Submitted to the Annals of Statistics

NONPARAMETRIC BANDITS WITH SINGLE-INDEX REWARDS: OPTIMALITY AND ADAPTIVITY

BY WANTENG MA [1,a] AND T. TONY CAI[2,b]

[1]*Department of Statistics and Data Science, University of Pennsylvania ,* [a]*wanteng@wharton.upenn.edu*

[2]*Department of Statistics and Data Science, University of Pennsylvania ,* [b]*tcai@wharton.upenn.edu*

Contextual bandits are a central framework for sequential decision-making, with applications ranging from recommendation systems to clinical trials. While nonparametric methods can flexibly model complex reward structures, they suffer from the curse of dimensionality. We address this challenge using a single-index model, which projects high-dimensional covariates onto a one-dimensional subspace while preserving nonparametric flexibility.

We first develop a nonasymptotic theory for offline single-index regression for each arm, combining maximum rank correlation for index estimation with local polynomial regression. Building on this foundation, we propose a single-index bandit algorithm and establish its convergence rate. We further derive a matching lower bound, showing that the algorithm achieves minimax-optimal regret independent of the ambient dimension d, thereby overcoming the curse of dimensionality.

We also establish an impossibility result for adaptation: without additional assumptions, no policy can adapt to unknown smoothness levels. Under a standard self-similarity condition, however, we construct a policy that remains minimax-optimal while automatically adapting to the unknown smoothness. Finally, as the dimension d increases, our algorithm continues to achieve minimax-optimal regret, revealing a phase transition that characterizes the fundamental limits of single-index bandit learning.

1. Introduction. Online decision-making has become central to modern data-driven systems, underpinning applications ranging from recommendation and advertising to clinical trials and resource allocation. A key objective in these settings is reward maximization under uncertainty, where an agent learns optimal decisions through feedback over time. The *bandit problem* provides a foundational framework for this challenge and has been extensively studied across multiple variants, including multi-armed bandits [47, 62], linear contextual bandits [1, 20], and adversarial bandits [6, 12].

Among these frameworks, *contextual bandits* stand out for their ability to incorporate side information, enabling personalized and adaptive decision-making. They have achieved remarkable success in applications such as recommendation systems [54], adaptive clinical trials [23], online advertising [18], and intelligent tutoring systems [65]. While many studies assume linear reward structures for tractability [7, 29], real-world feedback is often nonlinear. For instance, in the Netflix artwork selection problem [17, 30], a user's willingness to play a movie title follows a non-linear function of the user's preference over two different artworks that portray the title. Linear models fail to capture such complexity, motivating the study of *nonparametric contextual bandits*, which allow greater modeling flexibility. However, they suffer from the *curse of dimensionality*—learning rates deteriorate quickly as the context dimension increases. For the nonparametric bandit algorithms studied in the literature [30,

MSC2020 subject classifications: Primary 62G08; secondary 62L12.

Keywords and phrases: single-index model, contextual bandits, curse of dimensionality, minimax optimality, adaptivity.

62, 66], the sample complexity required to learn a good policy grows exponentially with the dimension d, making these methods challenging to deploy in practice.

Motivated by these challenges, we consider *single-index bandits*, where each reward function depends on covariates only through an unknown one-dimensional projection—an *index*—while retaining nonparametric flexibility in the link function. By reducing the effective dimension to one, single-index contextual bandits preserve the expressive power of nonparametric models while alleviating the curse of dimensionality, enabling efficient learning in complex, high-dimensional decision-making problems. The *single-index model* has been extensively studied in nonparametric regression and widely applied in practice [36, 41, 70]. We show that under single-index contextual bandits, it is possible to design an efficient algorithm that achieves minimax-optimal regret, matching the one-dimensional nonparametric bandit rate. Remarkably, this optimal rate is independent of the ambient dimension d, thereby achieving effective dimension reduction and overcoming a key barrier in high-dimensional online learning.

We briefly describe the formulation of the multi-armed contextual bandit problem as follows. At each time step $t = 1, \ldots, n$, a decision maker is presented with a covariate vector (or context) $X_t \in \mathbb{R}^d$ and selects one arm from a K-armed bandit $\{Y_t^{(k)}\}_{k=1}^K$ ($K \geq 2$). The decision maker then observes a reward $Y_t = Y_t^{(\pi_t)}$ corresponding to the chosen arm $\pi_t \in [K]$. The term "contextual" indicates that the decision at time t is made based on the contextual information X_t, which is observed prior to choosing the arm. Suppose further that the d-dimensional contexts X_t, together with the rewards for the K arms $\{Y_t^{(k)}\}_{k=1}^K$, are independent and identically distributed across time steps $t = 1, 2, \ldots, n$.

$$\left(X_t, Y_t^{(1)}, Y_t^{(2)}, \ldots, Y_t^{(K)}\right) \sim P_{X, \{Y^{(k)}\}_{k=1}^K}. \tag{1}$$

Define the (expected) reward function for each arm as

$$g_k(X_t) := \mathbb{E}[Y_t^{(k)} | X_t].$$

Our goal is to construct a policy $\pi = \{\pi_t\}_{t=1}^n$ that minimize the cumulative expected regret with respect to the optimal policy $\pi^\star(X_t) = \arg\max_{k \in K}\{g_k(X_t)\}$, i.e.,

$$\mathsf{Reg}(\pi) = \mathbb{E}\left[\sum_{t=1}^T \left(Y_t^{(\pi^\star(X_t))} - Y_t^{(\pi_t)}\right)\right] = \mathbb{E}\left[\sum_{t=1}^T \left(g_{\pi^\star(X_t)}(X_t) - g_{\pi_t}(X_t)\right)\right]. \tag{2}$$

In this paper, we focus on the setting where the reward functions $g_k(X_t)$ follow a single-index model (with the detailed definition provided in Section 2) for dimension reduction. We study the minimax regret for single-index bandits and develop rate-optimal online algorithms to achieve this bound.

1.1. *Main contribution.*

Our contribution is three-fold:

- *Minimax regret for single-index bandits.* To develop online algorithms for single-index bandits, we first propose an offline single-index regression method based on an index plug-in approach and establish a sharp non-asymptotic convergence rate. We consider a class of monotone link functions with Hölder smoothness level β, and employ maximum-rank correlation [25, 32] for index estimation. We show that the resulting estimator attains ℓ_2 accuracy of order $\widetilde{O}\left(\sqrt{d/n}\right)$ with overwhelming probability. The estimated index is then plugged into a one-dimensional local polynomial regression for reward estimation. Building on this regression framework, we design a batched online bandit algorithm featuring a two-step arm-elimination procedure: arms are first preselected based on estimates from the

previous epoch, and then further refined using the most recently collected data. We prove that our algorithm achieves a regret bound of order $\widetilde{O}\left(n^{1-\frac{(\alpha+1)\beta}{2\beta+1}}\right)$, which is minimax optimal up to logarithmic factors, where α denotes the margin parameter and n is the sample size. In the fixed-dimension setting, this rate is independent of d, demonstrating effective dimension reduction.

- *Adaptivity to unknown smoothness.* We next investigate adaptation when the smoothness level β is unknown. We show that, analogous to the classical nonparametric bandit setting, adaptation to an unknown β is impossible without additional structural assumptions [30]. However, under a standard self-similarity condition, adaptation becomes feasible by first estimating β prior to regret minimization. To this end, we propose a link-smoothness estimation procedure with index plug-in, which yields an "under-smoothed" estimator of β with accuracy $O\left(\frac{\log\log n}{\log n}\right)$. This estimation step enables our algorithm to achieve minimax-optimal regret even when β is unknown.
- *Single-index bandits in increasing dimensions.* Finally, we study the minimax regret when the dimension d grows with the sample size n. We show that, in the increasing-dimension regime, the minimax-optimal regret is of order $n\left(\left(\frac{1}{n}\right)^{\frac{(1+\alpha)\beta}{2\beta+1}}+\left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}}\right)$, which reveals a two-phase phenomenon: a nonparametric phase $n^{1-\frac{(\alpha+1)\beta}{2\beta+1}}$, corresponding to the optimal regret of one-dimensional nonparametric bandits, and a parametric phase $n\left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}}$, corresponding to the optimal regret of linear bandits [8, 59]. Moreover, we show that our proposed algorithm continues to achieve minimax-optimal regret (up to logarithmic factors) as d increases, thereby bridging the two phases.

1.2. *Related literature.* Bandit algorithms have been extensively studied in recent years; a comprehensive overview is provided in [49]. Here, we primarily review linear and nonparametric bandits, which are most closely related to our setting. Linear bandits represent arguably the simplest form of contextual bandits, where the reward function depends linearly on observed covariates [20]. For linear bandits, an optimal regret rate of order $O(\sqrt{dn})$ has been established and analyzed in [1, 7, 11, 34, 59], with extensions to generalized linear bandits under fixed link functions in [55]. However, due to the restrictive structure of linear reward models, techniques developed for linear bandits—such as upper confidence regions [44] and Thompson sampling [3]—are largely problem-specific and cannot be directly applied to bandit settings with more general, nonlinear reward functions.

To address such nonlinear settings, *nonparametric bandits*, pioneered by [66] and [62], have attracted increasing attention. Specifically, [66] introduced a UCB-type algorithm based on bin partitioning for two-armed bandits, while [62] proposed ABSE, a successive-arm-elimination method that achieves minimax-optimal regret for $\beta \leq 1$. An extension of ABSE to transfer learning was investigated in [13]. For smooth nonparametric bandits with $\beta \geq 1$, [39] designed an algorithm that utilizes information within bins to obtain optimal smooth regret rates. However, all of these approaches suffer from regret bounds that scale exponentially with the dimension d, typically of order $O\left((\mathrm{poly}(\beta))^d\right)$ due to bin partitioning. This exponential dependence underscores the importance of effective dimension reduction in high-dimensional settings.

The single-index model is a widely used dimension-reduction framework with applications in econometrics, biometrics, and medical science; see [36, 41, 70, 71]. Its dimension-reduction effect can be summarized as follows. In classical nonparametric regression, the minimax-optimal MISE rate is $O\left(n^{-\frac{2\beta}{2\beta+d}}\right)$, which deteriorates rapidly as d increases. However, when the regression function follows a single-index structure, the optimal rate improves

to $O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$ [26, 31], resolving the conjecture of [68] on achieving dimension reduction. A variety of estimation procedures for single-index models have since been developed, including [4, 46, 50, 61]. Recent reviews of estimation methods in single-index regression can be found in [48] and [72], as well as in the computational and learning-theoretic overviews of [21] and [9]. Despite these advances, single-index regression with adaptively sampled data has remained largely unexplored in the context of bandit problems.

Adaptation is another important topic in nonparametric bandits. As recognized in prior work [16, 30, 56], adaptation to unknown smoothness levels of reward functions is generally impossible without additional assumptions. This phenomenon is closely related—though not identical—to the impossibility of adaptive confidence interval construction in nonparametric regression [14, 57]. Despite these negative results, it has been shown in [63] and [27] that adaptive inference becomes possible under the self-similarity condition, a structural assumption that constrains the behavior of the unknown function. This idea has since been extended to nonparametric bandits [13, 64], where Lepski's method [51] is used to adapt to the smoothness level prior to decision-making. However, it remains an open question whether such adaptation remains feasible when the reward function additionally satisfies a single-index structure.

1.3. *Organization.* The rest of the paper is organized as follows. Section 2 introduces a formal mathematical formulation of the nonparametric bandit problem with single-index rewards. Section 3 presents our (offline) single-index regression method and establishes sharp non-asymptotic error bounds. Building on these results, Section 4 develops our online algorithm for single-index bandits. Section 5 provides the theoretical analysis of minimax-optimal regret, showing that our algorithm achieves the optimal rate. Section 6 investigates adaptation to unknown smoothness levels. Section 7 extends the problem to increasing dimensions and examines the resulting phase transition in minimax optimal rates. Finally, Section 8 presents numerical experiments that corroborate our theoretical findings.

1.4. *Notation.* The following notation will be used throughout the paper. For a function f, we use $f^{(i)}$ to denote its i-th derivative. For a random variable Y, $Y^{(k)}$ represents the variable corresponding to the k-th arm. We also use $g_{(i)}(x)$ to denote the i-th largest value among the functions $\{g_k\}_{k\in[K]}$ evaluated at x, with the precise definition provided in the next section. For parameters associated with each arm that share a common notation, such as v_k, we use $v_{k,i}$ to represent the i-th entry of the parameter in arm k. For vectors without an arm specification, such as x, x_i denotes the i-th entry of x. We denote $\lfloor\beta\rfloor$ as the largest integer that is strictly smaller than any positive β. The notation $\|\cdot\|_2$ represents the ℓ_2 norm. For two scalars a, b, we define $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_2 ball is denoted by $\mathbb{B}_2(x, r) = \left\{x' \in \mathbb{R}^d \mid \|x' - x\|_2 \le r\right\}$. For any positive integer K, we let $[K] = \{1, 2, \ldots, K\}$ denote the corresponding index set. For a set $\mathcal{X} \subseteq \mathbb{R}^d$, its projection onto a direction v is defined as $v \circ \mathcal{X} = v^\top x \mid x \in \mathcal{X}$. For a scalar x and $r > 0$, the r-neighborhood of x is denoted by $U(x, r) = [x - r, x + r]$. For an interval or general set $I \subseteq \mathbb{R}$, the r-neighborhood of I is defined as $U(I, r) = \{x \mid \mathrm{dist}(x, I) \le r\}$, where $\mathrm{dist}(x, I)$ represents the distance from x to the closest element in I. We use the standard asymptotic notation $O(\cdot)$ to describe the upper bound of a function's growth rate, which shares the same meaning as asymptotic inequality $\lesssim$. We also use the $\widetilde{O}(\cdot)$ notation to suppress additional polylogarithmic factors. Formally, $f(n) = \widetilde{O}(g(n))$ means $f(n) = O(g(n) \cdot \mathrm{poly}(\log n))$, where $\mathrm{poly}(\log n)$ denotes a polynomial in $\log n$.

