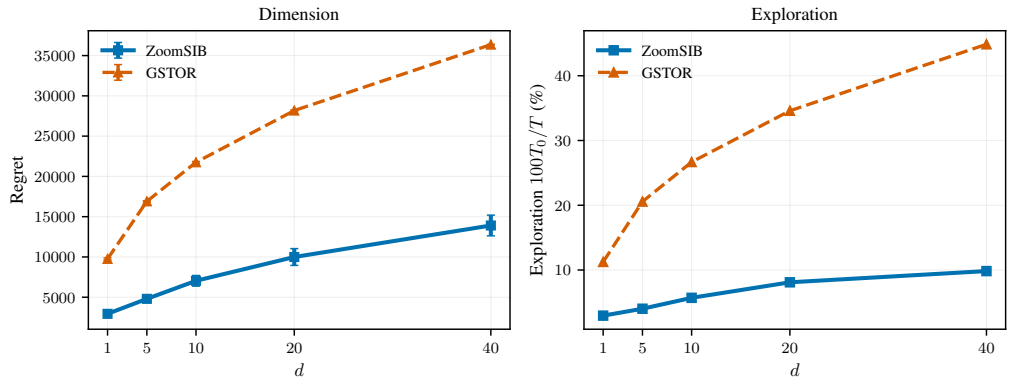

Figure 3: Dimension sensitivity at $T = 100{,}000$. Left: regret. Right: exploration percentage. The projected index distribution is fixed across d.

Table 3: Dimension sensitivity at $T = 100{,}000$. Exploration is reported as $100T_0/T$.

	Regret		Exploration (%)	
d	**ZoomSIB**	GSTOR	**ZoomSIB**	GSTOR
1	$2{,}940 \pm 227$	$9{,}761 \pm 133$	**2.95**	11.25
5	$4{,}800 \pm 174$	$16{,}904 \pm 77$	**4.02**	20.57
10	$7{,}039 \pm 661$	$21{,}751 \pm 106$	**5.70**	26.67
20	$9{,}994 \pm 1{,}043$	$28{,}174 \pm 79$	**8.10**	34.59
40	$13{,}906 \pm 1{,}287$	$36{,}359 \pm 41$	**9.84**	44.86

the one-dimensional distribution seen by the link. We use $T = 100{,}000$, $K = 20$, $f = f_1$, and

$$d \in \{1, 5, 10, 20, 40\}.$$

As expected, both algorithms become more expensive as d increases, but the adaptive procedure spends substantially fewer rounds in its estimation phase. Across $d = 1$ to 40, the fitted log–log slope of ZoomSIB's exploration length against d is approximately 0.34 under this practical Gaussian ℓ_2 width rule. This is an empirical property of the calibrated Gaussian implementation, not a claim about the worst-case d-dependence of Theorem 4.2.

F.4 Adaptation to unknown signal strength

The theorem allows arbitrary nonzero $\mu^\star$, but the learner is not given its value. We test this adaptivity using

$$f_\alpha(z) = \cos z + \alpha \sin z, \qquad \mu^\star = \alpha e^{-1/2}$$

under a standard Gaussian index. We fix $T = 200{,}000$ and vary

$$\alpha \in \{1, 0.5, 0.25, 0.125, 0.0625, 0.03\}.$$

The case $\alpha = 0$ is deliberately excluded because the current theorem assumes $\mu^\star \neq 0$.

The right panel of Figure 4 is the main point of the experiment: the algorithm reacts to a weaker Stein signal by spending more data on direction estimation, despite having no access to $\mu^\star$. This also exposes the finite-horizon cost of such adaptivity. For moderate and strong signals ($|\mu^\star| \gtrsim 0.15$ in this setup), ZoomSIB has lower regret than GSTOR; for the weakest signals, its longer exploration phase makes GSTOR better at this fixed horizon. We therefore view this experiment as evidence for adaptation, not as a claim of uniform finite-sample dominance over GSTOR.

F.5 Adaptive stopping versus hindsight tuning

The preceding experiment shows that T_0 changes substantially across instances. We next ask whether the selected value is close to what one would choose with hindsight. For each of the three links at

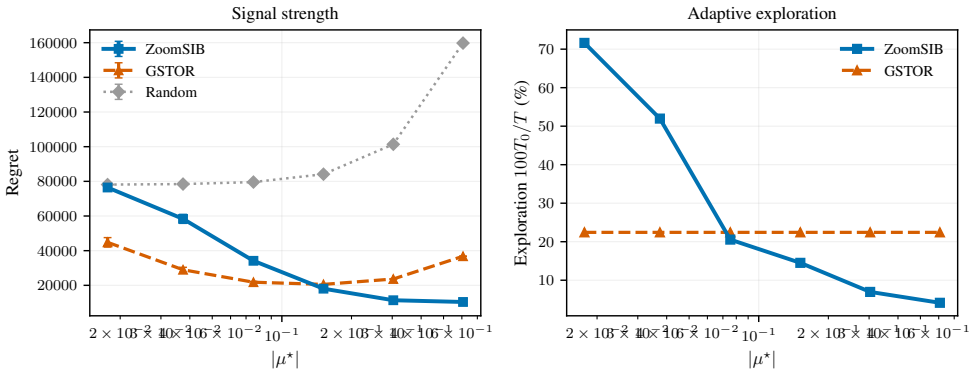

Figure 4: Dependence on the unknown Stein signal $|\mu^\star|$. Left: regret. Right: exploration fraction. ZoomSIB is never supplied with $\mu^\star$; its exploration time increases automatically as the signal weakens.

