The Annals of Statistics
2024, Vol. 52, No. 6, 2557–2582
https://doi.org/10.1214/24-AOS2395

STATISTICAL COMPLEXITY AND OPTIMAL ALGORITHMS FOR NONLINEAR RIDGE BANDITS

BY NIVED RAJARAMAN[1,a], YANJUN HAN[2,d], JIANTAO JIAO[1,b] AND KANNAN RAMCHANDRAN[1,c]

[1]*Department of Electrical Engineering and Computer Sciences, University of California, Berkely,*
[a]*nived.rajaraman@berkeley.edu,* [b]*jiantao@eecs.berkeley.edu,* [c]*kannanr@eecs.berkeley.edu*
[2]*Courant Institute of Mathematical Sciences and Center for Data Science, New York University,* [d]*yanjunhan@nyu.edu*

We consider the sequential decision-making problem where the mean outcome is a nonlinear function of the chosen action. Compared with the linear model, two curious phenomena arise in nonlinear models: first, in addition to the "learning phase" with a standard parametric rate for estimation or regret, there is an "burn-in period" with a fixed cost determined by the nonlinear function; second, achieving the smallest burn-in cost requires new exploration algorithms. For a special family of nonlinear functions named ridge functions in the literature, we derive upper and lower bounds on the optimal burn-in cost, and in addition, on the entire learning trajectory during the burn-in period via differential equations. In particular, a two-stage algorithm that first finds a good initial action and then treats the problem as locally linear is statistically optimal. In contrast, several classical algorithms, such as UCB and algorithms relying on regression oracles, are provably suboptimal.

1. Introduction. A vast majority of statistical modeling studies data analysis in a setting where the underlying data-generating process is assumed to be stationary. In contrast, sequential data analysis assumes an iterative model of interaction, where the predictions of the learner can influence the data-generating distribution. An example of this observation model is clinical trials, which require designing causal experiments to answer questions about treatment efficacy under the presence of spurious and unobserved counterfactuals [9, 68]. Sequential data analysis presents novel challenges in comparison with data analysis with i.i.d. observations. One, in particular, is the "credit assignment problem," where value must be assigned to different actions when the effect of only a chosen action was observed [55, 65]. This is closely related to the problem of designing good "exploration" strategies and the necessity to choose diverse actions in the learning process [6, 66].

Another observation model involving sequential data is manipulation with object interaction, which represents one of the largest open problems in robotics [11]. Intelligently interacting with previously unseen objects in open-world environments requires generalizable perception, closed-loop vision-based control, and dexterous manipulation [42, 44, 76]. This requires designing good sequential decision rules that continuously collect informative data, and can deal with sparse and nonlinear reward functions and continuous action spaces.

In this paper, we study a sequential estimation problem as follows. At each time $t = 1, 2, \ldots, T$, the learner chooses an action a_t in a generic action set $\mathcal{A}$, based on the observed history $\mathcal{H}_{t-1} = \{(a_s, r_s)\}_{s \le t-1}$. Upon choosing a_t, the learner obtains a noisy observation of $f_{\theta^*}(a_t)$, denoted as $r_t = f_{\theta^\star}(a_t) + z_t$, where $\{f_\theta : \mathcal{A} \to \mathbb{R}\}_{\theta \in \Theta}$ is a given function class, and the noise z_t follows a standard normal distribution. Here $\theta^\star \in \Theta \subseteq \mathbb{R}^d$ is an unknown parameter fixed across time, and the learner's target is to estimate the parameter $\theta^\star$ in the

Received March 2023; revised January 2024.
MSC2020 subject classifications. Primary 62L12; secondary 62C20, 62K05.
Key words and phrases. Bandit problems, regret bounds, adaptive sampling, sequential estimation, ridge functions, minimax rate.

high dimensional regime where d could be comparable to T. Here the learner needs to both design the sequential experiment (i.e., actions $a_1, \dots, a_T$) adapted to the history $\{\mathcal{H}_{t-1}\}_{t=1}^T$, and output a final estimator $\widehat{\theta}_T = \widehat{\theta}_T(\mathcal{H}_T)$ which is close to $\theta^\star$.

In the bandit literature, the observation r_t is often interpreted as the *reward* obtained for picking the action a_t. In addition to estimating parameter $\theta^\star$, another common target of the learner is to maximize the expected cumulative reward $\mathbb{E}[\sum_{t=1}^T r_t]$, or equivalently, to minimize the *regret* defined as

$$\mathfrak{R}_T(\Theta, \mathcal{A}) = \mathbb{E}_{\theta^\star}\left[T \cdot \max_{a^\star \in \mathcal{A}} f_{\theta^\star}(a) - \sum_{t=1}^T f_{\theta^\star}(a_t) \right].$$

Compared with the estimation problem, the regret minimization problem essentially requires that every action a_t is close to the maximizer of $f_{\theta^\star}(\cdot)$.