2. Nonparametric Bandits with Single-Index Rewards. We introduce in this section a formal framework for the nonparametric bandit problem with single-index rewards. Given the unknown bandit distribution in (1), our goal is to minimize the cumulative regret defined in (2). We denote the marginal distribution of X_t by P_X, with support $\mathcal{X}$, and assume that it satisfies certain regularity conditions (to be specified later). As is standard in bandit problems, once arm $k \in [K]$ is selected at time t, only $Y_t^{(k)}$ is observed (i.e., $Y_t = Y_t^{(k)}$), while the rewards for all other arms at time t remain unobserved. Under this partial-feedback structure, we aim to design a policy $\pi = \{\pi_t\}_{t=1}^n$ for arm selection, where each π_t is an *admissible* randomized decision rule $\pi_t : \mathcal{X} \to [K]$ that chooses which arm to pull at time t based on X_t, and depends only on observations strictly prior to time t. Throughout our analysis, we assume that the number of arms remains at a constant level, i.e., $K = O(1)$.

Without imposing additional structural assumptions, minimizing (2) is challenging due to the large functional space. In this paper, we assume that for each arm $k \in [K]$, the expected reward follows a single-index model [40, 58], i.e.,

$$g_k(X_t) = \mathbb{E}, [Y_t^{(k)} \mid X_t] = f_k(v_k^\top X_t), \tag{3}$$

where the *link* function $f_k : \mathbb{R} \to [0, 1]$ and the *index* vector $v_k \in \mathcal{V} \subseteq \mathbb{R}^d$ are both unknown to the decision-maker. We assume that $\mathcal{X}$ and the parameter space $\mathcal{V}$ are compact with bounded ℓ_2 norms.

However, due to the flexible choices of f_k and v_k, the model (3) is not identifiable without additional constraints. To ensure identifiability of both the link functions and index vectors, we impose the normalization $v_{k,1} = 1$ for each arm k. For further discussion on identifiability in single-index models, see, e.g., [41, 58], and Chapter 2 of [37].

For each arm $k \in [K]$, we pose the reward in a regression form:

$$Y_t^{(k)} = g_k(X_t) + \xi_t^{(k)} = f_k(v_k^\top X_t) + \xi_t^{(k)}, \tag{4}$$

where $\{\xi_t^{(k)}\}$ are noises with $\mathbb{E}\xi_t^{(k)} = 0$, and might be X_t-dependent. In this paper, we confine $\xi_t^{(k)}$'s dependency on X_t in the form that $\xi_t^{(k)}$ can be expressed as $f_k(v_k^\top X_t) + \xi_t^{(k)} = D(f_k(v_k^\top X_t) + u_t^{(k)})$ for some function $D(z)$, where $u_t^{(k)}$ is a random variable that is independent of X_t, and $D(z)$ is a non-degenerate monotonic function. Typical examples include Bernoulli response regression: $Y_t^{(k)} \mid X_t \sim \mathrm{Ber}(f_k(v_k^\top X_t))$, which is equivalent to taking $D(z) = \mathbb{I}\{z > \frac{1}{2}\}$, and $u_t^{(k)} \sim \mathrm{Unif}(-\frac{1}{2}, \frac{1}{2})$, and general i.i.d. noise regression when $\xi_t^{(k)}$ is independent of X_t by taking $D(z) = z$. Also, we assume that $\xi_t^{(k)}$ is sub-Gaussian such that conditional on X_t, $\mathbb{E}\left[\exp\left((\xi_t^{(k)})^2/C\right) \mid X_t\right] \le 2$, for some constant $C = O(1)$. To characterize f_k, we introduce the standard Hölder smoothness class as follows:

DEFINITION 2.1 (Smoothness class). We say that the link function f belongs to the Hölder class $\mathcal{H}_0(\beta, L)$ for $\beta > 0$ and $L > 0$, which is the collection of functions $f\colon \mathbb{R} \to [0, 1]$ that are $\lfloor\beta\rfloor$ times continuously differentiable and for any $x, x' \in \mathcal{X}$, satisfy the following definition:

$$\left|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(x')\right| \le L\left|x - x'\right|^{\beta - \lfloor\beta\rfloor}.$$

Let $\mathcal{A}_+$ be the non-degenerate monotonic function class. Define our function class of interest as

$$\mathcal{H}(\beta, L) = \begin{cases} \mathcal{H}_0(\beta, L), & \beta < 1, \\ \mathcal{H}_0(\beta, L) \cap \mathcal{H}_0(1, L), & \beta \ge 1. \end{cases} \quad \text{and } \mathcal{H}_+(\beta, L) = \mathcal{H}(\beta, L) \cap \mathcal{A}_+. \tag{5}$$

Here, $\mathcal{H}(\beta, L)$ is defined based on $\mathcal{H}_0(\beta, L)$ for minimax purpose, and $\mathcal{H}_+(\beta, L)$ is the monotonic function set of $\mathcal{H}_0(\beta, L)$. The Hölder smooth class for multivariate functions can be similarly defined using Taylor polynomials. See, e.g, [5]. To ease the notation, in what follows, we shall write the optimal arm and sub-optimal arm as

$$
\begin{aligned}
g_{(1)}(x) &= \max_{k\in[K]} \{g_k(x)\}, \\
g_{(2)}(x) &= \max_{k\in[K]} \left\{g_k(x) \mid g_k(x) < g_{(1)}(x)\right\},
\end{aligned}
\tag{6}
$$

or $g_{(2)}(x) = g_{(1)}(x)$ if $\left\{k \mid g_k(x) < g_{(1)}(x)\right\} = \emptyset$.

3. A Non-Asymptotic Theory for Single-Index Regression. In this section, we study the offline regression setting in which observations come from a single arm. In much of the bandit literature [12, 20, 39, 62], effective algorithms rely on estimating unobserved rewards and deriving sharp non-asymptotic confidence bounds. To develop algorithms for single-index bandits, it is therefore essential to establish uniform estimation guarantees for regression under single-index feedback.

Let f_0 and v_0 denote the target link function and index, and consider the model in (4):

$$
Y_t = g_0(X_t) + \xi_t = f_0(v_0^\top X_t) + \xi_t, \tag{7}
$$

where f_0, v_0, and ξ_t satisfy the assumptions in Section 2. We first briefly review existing methods for single-index regression and discuss why they are insufficient for our bandit setting.

The classical optimality result for single-index regression was established by [26] using an aggregation approach. They showed that the minimax-optimal MISE rate $O\left(n^{-\frac{2\beta}{2\beta+1}}\right)$ can be achieved via empirical risk minimization. However, this aggregation-based guarantee does not extend to pointwise or uniform error bounds. A similar limitation applies to the temperature method used to derive PAC-Bayesian bounds in [4]. To obtain pointwise and uniform estimation guarantees, a standard strategy is to consistently estimate the index v_0, followed by nonparametric regression using the resulting plug-in estimator. Two major approaches have been developed for estimating v_0: inverse regression and forward regression.

Inverse regression. Inverse regression exploits the idea that, when the distribution of X_t exhibits certain structural properties (e.g., elliptical symmetry), the conditional distribution $X_t \mid Y_t$ reveals information about the subspace $\mathrm{span}(v_0)$. Classical methods include sliced inverse regression (SIR) [53] and the sliced average variance estimator (SAVE) [22]. A key assumption underlying inverse regression is the linear conditional mean (LCM) condition, which requires $\mathbb{E}[X_t \mid v_0^\top X_t]$ to be linear in $v_0^\top X_t$. Unfortunately, this condition does not hold in contextual bandit settings, where the selection of arms alters the conditional distribution of X_t in a manner that breaks the geometric properties of P_X necessary for the LCM condition.

Forward regression. Forward regression encompasses several distinct methodologies. A prominent class is the Stein's Lemma-based methods [10], which exploit the fact that, under a Gaussian distribution of P_X, $\mathbb{E}[Y_t \cdot X_t] = \mathbb{E}\left[f'\left(v_0^\top X_t\right)\right] \cdot v_0$, thereby enabling the estimation of v_0. However, these methods require Gaussian covariates and differentiable link functions with smoothness parameter $\beta \geq 1$. In a similar spirit, many approaches estimate the gradient of $g(x)$, which contains information about v_0—for example, ADE [38] and rMAVE [71]. Nevertheless, these gradient-based approaches rely on well-behaved gradients $\nabla g(x)$ and typically require higher smoothness, such as $\beta \geq 2$ or even $\beta \geq 3$ for higher-order methods.

In contextual bandit problems, these existing approaches are unsuitable for online estimation: they either rely on assumptions (e.g., the LCM condition or Gaussian covariates) that fail under bandit sampling, or they apply only to restricted smoothness regimes. Motivated by these limitations, we instead study a rank-based estimator, which imposes fewer structural requirements and is well-suited to our online setting.

3.1. *Maximum rank correlation and index plug-in.* Rank-type estimators are a class of M-estimators that learn latent index structures by maximizing or minimizing rank-based objective functions. A canonical example is Han's maximum rank correlation (MRC) estimator [32], which recovers a single-index model under a monotonic link function. For brevity, we write $Z_t = (X_t, Y_t)$ and denote the regression sample by $\mathcal{S} = (X_t, Y_t) : t \in [n] = Z_t : t \in [n]$. We now describe Han's MRC estimator [32]. Define the kernel of the associated U-statistic as $J(Z_1, Z_2, v) = \mathbb{I}(Y_1 > Y_2)\,\mathbb{I}\left(X_1^\top v > X_2^\top v\right)$. The rank-sum objective is then given by

(8)
$$\Gamma_{\mathcal{S}}(v) = \frac{1}{|\mathcal{S}|\,(|\mathcal{S}|-1)} \sum_{Z_i \neq Z_j \in \mathcal{S}} J(Z_1, Z_2, v) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}(Y_i > Y_j)\,\mathbb{I}\left(X_i^\top v > X_j^\top v\right).$$

The MRC estimator is obtained by maximizing the empirical rank correlation:

(9)
$$\widehat{v}^{\mathsf{H}} = \operatorname*{argmax}_{u : u_1 = 1} \Gamma_{\mathcal{S}}(u).$$

Without loss of generality, we focus on monotonically increasing functions in $\mathcal{H}_+(\beta, L)$. The case of monotonically decreasing link functions can be handled analogously by replacing the kernel with $J'(Z_1, Z_2, v) = \mathbb{I}(Y_1 < Y_2)\,\mathbb{I}\left(X_1^\top v > X_2^\top v\right)$.

Our single-index regression procedure using MRC is summarized in Algorithm 1, where the estimator $\widehat{v}^{\mathsf{H}}$ is used as input for local polynomial estimation (LPE).

Algorithm 1 Single-Index Regression By MRC

Require: $\mathcal{S} = \{(X_t, Y_t) : t \in [n]\}$, where $n = |\mathcal{S}|$, smoothness level β, tuning parameter C_{H} .

1: Split $\mathcal{S}$ evenly into $\mathcal{S}_1, \mathcal{S}_2$

2: Compute $\widehat{v}^{\mathsf{H}} = \operatorname*{argmax}_{u : u_1 = 1} \Gamma_{\mathcal{S}_1}(u)$ following (8), (9)

3: For each $(X_t, Y_t) \in \mathcal{S}_2$, project X in onto direction $\widehat{v}^{\mathsf{H}}$ and get new $\mathcal{S}_2' = \{((\widehat{v}^{\mathsf{H}})^\top X_t, Y_t) : (X_t, Y_t) \in \mathcal{S}_2\}$.

4: Run local polynomial estimation on $\mathcal{S}_2'$ with level $\lfloor \beta \rfloor$, bandwidth $h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}} \vee C_{\mathsf{H}} \left(\frac{d+\log^2 n}{n}\right)^{\frac{1}{2}}$ and uniform kernel $\mathbb{I}\{|\cdot| \leq 1\}$, denote the regression function as $\widehat{f}(\cdot)$.

5: Plug $\widehat{v}^{\mathsf{H}}$ into $\widehat{f}(\cdot)$ and get $\widehat{g}(x) = \widehat{f}((\widehat{v}^{\mathsf{H}})^\top x)$.