Table 4: Signal-strength sweep at $T = 200{,}000$. The displayed $|\mu^\star|$ values are Monte Carlo estimates; the exact value is $\alpha e^{-1/2}$. Green marks the smaller regret among ZoomSIB and GSTOR; the blue column emphasizes ZoomSIB's adaptive exploration budget.

	Regret			Exploration T_0/T	
$|\mu^\star|$	**ZoomSIB**	GSTOR	Random	**ZoomSIB**	GSTOR
0.6070	**10,402**	36,775	159,769	**0.041**	0.224
0.3032	**11,400**	23,647	101,377	**0.070**	0.224
0.1513	**18,073**	20,548	84,127	**0.145**	0.224
0.0754	**34,155**	**21,805**	79,580	**0.205**	0.224
0.0374	**58,349**	**28,984**	78,427	**0.520**	0.224
0.0176	**76,444**	**44,810**	78,131	**0.716**	0.224

$T = 200{,}000$, we rerun the same practical ZoomSIB estimator, grid construction, warm start, and UCB stage while forcing

$$T_0/T \in \{0.01, 0.02, 0.04, 0.08, 0.16, 0.32\}.$$

The best point on this grid is a hindsight oracle for the exploration duration only; it is not given $\theta^\star$.

The selected exploration duration lies very close to the best fixed choice in all three cases. Its regret is within 12% of the hindsight optimum on the tested grid, and within 3% on two of the three links. This experiment directly tests the practical purpose of the stopping rule: removing the need to hand-tune T_0 separately for each link.

F.6 Robustness to reward noise

We next vary the reward-noise standard deviation over $\sigma \in \{0.25, 0.5, 1.0\}$ with $T = 100{,}000$, $d = 10$, $K = 20$, and $f = f_1$ fixed.

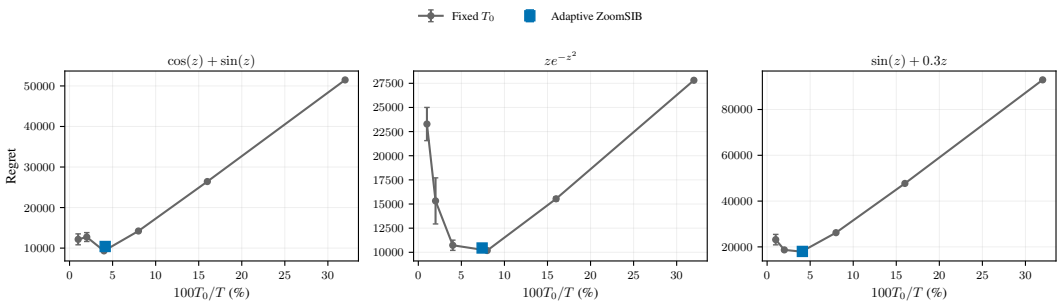

Figure 5: Adaptive stopping versus a hindsight grid of fixed exploration budgets. The square is the adaptive ZoomSIB run; the curve shows otherwise identical runs with forced T_0.

Table 5: Adaptive stopping versus hindsight tuning. Ratio = adaptive regret / best fixed-T_0 regret; values close to one are better.

Link	best fixed T_0/T	**adaptive** T_0/T	best fixed regret	**adaptive regret**	ratio
$\cos z + \sin z$	4.0%	**4.14%**	9,317	**10,401**	1.116
ze^{-z^2}	8.0%	**7.41%**	10,186	**10,446**	**1.026**
$\sin z + 0.3z$	4.0%	**4.10%**	17,921	**18,017**	**1.005**

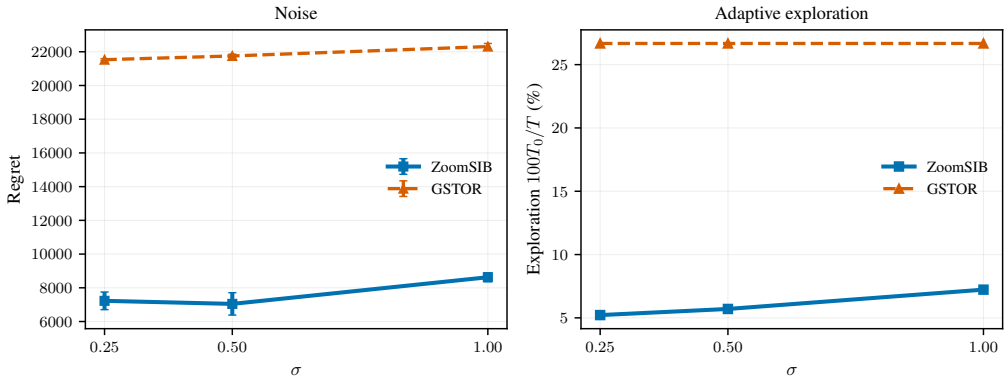

Figure 6: Noise robustness. ZoomSIB increases exploration as σ grows; GSTOR's prescribed exploration percentage is unchanged.