Throughout this paper, we are interested in both the estimation and regret minimization problems for the class of *ridge functions* [54]. More specifically, we assume that:

1. The parameter set $\Theta = \mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\|_2 = 1\}$ is the unit sphere in $\mathbb{R}^d$;
2. The action set $\mathcal{A} = \mathbb{B}^d = \{a \in \mathbb{R}^d : \|a\|_2 \le 1\}$ is the unit ball in $\mathbb{R}^d$;
3. The mean reward is given by $f_{\theta^\star}(a) = f(\langle \theta^\star, a \rangle)$, where $f : [-1, 1] \to [-1, 1]$ is a known link function.

The form of ridge functions also corresponds to the single index model [37] in statistics. We will be interested in characterizing the following two complexity measures.

DEFINITION 1 (Sample Complexity for Estimation). For a given link function f, dimensionality d, and $\varepsilon \in (0, 1/2]$, the sample complexity of estimating $\theta^\star$ within accuracy ε is defined as

$$T^\star(f, d, \varepsilon) = \min\left\{ T : \inf_{\widehat{\theta}_T \in \mathbb{S}^{d-1}} \sup_{\theta^\star \in \mathbb{S}^{d-1}} \mathbb{E}_{\theta^\star}\left[1 - |\langle \widehat{\theta}_T, \theta^\star \rangle|\right] \le \varepsilon \right\}, \tag{1.1}$$

where the infimum is taken over all possible actions a^T adapted to $\{\mathcal{H}_{t-1}\}_{t=1}^T$ and all possible estimators $\widehat{\theta}_T = \widehat{\theta}_T(\mathcal{H}_T)$.

DEFINITION 2 (Minimax Regret). For a given link function f, dimensionality d, and time horizon T, the minimax regret is defined as

$$\mathfrak{R}_T^\star(f, d) = \inf_{a^T} \sup_{\theta^\star \in \mathbb{S}^{d-1}} \mathbb{E}_{\theta^\star}\left[T \cdot \max_{a^\star \in \mathcal{A}} f\big(\langle \theta^\star, a^\star \rangle\big) - \sum_{t=1}^T f\big(\langle \theta^\star, a_t \rangle\big) \right], \tag{1.2}$$

where the infimum is taken over all possible actions a^T adapted to $\{\mathcal{H}_{t-1}\}_{t=1}^T$.

In this paper, we are mainly interested in the scenario where the link function f is nonlinear. If f is linear, that is, $f(x) = \mathrm{id}(x) = x$, this is called the linear bandit, and it is known [53, 69] that

$$T^\star(\mathrm{id}, d, \varepsilon) \asymp \frac{d^2}{\varepsilon}, \quad \text{and} \quad \mathfrak{R}_T^\star(\mathrm{id}, d) \asymp \min\{d\sqrt{T}, T\}.$$

Here and throughout, the symbol $\asymp$ ignores all constant and polylogarithmic factors in $(T, d, 1/\varepsilon)$. However, even for many specific choices of nonlinear functions f, much less is known about the above quantities. One of our main contributions in this paper is to identify

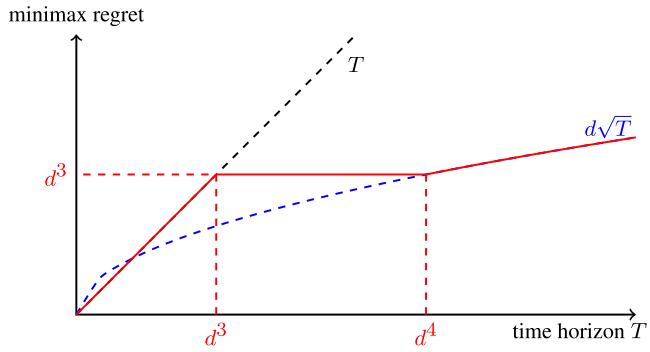

FIG. 1. *When $f(x) = x^3$ is the cubic function, the minimax regret scales as* $\min\{T, d^3 + d\sqrt{T}\}$ *(ignoring constant and polylogarithmic factors).*

a curious *phase transition* in the learning process for nonlinear link functions. Consider a toy example where $f(x) = \mathrm{cubic}(x) = x^3$. We will show that

$$T^\star(\mathrm{cubic}, d, \varepsilon) \asymp d^3 + \frac{d^2}{\varepsilon}, \quad \text{and} \quad \mathfrak{R}_T^\star(\mathrm{cubic}, d) \asymp \min\{d^3 + d\sqrt{T}, T\}.$$

A picture of the minimax regret as a function of T is displayed in Figure 1. We see that the minimax regret exhibits two elbows at $T \asymp d^3$ and $T \asymp d^4$: it grows linearly in T until $T \asymp d^3$, stabilizes for a long time during $d^3 \lesssim T \lesssim d^4$, and grows sublinearly in T in the end. Similarly, the sample complexity of achieving accuracy $\varepsilon = 1/2$ is already $\asymp d^3$, but improving the accuracy from $1/2$ to ε only requires $\asymp d^2/\varepsilon$ additional observations.