6: (Optionally) repeat 2-5 by switching $\mathcal{S}_1$ and $\mathcal{S}_2$ and get $\widehat{g}'$. Let $\widehat{g} \leftarrow \frac{1}{2}(\widehat{g} + \widehat{g}')$ for cross-fitting.

7: **return** Regression function $\widehat{g}(x)$

Instead of relying on assumptions that are impractical for contextual bandits—such as the LCM condition—Algorithm 1 requires only that the link function be monotonic, an assumption commonly adopted in many reward models [33, 43]. Moreover, monotonic reward structures arise naturally in a wide range of real-world applications, including the Netflix artwork selection problem discussed earlier [17, 30], as well as more general binary-outcome decision-making modeled using multinomial logit rewards [60]. Examples include product and brand selection [24], transportation choice modeling [73], among many others.

3.2. *Theoretical properties of MRC and index plug-in.* Despite the long history of MRC estimators in the literature—see, for example, [32], [67], [2], [33], and [25]—existing theories do not provide sharp non-asymptotic bounds for regression. The most recent analysis in [25] establishes only asymptotic convergence rates, which are insufficient for bandit regret control. Motivated by this gap, we develop in this section a new non-asymptotic theory for MRC that yields sharp finite-sample error bounds.

We write the covariate X as $X = [X_1, \widetilde{X}^\top]^\top$, where $\widetilde{X} \in \mathbb{R}^{d-1}$ contains the last $d-1$ components of X. Correspondingly, we write $x = [x_1, \widetilde{x}^\top]^\top$. Let $\mu_0(\cdot \mid \widetilde{X} = \widetilde{x}, Y^{(k)} = y)$

denote the conditional density of X_1 given $(\widetilde{X}, Y^{(k)})$, and let $\mu_0(\cdot)$ denote the marginal density of $v_0^\top X$. We introduce the following definition:

DEFINITION 3.1 ((c_0, r_0, r_1)-regularity condition). A set $A \subseteq \mathbb{R}$ is called (c_0, r_0, r_1)-*regular* if

$$\mathsf{Leb}(A \cap U(x, r)) \geq c_0 \cdot 2r, \qquad \forall, r_1 < r \leq r_0, \ \forall, x \in A. \tag{10}$$

When $r_1 = 0$, we refer to this simply as the (c_0, r_0)-regularity condition. Following [5], we say a distribution satisfies the *strong density* condition if its support is a (c_0, r_0)-regular set for some constants c_0, r_0, and its density is bounded away from zero and infinity on that support. The assumptions required for our analysis are provided in Assumption 3.2.

ASSUMPTION 3.2 (Conditions for estimation). We assume that:

3.2.1 The conditional density $\mu_0(x_1 \mid \widetilde{x}, y)$ exists and has uniformly bounded derivatives for x_1 up to order three, i.e., $\left|\mu_0^{(i)}(x_1 \mid \widetilde{x}, y)\right| \leq C$, for any x, y in the support and $i = 1, 2, 3$ almost everywhere.

3.2.2 $v_0^\top X$ is with (c_0, r_0)-regular support $I_0 := v_0 \circ \mathcal{X} = \left\{v_0^\top x \mid x \in \mathcal{X}\right\}$, and its marginal density $\mu_0(\cdot)$ satisfies $\underline{\mu} \leq \mu_0(\cdot) \leq \overline{\mu}$ on its support I_0.

DEFINITION 3.3 (Distribution class). Denote $\tau_P(z, v) = \mathbb{E}_P J(z, Z, v) + \mathbb{E}_P J(Z, z, v)$ for any $(X, Y) \sim P$. Define the second-order matrices as

$$\Delta_1(P) = \mathbb{E}_P\left[\nabla_P \tau(Z, v_0) \nabla_P^\top \tau(Z, v_0)\right], \quad \Delta_2(P) \ = -\mathbb{E}_P\left[\nabla_P^2 \tau(Z, v_0)\right]. \tag{11}$$

We denote by $\mathcal{P}_0$ the collection of distributions P for $Z = (X, Y)$ that satisfy Assumption 3.2 and additionally obey $\lambda_{\min}(\Delta_1(P)) \wedge \lambda_{\min}(\Delta_2(P)) \geq c_\Delta$ for some constant $c_\Delta > 0$.

Notice that, according to [67] and [25], the lower-bounded eigenvalue condition holds naturally for Han's MRC kernel $J(Z_1, Z_2, v)$, provided that the link function is non-degenerate and the support $\mathcal{X}$ spans the entire space. Here, we define $\mathcal{P}_0$ primarily for minimax analysis purposes; for each fixed distribution class, the lower-bounded eigenvalue condition automatically holds.

The theoretical result of our Algorithm 1 is given by the following theorem:

THEOREM 3.4 (Estimation error). *Under Assumption 3.2.1, there exists a constant $C_0 > 0$ independent of d, n such that with probability at least $1 - \varepsilon$,*

$$\left\|\widehat{v}^{\mathsf{H}} - v_0\right\|_2 \leq C_0 \sqrt{\frac{d + \log^2(\frac{1}{\varepsilon})}{n}}, \tag{12}$$

provided that $n \geq C_{v,0}\left(d + \log^2(\frac{1}{\varepsilon})\right)$ for some large $C_{v,0} > 0$ that is independent of d, n.

Moreover, if we further impose Assumption 3.2.2 and set $h_n = \left(\frac{\log n}{n}\right)^{\frac{1}{2\beta+1}} \vee C_{\mathsf{H}}\left(\frac{d+\log^2(\frac{1}{\varepsilon})}{n}\right)^{\frac{1}{2}}$ for some large C_{H}, then there exists a constant $c_{\mathsf{H}} \gtrsim C_0/C_{\mathsf{H}}$, such that when (12) holds, domain $\widehat{v}^{\mathsf{H}} \circ X \subseteq U(I_0, c_{\mathsf{H}} h_n)$, and with probability at least $1 - \varepsilon$ we have the uniform bound:

$$\begin{aligned}\left\|\widehat{f}_0 - f_0\right\|_\infty &= \sup_{U(I_0, c_{\mathsf{H}} h_n)} \left|\widehat{f}_0(a) - f_0(a)\right| \\ &\leq C_{0,\infty}\left[\sqrt{\frac{\log(\frac{n}{\varepsilon})}{\log n}}\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}} + \left(\frac{d + \log^2(\frac{1}{\varepsilon})}{n}\right)^{\frac{\beta \wedge 1}{2}}\right],\end{aligned}$$

some large $C_{0,\infty} > 0$ that is independent of d, n.

Notice that the cross-fitting in Algorithm 1 is for robustness and better data efficiency, which will not change the order of error rate. An immediate consequence of Theorem 3.4 is that when $d \lesssim n^{\frac{2\beta-1}{2\beta+1}}$, the following uniform bound holds for the final plug-in regression estimator $\widehat{g}_0(x) = \widehat{f}_0((\widehat{v}^{\mathsf{H}})^\top x)$.

COROLLARY 3.5 (Uniform bound). *Under the same condition as Theorem 3.4, if $d \leq Cn^{\frac{2\beta-1}{2\beta+1}}$ for some constant $C > 0$, then there exist constants $C_{0,\infty}, \widetilde{C}_1 > 0$ independent of d, n such that with probability $1 - n^{-10}$,*

$$\|\widehat{g}_0 - g_0\|_\infty = \sup_{x\in\mathcal{X}} \left|\widehat{f}_0((\widehat{v}^{\mathsf{H}})^\top x) - f_0(v_0^\top x)\right| \leq C_{0,\infty}\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}, \tag{13}$$

provided that $n \geq \widetilde{C}_1\left(d + \log^2 n\right)$.

Our Theorem 3.4 and Corollary 3.5 show that Algorithm 1 achieves optimal single-index regression with sharp uniform non-asymptotic error bound, and is therefore suitable for our single-index bandits. These results equivalently give the following uniform ℓ_∞ risk for single-index regression:

$$\sup_{\mathcal{P}_0,\mathcal{H}_+(\beta,L)} \mathbb{E}\left\|\widehat{f}(\widehat{v}^\top x) - f(v^\top x)\right\|_\infty \lesssim \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}.$$

This rate matches the minimax optimal rate for single-index regression [26, 50], which aligns with the minimax optimal rate for 1-dimensional nonparametric regression as in [68], [69].

We remark that our proof of the non-asymptotic bound is novel due to the presence of covariate distribution shift induced by index estimation. Unlike classical nonparametric bandit analyses [13, 39, 42, 62], which typically assume that the density of P_X is bounded both above and below over the entire $\mathbb{R}^d$, our theory only requires regularity of the marginal density of $v_0^\top X$. Consequently, for an estimated index $\widehat{v}^{\mathsf{H}}$, the strong density condition for $(\widehat{v}^{\mathsf{H}})^\top X$ may fail to hold, rendering classical results for local polynomial estimation (LPE) inapplicable. Addressing this challenge necessitates a careful re-examination of the theoretical properties of LPE under index estimation error. Moreover, in our theory (Corollary 3.5), the obtained non-asymptotic rate is dimension-free, revealing the intrinsic sufficient dimension reduction property of the single-index model. In contrast, the recent non-asymptotic analysis of [48] establishes only a MISE bound of order $d^2(\log n)^{\beta+2}\left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}}$, which depends explicitly on d and is unsuitable for bandit regret control.

4. Single-Index Bandits with Online Estimation.

In this section, we present our main algorithm for solving single-index bandits, along with the corresponding assumptions that support our theoretical analysis.

4.1. *Batched bandit algorithm without bin elimination.*

Theorem 3.4 provides a powerful tool to assess the uncertainty of offline single-index regression. It also gives us theoretical foundations to construct upper confidence bounds, a key tool to balance between exploration and exploitation in bandit problems, as long as we can perform online single-index regression following a similar fashion. Following this idea, we describe our batched online contextual bandit algorithm in Algorithm 2.

Our Algorithm 2 features a batched structure and an arm pre-selection approach before we update estimation for the next epoch. We will show later that this pre-selection approach plays a crucial role in learning rewards near the margin region, as the pre-selected arms, informed

Algorithm 2 Batched Contextual Bandit with Single-Index Rewards

Require: Epoch scheme $\{\mathcal{T}_m\}_{m=1}^{M}$, confidence bound $\{\epsilon_m\}_{m=0}^{M-1}$ smoothness level β, $\widehat{g}_{k,0} = 0$, active arm set function $\mathcal{K}_0(x) = [K]$

for $m \in \{1, \ldots, M\}$ **do**

  for $t \in \mathcal{T}_m$ **do**

    Receive covariate X_t, and compute active arm set $\mathcal{K}_{m-1}(X_t)$

    if $\left|\mathcal{K}_{m-1}(X_t)\right| \geq 2$ **then**

      Sample arm $\pi_t \in \mathcal{K}_{m-1}(X_t)$ by uniform sampling

    else

      Choose the only arm in $\mathcal{K}_{m-1}(X_t)$

    end if

  end for

  Log data for each arm k: $\mathcal{S}_{k,m} = \{(X_t, Y_t) \mid t \in \mathcal{T}_m, \pi_t = k\}$.

  For each arm k, update estimation $\widehat{g}_{k,m}$ by running Algorithm 1 on $\mathcal{S}_{k,m}$ with smoothness β.

  Pre-select arm by:

$$\mathcal{K}_{m-0.5}(x) = \left\{k \in \mathcal{K}_{m-1}(x) \mid \widehat{g}_{(1),m-1}(x, \mathcal{K}_{m-1}(x)) - \widehat{g}_{k,m-1}(x) \leq \frac{1}{2}\epsilon_{m-1}\right\}$$

  Update possible arm set

$$\mathcal{K}_m(x) = \left\{k \in \mathcal{K}_{m-0.5}(x) \mid \widehat{g}_{(1),m}(x, \mathcal{K}_{m-0.5}(x)) - \widehat{g}_{k,m}(x) \leq \epsilon_m\right\}$$

end for

return $\{\pi_t\}_{t=1}^{n}$

by data from previous epochs, help eliminate estimations that may fall within irregular areas. Unlike most existing methods [13, 30, 39, 62], our algorithm does not rely on bin partition and elimination. Instead, it implicitly defines decision regions through $\mathcal{K}_m(x)$, thereby avoiding the 2^d factor in the regret upper bound that arises from explicit bin partition.

We explain some notation and settings used in Algorithm 2. For any arm set $\mathcal{K}$, we define $\widehat{g}_{(1),m}(x, \mathcal{K}) = \max_{k \in \mathcal{K}} \{g_k(x)\}$, and $\widehat{g}_{(2),m}(x, \mathcal{K})$ is defined similarly following $g_{(2)}(x)$ in (6). Here, we select the batch and confidence bound with $\epsilon_m = c_\epsilon 2^{-m}$ with a small c_ϵ, and let

$$n_m = \left\lceil C_{\mathcal{T}} \left(\frac{d + \log^2 n}{\epsilon_m^{\frac{2}{1 \wedge \beta}}} + \left(\frac{\log n}{\epsilon_m^2}\right)^{\frac{2\beta+1}{2\beta}} \right) \right\rceil$$

be the length of each epoch for some large $C_{\mathcal{T}}$, with $n_0 = 0$. Let $M = \inf\{m \mid \sum_{i=1}^{m} n_i \geq n\}$ be the total number of epochs. We then define $\mathcal{T}_m = \{t \mid \sum_{i=1}^{m-1} n_{i-1} + 1 \leq t \leq \min\{\sum_{i=1}^{m} n_i, m\}\}$, for $m \in [M]$, and the number of accumulated observations after epoch $\mathcal{T}_m$ as $S_m = \sum_{i=1}^{m} n_i$.