Table 6: Noise robustness at $T = 100{,}000$.

	Regret		Exploration T_0/T	
σ	**ZoomSIB**	GSTOR	**ZoomSIB**	GSTOR
0.25	$7{,}227 \pm 523$	$21{,}530 \pm 53$	**0.052**	0.267
0.50	$7{,}046 \pm 664$	$21{,}751 \pm 106$	**0.057**	0.267
1.00	$8{,}626 \pm 264$	$22{,}309 \pm 184$	**0.072**	0.267

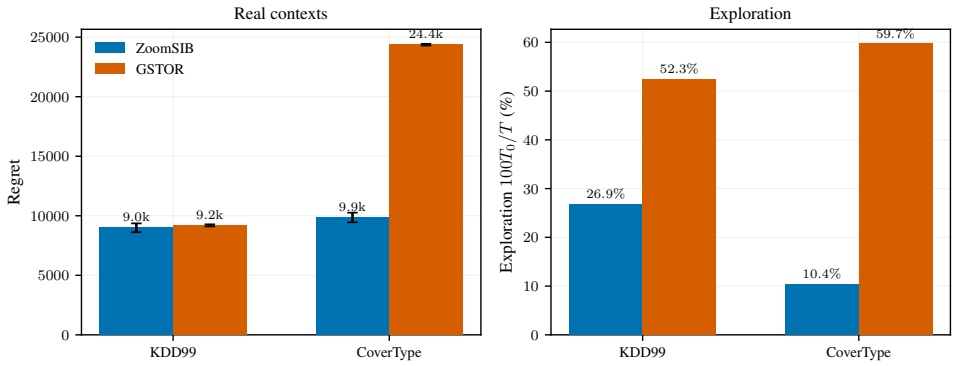

Figure 7: Semi-synthetic experiments with real context vectors. Left: regret. Right: exploration fraction. Rewards are generated from an exact single-index link, while the context distribution comes from the real datasets.

Table 7: Real-context semi-synthetic results at $T = 50{,}000$.

Dataset	d	K	**ZoomSIB regret**	GSTOR regret	**ZoomSIB** T_0/T	GSTOR T_0/T
KDD99	38	20	$8{,}991 \pm 370$	$9{,}194 \pm 85$	**0.269**	0.523
CoverType	54	20	$9{,}858 \pm 404$	$24{,}369 \pm 71$	**0.104**	0.597

Increasing the noise from 0.25 to 1 raises ZoomSIB's exploration fraction from 5.2% to 7.2% and increases its regret moderately. Across the tested noise levels, GSTOR has between $2.59\times$ and $3.09\times$ the regret of ZoomSIB.

F.7 Real-context semi-synthetic experiments

Finally, we test the algorithms with contexts drawn from two standard high-dimensional datasets, KDD Cup 99 and Forest CoverType. These are *semi-synthetic* experiments: the context vectors are real, while rewards are generated from the single-index model so that the ground-truth regret remains well defined. We remove constant columns, standardize each feature, whiten the resulting design, and at every round sample $K = 20$ real rows as the available menu. After preprocessing, the dimensions are $d = 38$ for KDD99 and $d = 54$ for CoverType. We use $T = 50{,}000$, $\sigma = 0.5$, $\theta^\star = e_1$, and $f(z) = \cos z + \sin z$.

Both methods use the fitted-Gaussian score $S(x) = x$ after whitening. The empirical feature distributions are not exactly Gaussian, so this score is a practical approximation; consequently this experiment is intended as a robustness/transfer study rather than a theorem-faithful test of the known-score assumption.

On KDD99 the two methods have similar final regret, with ZoomSIB reducing the exploration fraction by roughly one half. On CoverType, ZoomSIB attains about $2.47\times$ lower regret while exploring for only 10.4% of the horizon compared with 59.7% for GSTOR. The different behavior across the two datasets is useful: the real-context experiment does not merely reproduce a synthetic ordering, but shows that the benefit depends on how quickly the direction can be estimated under the approximate score model.

F.8 Summary of empirical findings

The experiments support four conclusions. First, on fixed smooth instances, ZoomSIB exhibits substantially slower regret growth with T than GSTOR and random exploration, while the true-direction oracle shows the remaining cost of direction estimation. Second, the stopping mechanism reacts in the intended direction to all three statistical difficulty parameters studied here: it explores longer as d increases, as $|\mu^\star|$ decreases, and as σ increases. Third, the resulting exploration duration is close to the hindsight-best fixed duration on the tested links, which suggests that the data-driven rule removes a meaningful tuning burden. Finally, the weak-signal and real-context experiments delimit the claim: ZoomSIB is not uniformly better at every finite horizon, and the semi-synthetic

real-data setting uses an approximate fitted-Gaussian score. These qualifications are consistent with the role of the experiments as finite-sample evidence for the mechanisms behind the regret guarantee rather than as a substitute for the worst-case theory.