This curious scaling is better motivated by understanding the behavior of an optimal learner. At the beginning of the learning process, the learner has very little information about $\theta^\star$ and tries to find actions having at least a constant inner product with $\theta^\star$. Finding such actions are necessary for the learner to eventually be able to get sublinear regret. As we will discuss later, loosely speaking, finding a single such action is also *sufficient* to get a sublinear regret. In other words, there is an additional *burn-in period* in the learning process:

1. In the burn-in period, the learner aims to find a good *initial action* a_0 such that $\langle a_0, \theta^\star \rangle \geq \text{const}$ (say $1/2$);
2. After the burn-in period, the learner views the problem as a linear bandit and starts learning based on the good initial action a_0.

As will be apparent later, the learning phase is relatively easy and could be solved in a similar manner to linear bandits. However, both the complexity analysis and the algorithm design in the burn-in period could be challenging and are the main focus of this paper. This burn-in period is not unique to f being cubic and occurs for many choices of the link function.

Understanding the above burn-in period is important for nonlinear bandits due to two reasons. First, the burn-in period results in a fixed *burn-in cost* which is independent of T or ε. This burn-in cost could be the dominating factor of our sequential problem in the high-dimensional setting—in our toy example, the burn-in cost $\Theta(d^3)$ dominates the learning cost $\Theta(d\sqrt{T})$ as long as $T = O(d^4)$, which is a reasonable range of acceptable sample sizes. Second, the long burn-in period requires new ideas of *exploration* or *experimental design*, which is a central problem in the current era of reinforcement learning. As a result, understanding the burn-in period provides algorithmic insights on where to explore when the learner has not gathered enough information. Similar burn-in costs were also observed in [34, 38].

The main contributions of this paper are as follows:

- We identify the existence of the burn-in period for general nonlinear ridge bandit problems, and show that the two-stage algorithm which first finds a good initial action and then treats the problem as linear is near optimal for both parameter estimation and regret minimization.
- We prove lower bounds for both the burn-in cost and the learning trajectory during the burn-in period, via a novel application of information-theoretic tools to establishing minimax lower bounds for sequential problems, which could be of independent interest.
- We provide a new algorithm that achieves a small burn-in cost, and establish an upper bound on the learning trajectory during the burn-in period.
- We show that other ideas of exploration, including the UCB and oracle-based algorithms, are provably suboptimal for nonlinear ridge bandits. This is also the first failure example of UCB in a general and noisy learning environment.

Notation. For $d \in \mathbb{N}$, let $\mathbb{B}^d$ and $\mathbb{S}^{d-1}$ denote the unit ball and sphere in d dimensions, respectively. For $n \in \mathbb{N}$, let $[n] \triangleq \{1, 2, \dots, n\}$. For probability measures P and Q over the same probability space, let $D_{\mathrm{KL}}(P\|Q) = \int \mathrm{d}P \log(\mathrm{d}P/\mathrm{d}Q)$ and $\chi^2(P\|Q) = \int(\mathrm{d}P)^2/\mathrm{d}Q - 1$ be the Kullback-Leibler (KL) divergence and χ^2 divergence between P and Q, respectively. For a random vector $(X, Y) \sim P_{XY}$, let $I(X;Y) = D_{\mathrm{KL}}(P_{XY}\|P_X \otimes P_Y)$ be the mutual information between X and Y, where P_X, P_Y are the respective marginals. For nonnegative sequences $\{a_n\}$ and $\{b_n\}$, the following asymptotic notation will be used: let $a_n = O(b_n)$ denote $\limsup_{n\to\infty} a_n/b_n < \infty$, and $a_n = \widetilde{O}(b_n)$ (or $a_n \lesssim b_n$) denote $a_n = O(b_n \log^c n)$ for some $c > 0$. Moreover, $a_n = \Omega(b_n)$ (resp. $a_n = \widetilde{\Omega}(b_n)$ or $a_n \gtrsim b_n$) means $b_n = O(a_n)$ (resp. $b_n \lesssim a_n$), and $a_n = \Theta(b_n)$ (resp. $a_n = \widetilde{\Theta}(b_n)$ or $a_n \asymp b_n$) means both $a_n = O(b_n)$ and $b_n = O(a_n)$ (resp. $a_n \lesssim b_n \lesssim a_n$).