4.2. *Assumptions for single-index bandits.* To move from offline to online estimation and bandit regret control, more assumptions on the optimal decision regions are required so that the learning problem itself is not pathological. We denote

$$\mathcal{Q}_k = \{x \in \mathcal{X} \mid g_k(x) = g_{(1)}(x)\}, \quad I_k := v_k \circ \mathcal{Q}_k = \{v_k^\top x \mid x \in \mathcal{Q}_k\}$$

as the optimal decision region and it projection on index v_k for each arm $k \in [K]$. We assume that $\mathcal{Q}_k \subseteq \mathbb{R}^d$ spans the full space, and $\mathcal{Q}_k$, I_k satisfy the following assumption:

ASSUMPTION 4.1 (Learnable margin area). We make the following assumptions:

4.1.1 For each k and any $x_1, \widetilde{x}, y$ in the support of $(X, Y^{(k)})$, $\mathbb{P}(X \in \mathcal{Q}_k \mid \widetilde{X} = \widetilde{x}, Y^{(k)} = y) \geq \underline{p}_K$.

4.1.2 Define margin area for arm k as $\mathcal{Q}_k^{\mathsf{m}}(\delta) := \{x \mid 0 < g_{(1)}(x) - g_k(x) \le \delta\}$, with it projection on v_k as $I_k'(\delta) := v_k \circ \mathcal{Q}_k^{\mathsf{m}}(\delta)$. There exist constants $\delta_1, c_{\mathsf{m}}, \underline{\mu}_{\mathsf{m}} > 0$ such, for any $n^{-\frac{\beta}{2\beta+1}} \le \delta \le \delta_1$, the set $I_k \cup I_k'(\delta)$ is $\left(c_0, c_{\mathsf{m}}\delta^{1/(\beta\vee 1)}, c_{\mathsf{m}} n^{-\frac{1}{2\beta+1}}\right)$-regular, and conditional on $X \in \mathcal{Q}_k \cup \mathcal{Q}_k^{\mathsf{m}}(\delta)$, the variable $v_k^\top X$ has density that satisfies $\mu_k(\cdot \mid \mathcal{Q}_k \cup \mathcal{Q}_k^{\mathsf{m}}(\delta)) \ge \underline{\mu}_{\mathsf{m}}$ on its support.

Assumption 4.1.1 requires that the (conditional) probability of $\{X \in \mathcal{Q}_k\}$ is bounded away from zero. This rules out the pathological cases where, conditional on optimal decisions $\{X \in \mathcal{Q}_k\}$, the distribution of $X_1|(\widetilde{X}, Y^{(k)})$ is degenerate. Due to the linear structure of single-index reward, this condition is relatively easy to satisfy. For example, when $Y^{(k)}$ is Bernoulli and $\min\{f_k, 1-f_k\} \ge c$, with $v_1 = \cdots = v_K$, a sufficient condition is that $\mathbb{P}(X_1 + \widetilde{v}^\top \widetilde{X} \in I_k \mid \widetilde{X}) \ge c$, which holds, for instance, when I_k is non-vanishing and X_1 has a density bounded away from zero conditional on $\widetilde{X}$.

Assumption 4.1.2 ensures that the margin area is regular and therefore learnable for LPE. This condition is satisfied when I_k is regular and either $I_k'(\delta)$ is connected to I_k, or $I_k'(\delta)$ is itself regular with length of order $\delta^{1/(\beta\wedge 1)}$. An illustration of Assumption 4.1.2 is given in Figure 1.

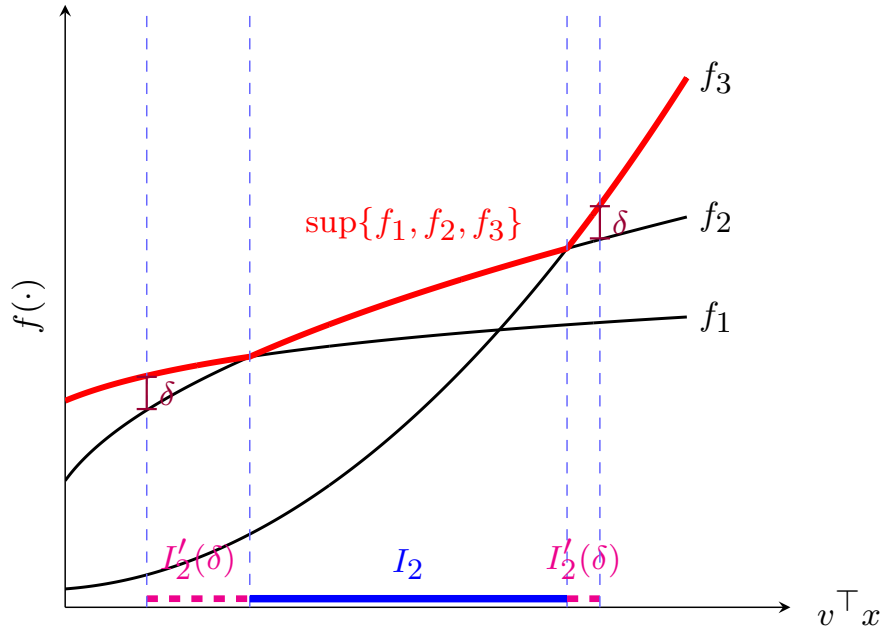

Fig 1: Illustration of regular margin area for Assumption 4.1.2 when $K = 3$ and $v_1 = v_2 = v_3 = v$. Here the red line represents the optimal reward $f_{(1)} = \max\{f_1, f_2, f_3\}$, and I_2 is the region where f_2 is optimal. The purple dashed intervals are $I_2'(\delta)$, representing the margin area for arm 2 at level δ. Assumption 4.1.2 can be naturally satisfied for $I_2 \cup I_2'(\delta)$ as long as I_2 itself is regular, and $I_2'(\delta)$ are pieces of intervals that are connected with I_2, as is shown in the figure.

We require Assumption 4.1.2 as establishing learnability guarantees in single-index bandits is more challenging than in general nonparametric bandits. In the latter case, the optimal decision region is a d-dimensional subset of $\mathcal{X} \subseteq \mathbb{R}^d$, which can be discretized using bin

partition. When P_X has strong density in $\mathbb{R}^d$, the discretized cubes are learnable at the expense of a 2^d factor in the regret upper bound; see, e.g., [39, 62]. By contrast, for single-index bandits, the optimal decision region lies in an *unknown* subspace of $\mathbb{R}^d$, rendering gridding schemes inefficient. Furthermore, since $\{v_k\}$ are unknown, it is impossible to identify irregular regions and thereby perform the screening procedure proposed in [39].

To reach small regret for nonparametric bandit, the following standard margin assumptions are required on the covariate distribution P_X [13, 39, 62]:

ASSUMPTION 4.2 (Margin condition). There exist constants $\alpha \geq 0, \delta_0 > 0, C_\alpha > 0$ such that the reward functions $\{g_k\}_{k=1}^K$ and marginal distribution P_X satisfy

$$\mathbb{P}_X\left(0 < g_{(1)}(X) - g_{(2)}(X) \leq \delta\right) \leq C_\alpha \delta^\alpha, \quad \forall \delta \in (0, \delta_0].$$

DEFINITION 4.3 (Bandit distribution class). We denote by $\mathcal{P}_{\alpha,\beta}$ (or simply $\mathcal{P}$ when there is no ambiguity) the class of bandit distributions such that, for each arm $(X, Y^{(k)}) \sim P^{(k)}$, Assumption 3.2 holds, the conditional distribution $P^{(k)} \mid \mathcal{Q}_k$ satisfies the eigenvalue lower-bound condition in Definition 3.3, and the joint distribution satisfies Assumptions 4.1 and 4.2.

5. Optimal Rates for the Regret. In this section, we study the performance of Algorithm 2, together with the minimax rate of single-index bandits to show that Algorithm 2 is simple yet optimal.

5.1. *Regret upper bound.* We first give a theoretical analysis on the regret of Algorithm 2 under the case when β is known. We will discuss the adaptation to unknown smoothness level in Section 6.

The tuning parameters are important for selecting epoch scheme and confidence bound. We selection our tuning parameters such that $C_{\mathcal{T}} \geq C_{\mathsf{gap}} C_{0,\infty}$ for some large constant C_{gap}, and $c_\epsilon \leq 0.6\,(\delta_0 \wedge \delta_1 \wedge 1)$, where δ_1 is from Assumption 4.1 and δ_0 from Assumption 4.2.

A regret upper bound of Algorithm 2, denoted by $\widehat{\pi}$, is described in Theorem 5.1:

THEOREM 5.1 (Upper bound). *Suppose $d \lesssim n^{\frac{1\wedge(2\beta-1)}{2\beta+1}}$. For distributions in $\mathcal{P}$ given in Definition 4.3, and link functions in $\mathcal{H}_+(\beta, L)$, our Algorithm 2 achieves the regret*

$$\sup_{\mathcal{P}, \mathcal{H}_+(\beta,L)} \mathsf{Reg}(\widehat{\pi}) \leq C(\log n)^{\iota_1(\alpha,\beta)} \cdot \left[n \left(\frac{1}{n}\right)^{\frac{\beta(1+\alpha)}{2\beta+1}} \vee 1 \right],$$

where $C > 0$ is a constant independent of n, d. Here ι_1 only depends on α, β, and $\iota_1(\alpha, \beta) = \frac{1}{2\beta} + \frac{\beta(1+\alpha)}{2\beta+1}$ when $\beta + 1 - \alpha\beta > 0$.

REMARK 1. Theorem 5.1 establishes a minimax upper bound for single-index bandits in the fixed-dimensional setting, which is independent of d. Compared with general nonparametric bandits, which attain a minimax-optimal regret of order $n^{1-\frac{\beta(1+\alpha)}{2\beta+d}}$, our regret upper bound is dimension-free—both in the constant factors and in the convergence rate. Moreover, it avoids the $(\log^{1/\beta} n \vee 2)^d$ dependence in the regret bound that arises from explicit bin partition [13, 30, 39, 62], by instead defining decision regions implicitly through $\mathcal{K}_m(x)$, rather than via bin partition.

We explain why our pre-selection procedure is important in regret control. To better understand how our algorithm explores arms, we define $\mathcal{D}_{k,m-1} = \{x \mid k \in \mathcal{K}_{m-1}(x)\}$ as the region that arm k might be pulled in epoch $m \in [M]$. To control regret, the key is to ensure that after epoch m, g_k in region $\mathcal{D}_{k,m-1}$ can be estiamted with finer accuracy. However,

$\mathcal{D}_{k,m-1}$ itself may not be regular for Algorithm 1 when applying LPE, leaving the estimation uncontrolled. To overcome this, we leverage the regular margin condition in Assumption 4.1.2: this condition ensures that for each $\mathcal{D}_{k,m-1}$, there exists a smaller margin area $\mathcal{Q}_k \cup \mathcal{Q}_k^{\mathsf{m}}(\frac{3\epsilon_{m-1}}{4}) \subseteq \mathcal{D}_{k,m-1}$ that is regular, and that accurate singe-index regression guarantees $\mathcal{D}_{k,m-1} \subseteq \mathcal{Q}_k \cup \mathcal{Q}_k^{\mathsf{m}}(\frac{5\epsilon_{m-1}}{4})$ for a slightly larger margin area. We pre-select arms using $\widehat{g}_{k,m-1}$, which is accurate in a broader margin $\mathcal{Q}_k \cup \mathcal{Q}_k^{\mathsf{m}}(\frac{5\epsilon_{m-1}}{4})$ that contains $\mathcal{D}_{k,m-1}$. After pre-selection, the region requiring finer decision-making is confined within $\mathcal{Q}_k \cup \mathcal{Q}_k^{\mathsf{m}}(\frac{3\epsilon_{m-1}}{4})$ as long as the estimation error is sufficiently smaller than ϵ_m.

5.2. *Regret lower bound.* Although Theorem 5.1 presents an exciting regret upper bound that can achieve sufficient dimension reduction and matches the minimax optimal regret for one-dimensional nonparametric bandits, it is still questionable on whether the upper bound is optimal for the distribution class $\mathcal{P}, \mathcal{H}_+(\beta, L)$. In this section, we construct a matching lower bound to show that the regret upper bound given in Theorem 5.1 is indeed minimax optimal (only up to logarithmic factors) We describe our matching lower bound in Theorem 5.2:

THEOREM 5.2 (Regret lower bound). *Suppose $\alpha\beta \leq 1$. For any admissible policy π, one has the minimax lower bound*

$$\inf_{\pi} \sup_{\mathcal{P}, \mathcal{H}_+(\beta,L)} \mathsf{Reg}(\pi) \geq c \left(n \left(\frac{1}{n} \right)^{\frac{\beta(1+\alpha)}{2\beta+1}} \vee 1 \right) \tag{14}$$

where $c > 0$ is a constant independent of n, d.