1.1. *Bounds on the burn-in cost*. This section provides upper and lower bounds on the burn-in cost, which we formally define below.

DEFINITION 3 (Burn-in Cost). For a given link function f and dimensionality d, the *burn-in cost* is defined as

$$T^\star_{\text{burn-in}}(f, d) = T^\star(f, d, 1/2),$$

where $T^\star$ is the sample complexity defined in Definition 1.

In other words, the burn-in cost is simply defined as the minimum amount of observations to achieve a constant correlation $\langle \widehat{\theta}, \theta^\star \rangle = \Omega(1)$. The constant $1/2$ in the definition is not essential and could be replaced by any constant bounded away from both 0 and 1. Next, we specify our assumptions on the link function f.

ASSUMPTION 1 (Regularity conditions for the burn-in period). We assume that the link function f satisfies the following conditions:

1. Normalized scale: $f(0) = 0$, $f(1) = 1$, and $|f| \le 1$;
2. Monotonicity: f is either (i) increasing on $[-1, 1]$; or (ii) even and increasing on $[0, 1]$.

We remark that Assumption 1 is very mild. The normalized scale is only for the scaling purpose. The monotonicity assumption ensures that $a = \theta^\star$ is a maximizer of $f(\langle a, \theta^\star \rangle)$, so that the task of regret minimization is aligned with the task of parameter estimation. Moreover, the additional benefit of the monotonicity condition during the burn-in period is that,

Algorithm 1: Iterative direction search algorithm

Input: link function f, dimensionality d, a noisy oracle $\mathcal{O} : a \in \mathbb{B}^d \mapsto \mathcal{N}(f(\langle \theta^\star, a\rangle), 1)$, error probability δ, target inner product $x_0 \in (0, 1)$.
Output: an action a_0 such that $\langle \theta^\star, a_0\rangle \geq x_0$ with probability at least $1 - \delta$.
Let numerical constants $(\kappa_1, \kappa_2, c_0, d_0)$ be given in Theorem 3.1.
Let $m \leftarrow \lceil x_0^2 d\rceil$, $V \leftarrow \{\mathbf{0}_d\}$, $L \leftarrow 2m \log(2m/\delta)/c_0$.
for *epoch* $i = 1, \ldots, d_0$ **do** `// Find initial few directions`
  while True **do**
    Sample $v \sim \mathrm{Unif}(V^\perp \cap \mathbb{S}^{d-1})$;
    if $\textit{INITIALACTIONHYPTEST}(v; f, d, \mathcal{O}, \delta/L, \kappa_1/4) = \textit{True}$ **then**
      $v_i \leftarrow v$; $V \leftarrow \mathrm{span}(V \cup \{v_i\})$; **break**;
for *epoch* $i = d_0 + 1, \ldots, m$ **do** `// Find subsequent directions`
  $v_{\mathrm{pre}} \leftarrow \frac{1}{\sqrt{i-1}} \sum_{j=1}^{i-1} v_j$; $x_{\mathrm{pre}} \leftarrow \sqrt{(i-1)/d}$;
  while True **do**
    Sample $v \sim \mathrm{Unif}(V^\perp \cap \mathbb{S}^{d-1})$;
    if $\textit{GOODACTIONHYPTEST}(v; f, d, \mathcal{O}, \delta/L, \kappa_1/4, \kappa_2, v_{\mathrm{pre}}, x_{\mathrm{pre}}) = \textit{True}$ **then**
      $v_i \leftarrow v$; $V \leftarrow \mathrm{span}(V \cup \{v_i\})$; **break**;
Output $a_0 \leftarrow \frac{1}{\sqrt{m}} \sum_{i=1}^m v_i$. `// Final action`

by querying the noisy values of $f(\langle a, \theta^\star\rangle)$, the learner could decide whether or not the inner product $\langle a, \theta^\star\rangle$ is improving. This turns out to be a crucial step in the algorithmic design.

Under Assumption 1, the next theorem provides an upper bound on the burn-in cost.

THEOREM 1.1 (Weaker version of Theorem 3.1). *In a ridge bandit problem with the link function f satisfying Assumption 1, for any $\kappa \in (0, 1/4)$, the following upper bound holds for the burn-in cost*:

$$T^\star_{\textit{burn-in}}(f, d) \lesssim d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \leq y \leq x} \min_{z \in [(1-\kappa)y, (1+\kappa)y]} [f'(z)]^2}, \tag{1.3}$$

with a hidden factor depending on κ. This is achieved by Algorithm 1 in Section 3.1.