This result shows that the derived regret rate is tight up to logarithmic factors. The proof is based on constructing instances that are difficult to distinguish and leveraging the inferior sampling rate established in [66] to control the regret under margin conditions. In addition to the instance construction that is distinct from those in [13, 39], we also need to verify that Assumptions 3.2 and 4.1 are satisfied.

REMARK 2. Together with Theorem 5.1, our lower bound in Theorem 5.2 reveals that, for the function class $\mathcal{H}_+(\beta, L)$, the fundamental difficulty of single-index bandits is the same as one-dimensional nonparametric bandits. The rationale behind this result is that, under monotonic link functions, the indices for all arms can be learned optimally even without explicit knowledge of the link functions themselves. Consequently, by applying the index plug-in, the problem effectively reduces to a one-dimensional nonparametric bandit.

5.3. *Threshold for static oracle policy.* The discussion of nonparametric bandits is often linked to a fundamental question: when is the oracle policy static? A static oracle policy refers to the existence of a fixed arm that is strictly better than all others. When such an oracle policy exists, the problem becomes arguably easier, as the contextual bandit setting effectively reduces to a standard multi-armed bandit problem. However, we emphasize two important remarks. (i) The existence of a static oracle policy does not imply that covariate information can be ignored, as it remains essential for identifying and excluding suboptimal arms in the margin region. (ii) The existence of a static oracle policy also does not guarantee logarithmic or constant regret. As shown by the classical instance-dependent lower bound in [47], the achievable regret fundamentally depends on the "optimal arm gap", which may still lead to large regret.

Pioneered by the study of fast learning rates in nonparametric classification [5], it is well understood that under a strong density condition on P_X over $\mathcal{X} \subseteq \mathbb{R}^d$, the threshold $\alpha(\beta \wedge$

$1) > 1$ characterizes the existence of a static optimal function. For general nonparametric bandits, [62, 66] established that a threshold of $\alpha(1 \wedge \beta) > d$ determines when a static oracle policy can be identified for K-armed problems under the same strong density condition. In the special case of $K = 2$, [30, 39] further refined this threshold to $\alpha(1 \wedge \beta) > 1$ under similar assumptions.

However, for single-index bandits with general $K \geq 2$, these results no longer apply because the strong density condition may not hold for the full distribution P_X. In our setting, we impose regularity only on the marginal distributions of $v_k^\top X$. To address this, we relax the strong density condition and consider a broader class of distributions where the strong density condition is violated to some extent, and we derive a new threshold for the existence of a static oracle policy under such relaxed conditions. We present our result in Proposition 5.3.

PROPOSITION 5.3. *Suppose that $g_k(x) : \mathbb{R}^d \to \mathbb{R}$ (not necessarily single-index) belong to the d-dimensional Hölder class with parameters (β, L), and that Assumption 4.2 holds for the bandit problem. Assume further that the functions $\{g_k\}_{k\in[K]}$ have no ties almost surely. We consider a class of covariate distributions for which there exist constants $\gamma \geq 0$ and $C_\gamma, c_\gamma > 0$ such that, for any $x \in \mathcal{X}$ and $\varepsilon \in (0, C_\gamma]$, $\mathbb{P}_X(\mathcal{X} \cap \mathbb{B}_2(x, \varepsilon)) \geq c_\gamma \varepsilon^{d+\gamma}$. Then, the following statements hold:*

5.3.1 *If $\alpha(1 \wedge \beta) > 1 + \gamma$, then a function g_k is either always or never optimal, and the oracle policy $\pi^\star$ in the bandit problem is static (only pulls one arm all the time);*

5.3.2 *If $\alpha(1 \wedge \beta) \leq 1 + \gamma$, then there exist distribution P_X with single-index rewards functions $g_k(x) = f_k(v_k^\top x)$ such that (i) each $v_k^\top X$ shares strong denstiy condition; and (ii) the optimal functions are not fixed, and the oracle policies are non-trivial.*

5.3.3 *If further $g_k(x)$ are single-index with shared index space, i.e., $g_k(x) = f_k(v_0^\top x)$, and $v_0^\top X$ has lower bounded density with (c_0, r_0)-regular support, then when $\alpha(1 \wedge \beta) > 1$, the oracle policy $\pi^\star$ is static; when $\alpha(1 \wedge \beta) \leq 1$ there exist distributions such that the oracle policies are non-trivial.*

Here, γ is a parameter measuring the decay speed of density. When $\gamma = 0$, we return to the classic strong density condition. Proposition 5.3.1 shows that, for any $K \geq 2$, the threshold for identifying static oracle policy it $\alpha(1 \wedge \beta) > 1 + \gamma$. Together with Proposition 5.3.2, we reveal that for single-index bandits with covariates that may not follow the strong density condition, the corresponding threshold will increase by γ. In the special case when $\gamma = 0$, this recovers the earlier findings of [30, 39], which were restricted only to $K = 2$. In another special case when $v_1 = \cdots = v_K = v_0$, the problem reduces to a one-dimensional nonparametric bandit, and Proposition 5.3.3 gives the threshold $\alpha(1 \wedge \beta) > 1$ under strong margin density for $v_0^\top X$, rather than a strong density condition for the whole $X \in \mathbb{R}^d$.

6. Adaptation to Unknown Smoothness. This section discusses the adaptation to the unknown smoothness level β in single-index bandits. Our Algorithm 2 and theoretical upper bound established in Section 5 necessarily require knowing the exact smoothness level β. However, in many applications, β is unavailable to decision makers. In this section, we discuss the adaptation to the unknown smoothness level when only a range of β, $[\underline{\beta}, \overline{\beta}]$ is given such that $\beta \in [\underline{\beta}, \overline{\beta}]$. We will show that under the standard self-similarity condition, achieving adaptation to unknown β is possible by an adaptive algorithm with regret that is still minimax optimal up to logarithmic factors.

6.1. *Impossibility of adaptation.* It is well established in the nonparametric multi-armed bandit literature that adaptation to an unknown smoothness level without incurring additional cost is impossible [16, 30, 56]. Without further restrictions, no algorithm can simultaneously achieve the minimax-optimal rate across a family of nonparametric bandit problems with varying smoothness parameters. This impossibility result naturally extends to single-index bandits by considering the special case $v_k^\top X = X_1$, which reduces the problem to a one-dimensional nonparametric bandit. A straightforward modification of the proof of Theorem 3.1 in [30] shows that cost-free adaptation is also impossible within the class $\mathcal{P}, \mathcal{H}_+(\beta, L)$. As discussed in [30], this failure of adaptation in nonparametric bandits arises partly from the incomplete feedback inherent to the bandit setting. This contrasts with nonparametric regression [15, 35], where adaptation can sometimes be achieved under shape constraints on the function class.

THEOREM 6.1. *Consider the case when $v_k^\top x = x_1$, $g_k(x) = f_k(x_1)$. Given two smoothness levels $\beta_1 < \beta_2$, and assume $0 < \alpha\beta_2 \le \beta_2 \vee 1$, $\beta_1(1+\alpha) \ge 1$. Then, for any admissible $\pi \in \Pi$ that achieves rate-optimal regret $\mathsf{opt}_{\mathsf{Reg}}(n, \alpha, \beta_2) := n \cdot (\frac{1}{n})^{\frac{\beta_2(\alpha+1)}{2\beta_2+1}}$ over $\mathcal{P}_{\alpha,\beta_2}, \mathcal{H}_+(\beta_2, L)$, there exists a constant $c > 0$ independent of n, d such that the following holds:*

1. *(At most Lipschitz smooth) If $0 < \beta_1 < \beta_2 \le 1$, then*

$$\sup_{\mathcal{P}_{\alpha,\beta_1}, \mathcal{H}_+(\beta_1, L)} \mathsf{Reg}(\pi) \ge c \cdot \mathsf{opt}_{\mathsf{Reg}}(n, \alpha, \beta_1) \cdot n^{\frac{(\alpha+1)\beta_1}{2\beta_1+1} - \frac{\beta_1+1}{(\alpha+1)(2\beta_1+1-\alpha\beta_1)} - \frac{\alpha(\beta_2+1)(1+\beta_2(1-\alpha))}{(\alpha+1)(2\beta_2+1-\alpha\beta_2)(2\beta_2+1)}}$$

2. *(At least Lipschitz smooth) If $\beta_1 = 1 < \beta_2$, then*

$$\sup_{\mathcal{P}_{\alpha,\beta_1}, \mathcal{H}_+(\beta_1, L)} \mathsf{Reg}(\pi) \ge c \cdot \mathsf{opt}_{\mathsf{Reg}}(n, \alpha, \beta_1) \cdot n^{\frac{(\alpha+1)\beta_1}{2\beta_1+1} + \frac{\alpha\beta_2}{2\beta_2+1} - 1}$$

Clearly, this theorem demonstrates that a rate-optimal single-index bandit algorithm designed for the function class $\mathcal{H}_+(\beta_2, L)$ cannot be minimax-optimal for $\mathcal{H}_+(\beta_1, L)$ without additional restrictions. To better illustrate this gap, consider the following examples: (i) when $\beta_1 = 0.18$, $\beta_2 = 0.2$, and $\alpha = 1/\beta_2 = 5$, we have $\sup_{\mathcal{P}_{\alpha,\beta_1}, \mathcal{H}_+(\beta_1, L)} \mathsf{Reg}(\pi) \gg \mathsf{opt}(n, \alpha, \beta_1) \cdot n^{0.0094}$; (ii) when $\beta_1 = 1$, $\beta_2 = 2$, $\alpha = 1$, we have $\sup_{\mathcal{P}_{\alpha,\beta_1}, \mathcal{H}_+(\beta_1, L)} \mathsf{Reg}(\pi) \gg \mathsf{opt}(n, \alpha, \beta_1) \cdot n^{0.06}$. Both cases clearly indicate the impossibility of cost-free adaptation within the function class $\mathcal{H}_+(\beta, L)$.

6.2. *Self-similarity condition.* Given the impossibility of adaptation to unknown β under general settings, we seek to explore the adaptive algorithms when more assumptions are imposed. As suggested by [13, 16, 30], when the reward functions are "self-similar", it is possible to extend the smoothness estimation technique to nonparametric bandits and achieve adaptation on β. We now show that adaptation to β is possible for single-index bandits under self-similar rewards. We describe the self-similarity condition introduced in [63], [27] by first defining the ℓ_2-projection to polynomials:

DEFINITION 6.2. Let $f(\cdot)$ be any function and let $l, p \ge 0$ be integers. We write $\mathrm{P}_h^p f(\cdot; U)$ for the ℓ_2-projection of f onto the space of polynomials of degree at most p over the hypercube U, measured with respect to P_X. Concretely, for each $x \in U$ we define

$$\mathrm{P}_h^p f(x; U) := g(x), \quad \text{s.t.} \quad g = \arg\min_{q \in \mathrm{Poly}(p)} \int_U |f(u) - q(u)|^2 K\left(\frac{x-u}{h}\right) p_X(u \mid U) du,$$

where the kernel is $K(\cdot) = \mathbb{I}\{|\cdot| \le 1\}$ with bandwidth h, and $\mathrm{Poly}(p)$ represents the collection of polynomials of degree no greater than p.

To facilitate the definition of self-similarity, we denote the lattice in $\mathbb{R}$ as $\mathcal{B}_{2,l} = \{B_i \mid B_i = [\frac{i-1}{2^l}, \frac{i}{2^l}], i \in \mathbb{Z}\}$. The self-similar functions are given in Definition 6.3.

DEFINITION 6.3 (Self-similar functions). Given some $\beta \in [\underline{\beta}, \overline{\beta}]$, and finite constants $l_0 \geq 0$ and $b > 0$, define the class of self-similar sets of function, $\mathcal{H}^{ss}(\beta, L, b, l_0)$, to be the collection of the sets of functions $\{f_k\}_{k\in[K]}$ such that $f_k \in \mathcal{H}(\beta, L), k \in [K]$, and for which for any integers $l \geq l_0$, and $\lfloor\beta\rfloor \leq p \leq \lfloor\overline{\beta}\rfloor$, one has

$$\max_{B\in\mathcal{B}_{2,l}} \max_{k\in[K]} \sup_{x\in B} \left|\mathrm{P}^p_{2^{-l}} f_k(x; B) - f_k(x)\right| \geq b \cdot 2^{-l\beta}.$$

The self-similarity condition complements the Hölder smoothness in the sense that the function class is hard to estimate with a bias lower bounded by h^β. Similar to $\mathcal{H}_+(\beta, L)$ defined in (5), we also define the monotonic self-similar reward class $\mathcal{H}^{ss}_+(\beta, L, b, l_0)$ analogously. Our first result for the self-similar reward class is that the self-similarity condition will not reduce the complexity of the problem.

THEOREM 6.4 (Regret lower bound under self-similarity). *Under the assumptions of Theorem 5.2, if further* $(1+\alpha)\beta \geq 1$*, the minimax lower bound satisfies*

$$\inf_{\pi} \sup_{\mathcal{P}, \mathcal{H}^{ss}_+(\beta,L,b,l_0)} \mathsf{Reg}(\pi) \geq c\left(n\left(\frac{1}{n}\right)^{\frac{\beta(1+\alpha)}{2\beta+1}} \vee 1\right) \tag{15}$$

Here, the condition $(1+\alpha)\beta \geq 1$ is required only for the strictly monotonic function class due to the standard α-margin instance construction [5, 13, 30]. For the general self-similar (β, L)-Hölder smooth class reward functions $\mathcal{H}^{ss}(\beta, L, b, l_0)$, this condition is not needed, as is shown in [30].

6.3. *Adaptation under self-similarity condition.* We study the problem of adaptation to an unknown smoothness parameter β, assuming that only an interval $[\underline{\beta}, \overline{\beta}]$ is known such that $\beta \in [\underline{\beta}, \overline{\beta}]$. Without loss of generality, we assume $\underline{\beta} < 1 < \overline{\beta}$. Our idea is to first estimate β by leveraging Lepski's method [51], which identifies the optimal bandwidth (corresponding to the true smoothness) by comparing the estimation bias and stochastic error across different bandwidths during the exploration phase. We then use an undersmoothed estimator based on the estimated β as an input to Algorithm 2 for regret control. The MRC and index plug-in components are employed for dimension reduction.

We introduce several notations and settings. We denote

$$l_1 = l := \left\lceil \frac{\underline{\beta}\log_2 n}{(2\overline{\beta}+1)^2} \right\rceil, \quad l_2 = l + \left\lceil \frac{1}{\underline{\beta}} \log_2 \log n \right\rceil. \tag{16}$$

Our apporach focuses on comparing $\max_{B\in\mathcal{B}_{2,l},k\in[K]} \left|\widehat{f}_k(z; B, 2^{-l_1}) - \widehat{f}_k(z; B, 2^{-l_2})\right|$. We consider first estimating the smoothness parameter with N_0 data, and then plug the estimate β_{est} into Algorithm 2. To this end, we define the grid for comparison as $\mathcal{M} = \left\{\frac{k}{2^{-l_3}}, k \in \mathbb{Z}\right\}$, where $l_3 = \frac{\overline{\beta}}{\underline{\beta}} l_1 + \frac{1}{\underline{\beta}} \log_2 \log n$, and $\mathcal{M}(B) = \mathcal{M} \cap B$ for any $B \in \mathcal{B}_{2,l_1}$ We propose the estimation procedures for smoothness of link functions in Algorithm 3.

The performance of Algorithm 3 is guaranteed in the following Proposition 6.5:

Algorithm 3 Link Smoothness Estimation

Require: Smoothness ranges $\underline{\beta}, \overline{\beta}$.

Initialize $N_0 = \lceil C_{\mathsf{gap}}(d+\log^2 n) n^{\frac{2\underline{\beta}(\overline{\beta}+1)}{(2\overline{\beta}+1)^2}} \rceil$, l_1, l_2, l_3 in (16), $C_l > 0$

for $k \in \{1, \dots, K\}$ **do**

Pull arm k in the following $2N_0$ round, and denote the first and second N_0 data as $\mathcal{S}_{1,k}, \mathcal{S}_{2,k}$.

Compute MRC $\widehat{v}_k^{\mathsf{H}}$ from $\mathcal{S}_{1,k}$, and project covariates in $\mathcal{S}_{2,k}$ onto direction $\widehat{v}_k^{\mathsf{H}}$: $\mathcal{S}'_{2,k} = \{((\widehat{v}_k^{\mathsf{H}})^\top X_t, Y_t) \mid (X_t, Y_t) \in \mathcal{S}_{2,k}\}$.

for $B \in \mathcal{B} \cap U\left(\widehat{v} \circ \mathcal{X}, C_{\mathsf{H}}\sqrt{\frac{d}{N_0}}\right)$ where $B \cap \{X'_t \mid (X'_t, Y_t) \in \mathcal{S}'_{2,k}\} \neq \emptyset$ **do**

Run LPE on $\{(X'_t, Y_t) \in \mathcal{S}'_{2,k} \mid X'_t \in B\}$ with level $\lfloor \overline{\beta} \rfloor$, bandwidth 2^{-l_1}, 2^{-l_2} and uniform kernel $\mathbb{I}\{|\cdot| \le 1\}$, denote the regression function as $\widehat{f}_k(z; B, 2^{-l_1}), \widehat{f}_k(z; B, 2^{-l_2})$

end for

end for

Compute

$$b_{\max} = \max_{k\in[K], B, a \in \mathcal{M}(B)} \left| \widehat{f}_k(a; B, 2^{-l_1}) - \widehat{f}_k(a; B, 2^{-l_2}) \right|.$$

$$\beta_{\mathsf{est}} = -\frac{1}{l_1}\log_2 b_{\max} - C_l \frac{\log_2 \log n}{\log_2 n}$$

return $\beta_{\mathsf{est}} \leftarrow (\beta_{\mathsf{est}} \vee \overline{\beta}) \wedge \underline{\beta}$

PROPOSITION 6.5. *Assume that $\overline{\beta} > 2/(\underline{\beta} \wedge (1-\underline{\beta}))$. With the choice $N_0 = \lceil C_{\mathsf{gap}}(d + \log^2 n) n^{\frac{2\underline{\beta}(\overline{\beta}+1)}{(2\overline{\beta}+1)^2}} \rceil$ for some large $C_{\mathsf{gap}} > 0$, we have that*

$$\mathbb{P}\left(\beta_{\mathsf{est}} \in \left[\beta - \frac{C_{\mathsf{est}}(2\overline{\beta}+1)^2 \log_2 \log n}{\underline{\beta} \log_2 n}, \beta\right]\right) \ge 1 - n^{-2},$$

for some constant $C_{\mathsf{est}} > 0$ indepnedent of $\alpha, n, d, \beta, \overline{\beta}, \underline{\beta}$.

Proposition 6.5 reveals that our β_{est} is a good "undersmooth" estimator for β, which ensures that with high probability, we have $\mathcal{H}_+(\beta, L) \subseteq \mathcal{H}_+(\beta_{\mathsf{est}}, L)$. This is important for the subsequent plug-in approach. Based on the smoothness estimation, we present our smoothness adaptive algorithm in Algorithm 4, with the corresponding regret control in Theorem 6.6.

Algorithm 4 Smoothness Adaptive Single-Index Bandits

Require: $\underline{\beta}, \overline{\beta}, N_0$

for $t = 1, \dots, KN_0$ **do**

Run Algorithm 3 and get β_{est}.

end for

for $t = KN_0 + 1, \dots, n$ **do**

Run Algorithm 2 with smoothness level β_{est} with sample index $t' \leftarrow t - KN_0$.

end for

return $\{\pi_t\}_{t=1}^n$

THEOREM 6.6. *Under the conditions of Proposition 6.5, suppose when run Algorithm 3 for the first $2KN_0$ rounds, and run Algorithm 2 for the rest with smoothness level β_{est}, then we have*

$$\sup_{\mathcal{P},\mathcal{H}_+^{ss}(\beta,L,b,l_0)} \mathsf{Reg}(\pi) \le C(\log n)^{\iota_2(\alpha,\beta,\overline{\beta},\underline{\beta})} \cdot \left[n \left(\frac{1}{n}\right)^{\frac{\beta(1+\alpha)}{2\beta+1}} \vee 1 \right],$$

provided that $\alpha(\beta \wedge 1) \le 1$ *and* $d \lesssim n^{\frac{2(1-\underline{\beta})\overline{\beta}+1-2\underline{\beta}}{(2\overline{\beta}+1)^2}}$*, for some constant* C *indepnedent of* n, d*. Here* $\iota_2(\alpha,\beta,\overline{\beta},\underline{\beta})$ *is a function that is only related to* $\alpha,\beta,\overline{\beta},\underline{\beta},C_{\mathsf{est}}$*, and is independent of* n, d*, and* $\iota_2(\alpha,\beta,\overline{\beta},\underline{\beta}) = \frac{1}{2\underline{\beta}} + \frac{\beta(1+\alpha)}{2\beta+1} + C_{\mathsf{est}}(1+\alpha)\frac{(2\overline{\beta}+1)^2}{(2\underline{\beta}+1)^2}$ *when* $\beta + 1 - \alpha\beta > 0$*.*

7. Single-Index Bandits in Increasing Dimensions. A large body of the literature on single-index models focuses on the case where the dimension d is fixed with respect to n. However, in many practical problems, the dimension may increase with n (while still satisfying $d \ll n$ for learnability). In this section, we study the fundamental difficulty of single-index bandits in increasing-dimensional settings, where d can grow with n but remains much smaller than n.

7.1. *Phase transition in increasing dimensions.* Interestingly, in the case when d is increasing, the minimax rate exhibits a phase transition between the nonparametric and linear regimes. Specifically, when d is small, the optimal rate is governed by the nonparametric phase, where the smoothness parameter β determines the regret rate. In contrast, when d becomes large, the optimal rate enters the parametric phase, as learning the index becomes more challenging; in this case, the optimal regret corresponds to that of linear bandits [8, 59], which is independent of β.

We show that in the increasing dimensions case, our Algorithm 2 shares a two-phase regret upper bound. For simplicity, we assume $\beta \ge 1$ and is known.

THEOREM 7.1 (Regret upper bound). *Suppose* $\beta \ge 1, d \ll n$*. Our Algorithm 2, denoted by* $\widehat{\pi}$*, achieves the regret*

$$\sup_{\mathcal{P},\mathcal{H}_+(\beta,L)} \mathsf{Reg}(\widehat{\pi}) \le C(\log n)^{\iota_3(\alpha,\beta)} \cdot \left[n \left(\left(\frac{1}{n}\right)^{\frac{\beta(1+\alpha)}{2\beta+1}} + \left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}\wedge 1} \right) \vee 1 \right],$$

where $C > 0$ *is a constant independent of* n, d*. Specifically, we have*

$$\sup_{\mathcal{P},\mathcal{H}_+(\beta,L)} \mathsf{Reg}(\widehat{\pi}) \lesssim \begin{cases} (\log n)^{\iota_1(\alpha,\beta)} \cdot \left[n^{1-\frac{(1+\alpha)\beta}{2\beta+1}} \vee 1 \right] & \text{if } d \lesssim n^{\frac{1}{2\beta+1}} \\ (\log n)^{\iota_3(\alpha,\beta)} \cdot n \left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}} & \text{if } d \gtrsim n^{\frac{1}{2\beta+1}}, \alpha \le 1 \end{cases}. \tag{17}$$

Here, $\iota_1(\alpha,\beta)$ *is the same as the function defined in Theorem 5.1.*

Notably, our algorithm achieves this two-phase regret bound without prior knowledge of which phase the problem belongs to. The rate $n^{1-\frac{(1+\alpha)\beta}{2\beta+1}}$ corresponds to the optimal regret in the nonparametric phase, consistent with the findings in Section 5. In contrast, the rate $n\left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}}$ characterizes the optimal regret in the linear bandit phase, as established in [8, 52, 59]. For linear bandits, when $\alpha > 1$, the regret upper bound becomes $\widetilde{O}(d)$, representing the unavoidable learning cost in $\mathbb{R}^d$ for most algorithms. A matching lower bound is provided in Theorem 7.2, demonstrating that Algorithm 2 is, in fact, minimax-optimal up to logarithmic factors in most cases.

THEOREM 7.2 (Regret lower bound). *Suppose $\alpha\beta \le 1$ For any admissible policy π, one has the minimax lower bound*

$$\inf_{\pi} \sup_{\mathcal{P},\mathcal{H}_+(\beta,L)} \mathsf{Reg}(\pi) \ge c \cdot \begin{cases} n^{1-\frac{(1+\alpha)\beta}{2\beta+1}} \vee 1 & \text{if } d \lesssim n^{\frac{1}{2\beta+1}} \\ n\left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}} & \text{if } d \gtrsim n^{\frac{1}{2\beta+1}}, \alpha \le 1 \end{cases}. \tag{18}$$

where $c > 0$ is a numerical constant independent of n, d.

Theorem 7.2 establishes two phases of the regret lower bound that match the upper bound in (17), and thus bridge the gap between linear bandits and nonparametric bandits through a single-index reward model. An illustration of this phase transition is provided in Figure 2. This result reveals that the threshold between the nonparametric and parametric regimes in single-index bandits occurs at $d \asymp n^{\frac{1}{2\beta+1}}$. Specifically, when $d \lesssim n^{\frac{1}{2\beta+1}}$, learning the nonparametric link function dominates the difficulty of the problem, and the setting lies in the nonparametric phase. Conversely, when $d \gtrsim n^{\frac{1}{2\beta+1}}$, learning the parametric index becomes the bottleneck, and the problem transitions into the parametric phase.