We remark that the hidden constant does not depend on f, so Theorem 1.1 establishes an upper bound on the burn-in cost which is *pointwise* in f. Also, this upper bound depends on f through some integral involving the derivative of f, suggesting that the behavior of f at all points is important to determine the burn-in cost. We note that although Theorem 1.1 is stated in terms of the derivative f', in general we do not need to assume that f is differentiable, and our general result (cf. Theorem 3.1) is stated in terms of finite differences of f.

In the integral (1.3), the variable x captures the progress of the learner in terms of the inner product $\langle a_t, \theta^\star\rangle$, and therefore the upper and lower limits of the integral means that the inner product grows from $\Theta(1/\sqrt{d})$ to $\Theta(1)$. In addition, when κ is small, the integrand can be interpreted as the signal-to-noise ratio (SNR) witnessed by the learner

$$\max_{1/\sqrt{d} \leq y \leq x} \min_{z \in [(1-\kappa)y, (1+\kappa)y]} \big[f'(z)\big]^2 \approx \frac{1}{d} \max_{1/\sqrt{d} \leq y \leq x} \big[f(y) - f(y - 1/\sqrt{d})\big]^2. \tag{1.4}$$

Here $f(y) - f(y - 1/\sqrt{d})$ is the increment of the function value if the inner product $\langle a_t, \theta^\star\rangle$ changes by $1/\sqrt{d}$, and taking the maximum over $y \leq x$ corresponds to evaluating f at points offering the highest SNR below the current inner product $\langle a_t, \theta^\star\rangle = x$. The total burn-in cost

is naturally upper bounded by the integral (1.4) of real-time costs using the best available SNR. This intuition will become clearer when we characterize the learning trajectory during the burn-in period in Section 1.2.

The next theorem shows a lower bound on the burn-in cost in terms of a different integral.

THEOREM 1.2 (Weaker version of Theorem 2.1). *Suppose f is even or odd. In a ridge bandit problem with the link function f satisfying Assumption 1, the following lower bound holds for the burn-in cost: whenever $T^\star_{\text{burn-in}}(f,d) \le T$, then*

$$T^\star_{\text{burn-in}}(f,d) \gtrsim d \cdot \int_{\sqrt{c\log(T)/d}}^{1/2} \frac{\mathrm{d}(x^2)}{(f(x))^2}.$$

Here $c>0$ is an absolute constant independent of (f,d).

Again, the hidden constant in Theorem 1.2 also does not depend on f, so the above lower bound is also pointwise in f. However, ignoring the logarithmic factor in the lower limit, the specific form of the integrand is also different. Compared with Theorem 1.2 proves an upper bound $f(x)^2/d$ on the real-time SNR, which by the monotonicity of f is no smaller than the SNR lower bound in (1.4). It is an interesting question to close this gap, while we remark that even proving the above weaker SNR upper bound is highly nontrivial and possibly requires new information-theoretic ideas. We also conjecture the SNR lower bound in (1.4) is essentially tight, and we defer these discussions to Section 4.4.

We also note that the assumption that f is even or odd is only for the simplicity of presentation and not required in general. As will become clear in Theorem 2.1, the general lower bound simply replaces $f(x)$ in Theorem 1.2 by $g(x) := \max\{|f(x)|, |f(-x)|\}$.

EXAMPLE 1. For $f(x) = |x|^p$ with $p>0$ (or $f(x)=x^p$ for $p\in\mathbb{N}$), Theorems 1.1 and 1.2 show that

$$\max\{d, d^p\} \lesssim T^\star_{\text{burn-in}}(f,d) \lesssim \max\{d^2, d^p\}.$$

Therefore, the upper and lower bounds match unless $p\in(1,2)$. However, this does not cause any discrepancy for the overall sample complexity $T^\star(f,d,\varepsilon)$, as the sample complexity $\Theta(d^2/\varepsilon)$ in the learning phase will dominate the burn-in cost if $p\in(1,2)$. Therefore, in many scenarios, Theorems 1.1 and 1.2 are sufficient to give tight results on the sample complexity within logarithmic factors.

1.2. *Learning trajectory during the burn-in period.* In addition to the burn-in cost, which is the sample complexity required to achieve a constant inner product $\langle \widehat{\theta}_T, \theta^\star\rangle$, we can also provide a fine-grained analysis of the learning trajectory during the burn-in period. Specifically, we have the following definition.