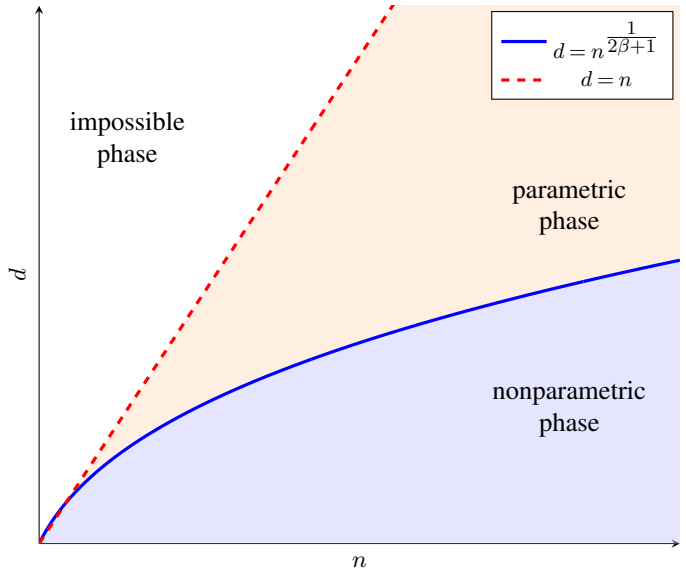

Fig 2: Phase transition of optimal regret in single-index bandits. When $d \gtrsim n^{\frac{1}{2\beta+1}}$, the optimal rate is the nonparametric rate $n^{1-\frac{(1+\alpha)\beta}{2\beta+1}}$. When $d \lesssim n^{\frac{1}{2\beta+1}}$, the optimal rate is the same as linear bandits, which is of order $n\left(\frac{d}{n}\right)^{\frac{1+\alpha}{2}}$. When $d \gtrsim n$, controlling regret is impossible as there is not enough data to learn indices in $\mathbb{R}^d$, leaving the problem indeterminate.

The parametric phase of single-index bandits shares certain similarities with linear bandits, yet it remains distinct due to the presence of unknown link functions. The inclusion of link functions enables the construction of lower bounds in regimes where d can be large, but the problem remains learnable, i.e., $n^{\frac{1}{2\beta+1}} \lesssim d \ll n$. In contrast, the lower bound for linear bandits in [59] holds only when $d \ge Cn$ for some sufficiently large constant $C > 0$. Similar to the linear bandit setting, our lower bound does not cover the regime $\alpha > 1$ as the optimality in this case remains an open question in the linear bandit literature; see, for example, [28, 52, 59].

7.2. *Sub-optimality of plug-in estimators in data-poor regime.* In Section 7.1, we discuss the optimal algorithm and the associated phase transition when $\beta \geq 1$, showing that the index plug-in idea leads to the minimax-optimal Algorithm 2 for single-index bandits in both the parametric and nonparametric phases. This result follows from the fact that our single-index regression procedure, Algorithm 1, achieves rate optimality in both regimes when $\beta \geq 1$. We define the optimal single-index estimation rate as $\mathsf{opt}_{\mathsf{est}}(n, d, \beta) := (\frac{1}{n})^{\frac{\beta}{2\beta+1}} \vee \sqrt{\frac{d}{n}}$ (as is revealed in [4, 26]), then the optimal regret for bandits is basically of order

$$\mathsf{opt}_{\mathsf{Reg}}(n, d, \beta) = n \cdot [\mathsf{opt}_{\mathsf{est}}(n, d, \beta)]^{1+\alpha}.$$

However, when $\beta < 1$, the index plug-in approach may become suboptimal, particularly in the data-poor regime—that is, when n is small but d is large ($d \gtrsim n^{\frac{2\beta-1}{2\beta+1}}$). In this setting, even under the distributional class $\mathcal{P}_0$ defined in Assumption 3.2, the index plug-in estimator cannot achieve the estimation lower bound of $\mathsf{opt}_{\mathsf{est}}(n, d, \beta) = \sqrt{\frac{d}{n}}$ when $d \gtrsim n^{\frac{2\beta-1}{2\beta+1}}$. To formalize this issue, we define $\mathcal{R}_\infty$ as the ℓ_∞ risk incurred when estimating the single-index model using the index plug-in approach: $\mathcal{R}_\infty(\widehat{f}, \widehat{v}) = \inf_{\widehat{f},\widehat{v}} \sup_{\mathcal{P}_0} \mathbb{E} \left\| \widehat{f}(\widehat{v}^\top x) - f(v^\top x) \right\|_\infty$. We describe our lower bound for plug-in estimators in Proposition 7.3.

PROPOSITION 7.3. *Suppose we estimate the single index regression function by a plug-in estimator: $\widehat{g}(x) = \widehat{f}(\widehat{v}^\top x)$, then we have*

$$\mathcal{R}_\infty(\widehat{f}, \widehat{v}) = \inf_{\widehat{f},\widehat{v}} \sup_{\mathcal{P}_0} \mathbb{E} \left\| \widehat{f}(\widehat{v}^\top x) - f(v^\top x) \right\|_\infty \gtrsim \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}} + \left(\frac{d}{n}\right)^{\frac{\beta \wedge 1}{2}}.$$

When further $\beta < 1$ and $d \gtrsim n^{\frac{2\beta-1}{2\beta+1}}$, we have

$$\mathcal{R}_\infty(\widehat{f}, \widehat{v}) \gtrsim \left(\frac{d}{n}\right)^{\frac{\beta}{2}} \gg \mathsf{opt}_{\mathsf{est}}(n, d, \beta) = \sqrt{\frac{d}{n}}.$$

These results indicate that no simple single-index estimator with a fixed index plug-in can achieve rate-optimal estimation in the data-poor regime. This limitation arises partly because, when $\widehat{v}$ is fixed for all x, the poor smoothness of the link function prevents the final estimator from attaining optimal performance.

To achieve optimality in single-index regression (and consequently in single-index bandits), the index estimation must take a more sophisticated form, e.g., varying with respect to x. However, adapting the index v locally for each x is inherently challenging in single-index regression. For example, [50] shows that such pointwise adaptation for $\beta < 1$ is feasible only when $d = 2$. Their approach, however, requires solving a complex optimization problem that must be recomputed for every x, making it computationally infeasible in large-scale settings where numerous queries are required, such as in bandit problems. Moreover, their technique does not extend to $d > 2$, since constructing the matrix necessary for kernel estimation and optimization becomes intractable in higher dimensions.

Since most rate-optimal bandit algorithms rely critically on achieving the optimal estimation rate of the underlying regression function, it becomes challenging to design rate-optimal algorithms for $\beta < 1$ based on the index plug-in approach when the problem lies in the data-poor regime. This phenomenon in the data-poor regime is also observed in the literature on single-index regression with index plug-in estimators, as discussed in [19, 45, 48].

8. Numerical Experiments. In this section, we investigate the numerical performance of the proposed algorithms through a series of simulations on single-index bandits. We begin by evaluating Algorithm 2 under known β, presenting results on both online index estimation and cumulative regret. The experiments demonstrate that Algorithm 2 effectively learns the latent index via online MRC. In our simulations, we compare our method with the smooth-bin successive elimination algorithm (SmoothBandit), studied in [30, 39] for general d-dimensional nonparametric bandits. We then examine our adaptive approach for the setting where β is unknown by running Algorithm 4. Specifically, we apply Algorithm 3 to obtain an estimate β_{est}, which is subsequently used in Algorithm 2 for regret control. These simulations highlight the benefits of dimension reduction in single-index bandits and demonstrate the empirical effectiveness of our approach.

Set up. We consider a 3-armed contextual bandit setting with $K = 3$ and $d = 4$. The vectors v_1, v_2, v_3 are randomly generated to satisfy $\|v_k\|_2 \le 2$ and $v_{k,1} = 1$, and linear independence. The covariates X_t are drawn from a truncated standard Gaussian distribution, $X_t \sim \mathcal{N}(0, I_d) \mid \|X_t\|_2 \le 1$. For the reward distributions, each $Y_t^{(k)}$ follows the regression model $Y_t^{(k)} = f_k(v_k^\top X_t) + \xi_t$, where $\xi_t \sim \mathcal{N}(0, 0.1)$, and is independent of X_t. The link functions are chosen as

$$f_1(z) = 0.8 \cdot \mathsf{sgn}(z) \left(\frac{z}{2}\right)^\beta, \quad f_2(z) = 0.5 \cdot \mathsf{sgn}(z) \left(\frac{z}{2}\right)^\beta + 0.1 \cdot z, \quad f_3(z) = 1.5 \cdot \mathsf{sgn}(z) \left(\frac{z}{2}\right)^\beta.$$

All results in this section are averaged over 50 independent Monte Carlo trials.

Online MRC and regret control. We run Algorithm 2 to control the cumulative regret under $\beta = 1.5$ and $\beta = 2.5$, with randomly generated vectors v_1, v_2, v_3, and a total sample size of $n = 12{,}000$. We first present the online MRC estimation after each epoch of the bandit algorithm in Figure 3. As shown in the figure, Algorithm 2 consistently learns the index of each arm, demonstrating the dimension-reduction phenomenon. Thanks to the exponential epoch schedule used in our algorithm, the accuracy of the MRC estimator improves as the epoch number increases, enabling sharper regret control.

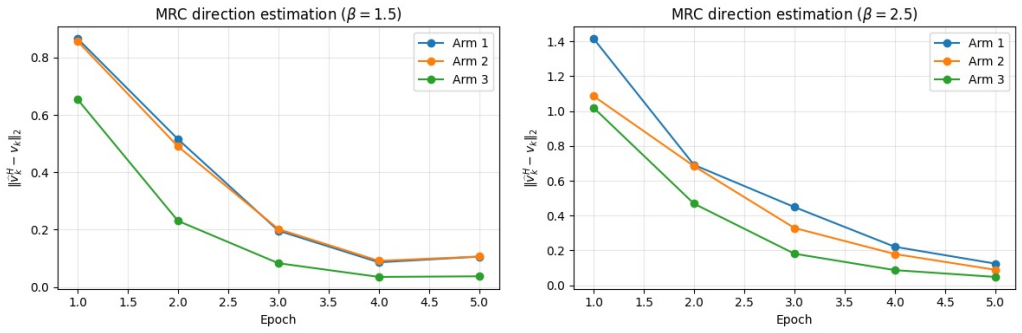

Fig 3: Performance of direction estimation by MRC in Algorithm 2. We show results for $\beta = 1.5$ and $\beta = 2.5$ for comparison. It can be observed that MRC consistently learns the direction before link functions are estimated.

The regret growth of Algorithm 2 is shown in Figure 4. For comparison, we also evaluate SmoothBandit, which is designed for general d-dimensional smooth nonparametric bandits [30, 39]. We report results for $\beta = 1.5$ and $\beta = 2.5$. As expected, our algorithm consistently

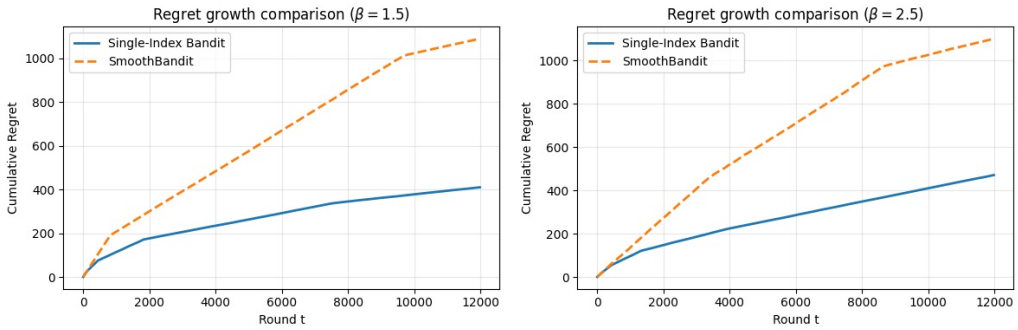

Fig 4: Regret growth of Algorithm 2 (Single-Index Bandit). Here, we set $\beta = 1.5$ and $\beta = 2.5$, and compare our algorithm with the SmoothBandit studied in [30, 39]. Our algorithm shows a significant advantage over SmoothBandit due to successfully learning the latent low-dimensional structure.

outperforms SmoothBandit due to effective dimension reduction. In SmoothBandit, the learning bandwidth is chosen as $h = \widetilde{O}\left(n_m^{-\frac{1}{2\beta+d}}\right)$, where each $n_m = \widetilde{O}\left(\left(\frac{1}{\epsilon_m^2}\right)^{\frac{2\beta+d}{2\beta}}\right)$. This rate is optimal for general nonparametric bandits [39], but is suboptimal when the underlying structure is single-index. Because SmoothBandit relies on an unnecessarily large epoch schedule and a suboptimal nonparametric estimator, it is unable to achieve optimal regret control under the single-index structure.