DEFINITION 4 (Learning trajectory). For a given link function f, dimensionality d, and $\varepsilon\in(0,1/2]$, the burn-in cost for achieving ε inner product is defined as

$$T^\star_{\text{burn-in}}(f,d,\varepsilon) = T^\star(f,d,1-\varepsilon),$$

where $T^\star$ is the sample complexity defined in Definition 1. We will call the function $\varepsilon \mapsto T^\star_{\text{burn-in}}(f,d,\varepsilon)$ as the *minimax learning trajectory* during the burn-in period.

In other words, the learning trajectory concerns the sample complexity of achieving inner products $\langle \widehat{\theta}_T, \theta^\star\rangle \ge \varepsilon$, simultaneously for all $\varepsilon\in(0,1/2]$. The following theorem is a strengthening of Theorems 1.1 and 1.2 in terms of the learning trajectory.

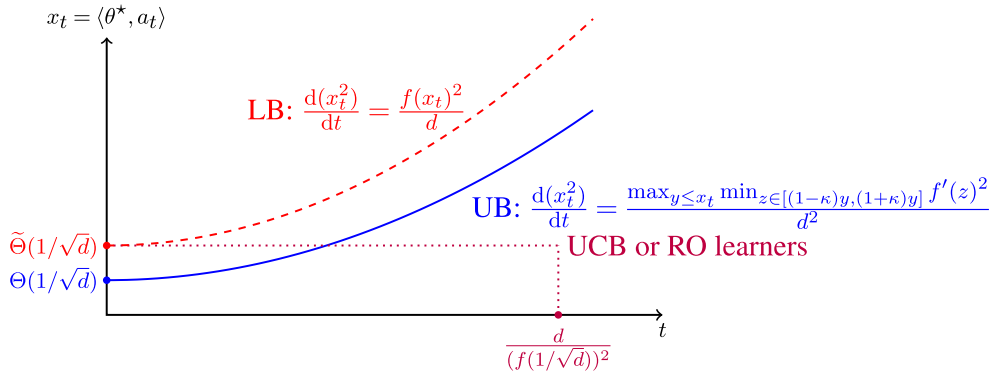

FIG. 2. *Upper and lower bounds on the minimax learning trajectory. Here UB stands for upper bound, LB stands for lower bound, and RO stands for regression oracles.*

THEOREM 1.3. *Consider a ridge bandit problem with a link function f satisfying Assumption 1. In what follows $\kappa \in (0, 1/4)$ is any fixed constant, and $c_1, c_2 > 0$ are absolute constants independent of (f, d, ε).*

- *For $\varepsilon \in [c_1/\sqrt{d}, 1/2]$, the following upper bound holds on the learning trajectory:*

$$T^\star_{\text{burn-in}}(f, d, \varepsilon) \lesssim d^2 \cdot \int_{1/\sqrt{d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \le y \le x} \min_{z \in [(1-\kappa)y, (1+\kappa)y]} [f'(z)]^2}.$$

- *In addition assume that f is even or odd. Then for $\varepsilon \in [\sqrt{c_2 \log(T)/d}, 1/2]$, the following lower bound holds on the learning trajectory: if $T^\star_{\text{burn-in}}(f, d, \varepsilon) \le T$, then*

$$T^\star_{\text{burn-in}}(f, d, \varepsilon) \gtrsim d \cdot \int_{\sqrt{c_2 \log(T)/d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{(f(x))^2}.$$

Theorem 1.3 shows that the integrals in Theorems 1.1 and 1.2 are not superfluous: when the target inner product changes from $1/2$ to ε, in the sample complexity we simply replace the upper limits of the integrals with ε as well. Note that in the above theorem we always assume that $\varepsilon \gtrsim 1/\sqrt{d}$, as a uniformly random action $a \in \mathbb{S}^{d-1}$ achieves $\langle a, \theta^\star \rangle = \Omega(1/\sqrt{d})$ with a constant probability, and thus the sample complexity for smaller ε is $\Theta(1)$. This result leads to a characterization of the learning trajectory using *differential equations* displayed in Figure 2. As a function of t, there is an algorithm where the inner product $x_t = \langle \theta^\star, a_t \rangle$ can start from $\Theta(1/\sqrt{d})$ and follow the differential equation shown in the blue solid line. Moreover, for every algorithm, with high probability the start point of $x_t = \langle \theta^\star, a_t \rangle$ cannot exceed $\widetilde{\Theta}(1/\sqrt{d})$, and the entire learning trajectory must lie below the differential equation shown in the red dashed line. The purple dotted line displays the performances of other exploration algorithms such as UCB and regression oracle (RO) based algorithms, showing that these algorithms make no progress until the time point $t \asymp d/[f(1/\sqrt{d})]^2$. This last part is the central theme of the next section.