REMARK 3. Our Algorithm 2 requires maximizing the rank-correlation objective defined in (8). Although this MRC optimization problem is inherently non-convex, it is solved only once per epoch, resulting in a small total number of optimization steps. In particular, the number of epochs satisfies $M \lesssim \log n$ (a consequence of the exponential decay of ϵ_m; see the supplement for a detailed derivation), which keeps the overall computational cost manageable even for large n. In our simulations, we use differential evolution to optimize the MRC objective, which provides an efficient global search procedure in practice, as illustrated

Adaptive regret control. We evaluate the adaptive regret performance of Algorithm 4, in which the smoothness level is automatically estimated using only the bounds $\underline{\beta}$ and $\overline{\beta}$. We consider two settings, $\beta = 1.5$ and $\beta = 2.5$, with randomly generated v_1, v_2, v_3 and a total sample size of $n = 12{,}000$. For $\beta = 1.5$, we set $\underline{\beta} = 0.9$ and $\overline{\beta} = 1.9$; for $\beta = 2.5$, we set $\underline{\beta} = 1.9$ and $\overline{\beta} = 2.9$. Following [30], we compare the performance of Algorithm 4 with Algorithm 2 under a range of misspecified smoothness levels β'. The results for Algorithm 4 and for Algorithm 2 with varying β' are shown in Figure 5.

As illustrated in Figure 5, Algorithm 4 performs comparably to Algorithm 2 even when the latter is supplied with the true β. Moreover, its performance closely matches that of Algorithm 2 with a slightly misspecified $\beta' < \beta$, reflecting the intended undersmoothing-based adaptation. However, when β' becomes too small, the regret deteriorates as the cost of smoothness misspecification outweighs the benefits of adaptation. These patterns align with the observations in [30] for one-dimensional settings.

Funding. The research was supported in part by NSF Grant DMS-2413106 and NIH grants R01-GM129781 and R01-GM123056.

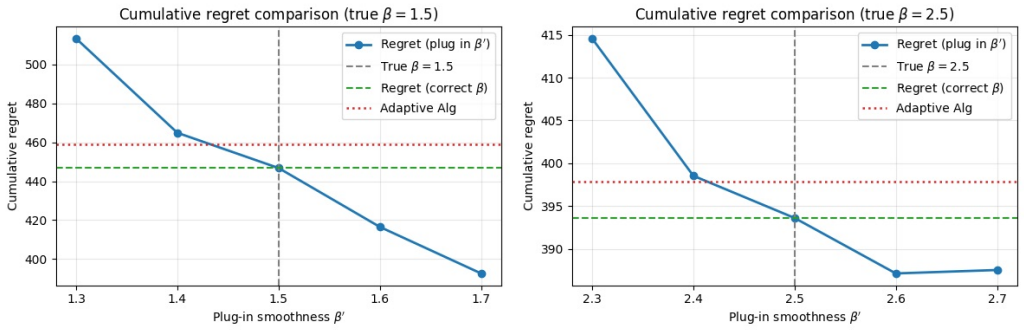

Fig 5: Adaptive regret control comparison. The results of our adaptive Algorithm 4 are denoted using vertical red dashed lines. Here, we vary the plug-in smoothness level β' between $[1.3, 1.7]$ for the true $\beta = 1.5$, $\underline{\beta} = 0.9, \overline{\beta} = 1.9$ (and $[2.3, 2.7]$ for true $\beta = 2.5$, $\underline{\beta} = 1.9, \overline{\beta} = 2.9$). Regrets of Algorithm 2 with plug-in β' are denoted by blue solid lines.

SUPPLEMENTARY MATERIAL

Supplementary Materials to " Nonparametric Bandits with Single-Index Rewards: Optimality and Adaptivity"

The supplementary file contains (i) a detailed proof of the nonasymptotic single-index regression bound and the corresponding minimax regret upper bound for single-index bandits; (ii) a constructive proof of the minimax lower bound, together with the proof of adaptivity results and other theorems; and (iii) additional propositions and technical lemmas that support the main results.

REFERENCES

[1] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

[2] ABREVAYA, J. and SHIN, Y. (2011). Rank estimation of partially linear index models. *The Econometrics Journal* **14** 409–437.

[3] AGRAWAL, S. and GOYAL, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning* 127–135. PMLR.

[4] ALQUIER, P. and BIAU, G. (2013). Sparse single-index model. *Journal of Machine Learning Research* **14**.

[5] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast Learning Rates for Plug-In Classifiers. *The Annals of Statistics* 608–633.

[6] AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing* **32** 48–77.

[7] BASTANI, H. and BAYATI, M. (2020). Online decision making with high-dimensional covariates. *Operations Research* **68** 276–294.

[8] BASTANI, H., BAYATI, M. and KHOSRAVI, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Management Science* **67** 1329–1349.

[9] BIETTI, A., BRUNA, J., SANFORD, C. and SONG, M. J. (2022). Learning single-index models with shallow neural networks. *Advances in neural information processing systems* **35** 9768–9783.

[10] BRILLINGER, D. R. (2011). A generalized linear model with a Gaussian regressor variables. In *Selected Works of David Brillinger* 589–606. Springer.

[11] BUBECK, S., CESA-BIANCHI, N. and KAKADE, S. M. (2012). Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory* 41–1. JMLR Workshop and Conference Proceedings.

[12] BUBECK, S., CESA-BIANCHI, N. et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* **5** 1–122.

[13] CAI, C., CAI, T. T. and LI, H. (2024). Transfer learning for contextual multi-armed bandits. *The Annals of Statistics* **52** 207–232.

[14] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics* **32** 1805–1840. https://doi.org/10.1214/009053604000000049

[15] CAI, T. T., LOW, M. G. and XIA, Y. (2013). Adaptive Confidence Intervals for Regression Functions Under Shape Constraints. *The Annals of Statistics* **41** 722–750.

[16] CAI, T. T. and PU, H. (2022). Stochastic continuum-armed bandits with additive models: Minimax regrets and adaptive algorithm. *The Annals of Statistics* **50** 2179–2204.

[17] CHANDRASHEKAR, A., AMAT, F., BASILICO, J. and JEBARA, T. (2017). Artwork Personalization at Netflix. https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76. Blog post.

[18] CHAPELLE, O. and LI, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems* **24** 2249–2257.

[19] CHEN, D., HALL, P. and MÜLLER, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics* **39** 1720–1747. https://doi.org/10.1214/11-AOS882

[20] CHU, W., LI, L., REYZIN, L. and SCHAPIRE, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* 208–214. JMLR Workshop and Conference Proceedings.

[21] DAMIAN, A., PILLAUD-VIVIEN, L., LEE, J. and BRUNA, J. (2024). Computational-Statistical Gaps in Gaussian Single-Index Models. In *The Thirty Seventh Annual Conference on Learning Theory* 1262–1262. PMLR.

[22] DENNIS COOK, R. (2000). SAVE: a method for dimension reduction and graphics in regression. *Communications in statistics-Theory and methods* **29** 2109–2121.

[23] DURAND, A., GELLY, S. and TEYTAUD, O. (2018). Contextual bandits for adaptive clinical trial design. *arXiv preprint arXiv:1806.03881*.

[24] ELSHIEWY, O., GUHL, D. and BOZTUĞ, Y. (2017). Multinomial logit models in marketing–from fundamentals to state-of-the-art. *Marketing: ZFP–Journal of Research and Management* **39** 32–49.

[25] FAN, Y., HAN, F., LI, W. and ZHOU, X.-H. (2020). On rank estimators in increasing dimensions. *Journal of Econometrics* **214** 379–412.

[26] GAIFFAS, S. and LECUÉ, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics* **1** 538–573.

[27] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *The Annals of Statistics* **38**.

[28] GOLDENSHLUGER, A. and ZEEVI, A. (2009). Woodroofe's One-Armed Bandit Problem Revisited. *The Annals of Applied Probability* 1603–1633.

[29] GOLDENSHLUGER, A. and ZEEVI, A. (2013). A linear response bandit problem. *Stochastic Systems* **3** 230–261.

[30] GUR, Y., MOMENI, A. and WAGER, S. (2022). Smoothness-adaptive contextual bandits. *Operations Research* **70** 3198–3216.

[31] GYÖRFI, L., KOHLER, M., KRZYZAK, A. and WALK, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

[32] HAN, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* **35** 303–316.

[33] HAN, F., JI, H., JI, Z. and WANG, H. (2017). A provable smoothing approach for high dimensional generalized regression with applications in genomics. *Electronic Journal of Statistics* **11** 4347–4403.

[34] HE, J., ZHANG, J. and ZHANG, R. Q. (2022). A reduction from linear contextual bandits lower bounds to estimations lower bounds. In *International Conference on Machine Learning* 8660–8677. PMLR.

[35] HENGARTNER, N. W. and STARK, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *The Annals of Statistics* 525–550.

[36] HOROWITZ, J. L. (1998). Semiparametric Methods in Econometrics. *Lecture Notes in Statistics*.

[37] HOROWITZ, J. L. (2009). *Semiparametric and nonparametric methods in econometrics* **12**. Springer.

[38] HRISTACHE, M., JUDITSKY, A. and SPOKOINY, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics* 595–623.

[39] HU, Y., KALLUS, N. and MAO, X. (2022). Smooth Contextual Bandits: Bridging the Parametric and Non-differentiable Regret Regimes. *Operations Research* **70** 3261–3281.

[40] ICHIMURA, H. (1991). Semiparametric least squares estimation of multiple index models: single equation estimation. *Nonparametric and semiparametric methods in econometrics and statistics* 3.

[41] ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58** 71–120.

[42] JIANG, R. and MA, C. (2025). Batched nonparametric contextual bandits. *IEEE Transactions on Information Theory*.

[43] KAKADE, S. M., KANADE, V., SHAMIR, O. and KALAI, A. (2011). Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems* **24**.

[44] KAUFMANN, E., CAPPÉ, O. and GARIVIER, A. (2012). On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics* 592–600. PMLR.
[45] KLOCK, T., LANTERI, A. and VIGOGNA, S. (2021). Estimating multi-index models with response-conditional least squares. *Electronic Journal of Statistics* **15** 589–629.
[46] KUCHIBHOTLA, A. K., PATRA, R. K. and SEN, B. (2023). Semiparametric efficiency in convexity constrained single-index model. *Journal of the American Statistical Association* **118** 272–286.
[47] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* **6** 4–22.
[48] LANTERI, A., MAGGIONI, M. and VIGOGNA, S. (2022). Conditional regression for single-index models. *Bernoulli* **28** 3051–3078.
[49] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.
[50] LEPSKI, O. and SERDYUKOVA, N. (2014). Adaptive estimation under single-index constraint in a regression model. *Annals of Statistics* **42** 1–28.
[51] LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics* 929–947.
[52] LI, K., YANG, Y. and NARISETTY, N. N. (2021). Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. *Electronic Journal of Statistics* **15** 5652–5695.
[53] LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327.
[54] LI, L., CHU, W., LANGFORD, J. and SCHAPIRE, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web* 661–670.
[55] LI, L., LU, Y. and ZHOU, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning* 2071–2080. PMLR.
[56] LOCATELLI, A. and CARPENTIER, A. (2018). Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory* 1463–1492. PMLR.
[57] LOW, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics* **25** 2547–2554.
[58] MANSKI, C. F. (1988). Identification of binary response models. *Journal of the American statistical Association* **83** 729–738.
[59] MAO, X., TANG, Z. and WANG, Y. (2025). On the Optimal Regret of Linear Contextual Bandit under Generalized Margin Condition. *Available at SSRN 5237709*.
[60] MCFADDEN, D. (1987). Regression-based specification tests for the multinomial logit model. *Journal of econometrics* **34** 63–82.
[61] NOVO, S., ANEIROS, G. and VIEU, P. (2019). Automatic and location-adaptive estimation in functional single-index regression. *Journal of Nonparametric Statistics* **31** 364–392.
[62] PERCHET, V. and RIGOLLET, P. (2013). The multi-armed bandit problem with covariates. *The Annals of Statistics* **41** 693–721.
[63] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Annals of statistics* 298–335.
[64] QIAN, W. and YANG, Y. (2016). Randomized allocation with arm elimination in a bandit problem with covariates. *Electronic Journal of Statistics* **10** 242–270.
[65] RAFFERTY, A. N., BRUNSKILL, E. and GRIFFITHS, T. L. (2011). Faster teaching by POMDP planning. In *Proceedings of the 15th international conference on artificial intelligence in education* 280–287.
[66] RIGOLLET, P. and ZEEVI, A. (2010). Nonparametric Bandits with Covariates. *COLT 2010* 54.
[67] SHERMAN, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society* 123–137.
[68] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics* 1040–1053.
[69] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer. https://doi.org/10.1007/b13794
[70] XIA, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association* **103** 1631–1640.
[71] XIA, Y. and HÄRDLE, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis* **97** 1162–1184.
[72] YUAN, G., XU, M., KPOTUFE, S. and HSU, D. (2023). Efficient estimation of the central mean subspace via smoothed gradient outer products. *arXiv preprint arXiv:2312.15469*.
[73] ZHAO, X., YAN, X., YU, A. and VAN HENTENRYCK, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel behaviour and society* **20** 22–35.