1.3. *Suboptimality of existing exploration algorithms.* As we discussed in the introduction, learning in the burn-in period is essentially *exploration*, where the learner has not found a good action but aims to do so. In the literature of sequential decision making or bandits, several exploration ideas have been proposed and shown to work well for many problems. In this section, we review two well-known exploration algorithms, that is, algorithms based on upper confidence bounds (UCB) or regression oracles, and show that they can be strictly suboptimal for general ridge bandits.

1.3.1. *Eluder-UCB*. The UCB adopts a classical idea of "optimism in the face of uncertainty," that is, the algorithm maintains for each action an optimistic upper bound on its reward, and then chooses the action with the largest optimistic upper bound. The core of the UCB algorithm is the construction of the upper confidence bound, and the Eluder-UCB algorithm [60] proposes a general way to do so. In the Eluder-UCB algorithm specialized to ridge bandits, at each time t the learner computes the least squares estimate of $\theta^\star$ based on past history:

$$\widehat{\theta}_t^{\mathrm{LS}} := \operatorname*{arg\,min}_{\theta \in \mathbb{S}^{d-1}} \sum_{s<t} \big(r_s - f\big(\langle \theta, a_s \rangle\big)\big)^2.$$

Then using standard theory of least squares, one can show that the true parameter $\theta^\star$ belongs to the following confidence set $\mathbb{C}_t$ with high probability:

$$\mathbb{C}_t = \left\{ \theta \in \mathbb{S}^{d-1} : \sum_{s<t} \big(f\big(\langle a_s, \theta \rangle\big) - f\big(\langle a_s, \widehat{\theta}_t^{\mathrm{LS}} \rangle\big)\big)^2 \le \mathbf{Est}_t \right\}, \tag{1.5}$$

where $\mathbf{Est}_t \asymp d$ is an upper bound on the estimation error and known to the learner. Conditioned on the high probability event that $\theta^\star \in \mathbb{C}_t$, the quantity $\max_{\theta \in \mathbb{C}_t} f(\langle a, \theta \rangle)$ is an upper bound of $f(\langle a, \theta^\star \rangle)$ for every action a, and the Eluder-UCB algorithm chooses the action

$$a_t \in \operatorname*{arg\,max}_{a \in \mathcal{A}} \max_{\theta \in \mathbb{C}_t} f\big(\langle a, \theta \rangle\big). \tag{El-UCB}$$

If there are ties, they can be broken in an arbitrary manner. The next theorem presents a lower bound on the burn-in cost for the Eluder-UCB algorithm.

THEOREM 1.4. *For every Lipschitz link function f satisfying Assumption 1, there exists a tie-breaking rule for* (El-UCB) *such that for the Eluder-UCB algorithm, the following lower bound holds for its sample complexity $T^\star_{\mathrm{UCB}}$ of achieving inner product at least ε: whenever $T^\star_{\mathrm{UCB}} \le T$ and $\varepsilon \ge \sqrt{c \log(T)/d}$, it holds that*

$$T^\star_{\mathrm{UCB}} \gtrsim \frac{d}{g(\sqrt{c \log(T)/d})^2}.$$

For an absolute constant $c > 0$ independent of (f, d, ε), and $g(x) := \max\{|f(x)|, |f(-x)|\}$.

Compared with Theorem 1.2, the lower bound for the Eluder-UCB algorithm only depends on the function value of f at a single point $\widetilde{\Theta}(1/\sqrt{d})$, even for achieving an inner product $\widetilde{O}(1/\sqrt{d})$. Since f is monotone (cf. Assumption 1), the lower bound in Theorem 1.4 is always no smaller than the minimax lower bound in Theorem 1.2, and this gap could be arbitrarily large for carefully chosen f. Note that the lower bound in Theorem 1.4 is again pointwise in f, meaning that the suboptimality of the Eluder-UCB algorithm in ridge bandits is general.

1.3.2. *Regression oracle based algorithms*. Algorithms based on regression oracles follow a different idea: instead of observing the noisy observation $r_t = f(\langle \theta^\star, a_t \rangle) + z_t$, suppose the learner receives an estimate $\widehat{\theta}_t$ from an oracle treated as a black box. There are two types of such oracles:

- Online regression oracle: the oracle outputs $\widehat{\theta}_t$ at the beginning of time t which satisfies

$$\sum_{s \le t} \big(f\big(\langle \theta^\star, a_s \rangle\big) - f\big(\langle \widehat{\theta}_s, a_s \rangle\big)\big)^2 \le \mathbf{Est}_t^{\mathbf{On}} \tag{1.6}$$

with high probability, where $\mathbf{Est}_t^{\mathbf{On}} \asymp d$ is a known quantity.

- Offline regression oracle: the oracle outputs $\widehat{\theta}_t$ at the end of time t which satisfies

$$\sum_{s\le t}\big(f\big(|\langle\theta^\star, a_s\rangle|\big) - f\big(|\langle\widehat{\theta}_t, a_s\rangle|\big)\big)^2 \le \mathbf{Est}_t^{\mathbf{Off}} \tag{1.7}$$

with high probability, where $\mathbf{Est}_t^{\mathbf{Off}} \asymp d$ is a known quantity.

Under the oracle model, instead of observing $(a_1, r_1, a_2, r_2, \ldots)$, the learner only observes $(a_1, \widehat{\theta}_1, a_2, \widehat{\theta}_2, \ldots)$, where the learner has no control over $\{\widehat{\theta}_t\}$ except for the error bound (1.6) or (1.7). Note that the observational model can be reduced to an oracle model with the help of certain oracles $\widehat{\theta}_t = \widehat{\theta}_t(\{a_s, r_s\}_{s=1}^t)$, but the converse may not be true. Over the recent years, an interesting line of research in the bandit literature [28, 29, 32, 46, 63] is the development of learning algorithms under only the oracle models.

Despite the success of oracle models, we show that for ridge bandits, the oracle models could be strictly less powerful than the original observational model. In particular, *any* algorithm under the oracle model could have a suboptimal performance. The exact statement is summarized in the next theorem, where we call an oracle "proper" if we require that $\widehat{\theta}_t \in \mathbb{S}^{d-1}$ for every t, and "improper" otherwise.

THEOREM 1.5. *For every Lipschitz link function f satisfying Assumption 1, there exists improper online regression oracles satisfying (1.6) or proper offline regression oracles satisfying (1.7) such that: for any algorithm under the oracle model, its sample complexity $T^\star_{\mathrm{RO}}$ of achieving inner product at least ε satisfies: whenever $T^\star_{\mathrm{RO}} \le T$ and $\varepsilon \ge \sqrt{c\log(T)/d}$, then*

$$T^\star_{\mathrm{RO}} \gtrsim \frac{d}{g(\sqrt{c\log(T)/d})^2}.$$

For an absolute constant $c > 0$ independent of (f, d, ε), and $g(x) := \max\{|f(x)|, |f(-x)|\}$.

The lower bound in Theorem 1.5 again holds for every f, and is the same as the lower bound in Theorem 1.4. Therefore, for general link function f, every algorithm could only achieve a strictly suboptimal performance under the oracle model. Note that this result does *not* rule out the possibility that some algorithm based on a particular oracle has a smaller sample complexity than Theorem 1.5; instead, Theorem 1.5 only means that even if an algorithm works, its analysis cannot treat the oracle as a black box.

EXAMPLE 2. Consider again the example where $f(x) = |x|^p$ with $p > 0$ (or $f(x) = x^p$ for $p \in \mathbb{N}$). Theorems 1.4 and 1.5 shows that the Eluder-UCB or regression oracle based algorithms can only achieve a burn-in cost $\widetilde{\Omega}(d^{p+1})$, which is strictly suboptimal compared with Example 1 if $p > 1$. In particular, if $p \ge 2$, the suboptimality gap is as large as $\widetilde{\Omega}(d)$.

1.4. *Complexity of the learning phase.* Next, we proceed to understand the learning performance after a good initial action a_0 is found with $\langle\theta^\star, a_0\rangle \ge 1/2$. To this end, we need a few additional assumptions on the link function f.

ASSUMPTION 2 (Regularity conditions for the learning phase). The link function f is differentiable with derivative f', and locally linear on some interval $[1-\gamma, 1]$ around 1:

$$c_f \le \min_{x\in[1-\gamma,1]} f'(x) \le \max_{x\in[1-\gamma,1]} f'(x) \le C_f. \tag{1.8}$$

The local linearity condition may appear to be strong at the first sight, as it forces f to come close to being linear. The crucial feature of (1.8) is that we only require it for x bounded away

from zero, thus it does not help alleviate the challenge in the burn-in period. This assumption also holds for many link functions, such as $f(x) = |x|^p$ for any fixed $p > 0$.

The following theorem establishes an upper bound on the sample complexity and the regret in the learning phase. It essentially states that, every ridge bandit problem becomes a linear bandit given a good initial action, provided that Assumption 2 holds.

THEOREM 1.6. *Suppose the link function f satisfies Assumption 2, and the learner is given an action a_0 with $\langle \theta^\star, a_0 \rangle \geq 1 - 3\gamma/4$. Then for every $\varepsilon < \gamma$, the output $\widehat{\theta}_T$ of Algorithm 4 in Section 3.2 satisfies $\mathbb{E}[\langle \widehat{\theta}_T, \theta^\star \rangle] \geq 1 - \varepsilon$ with $T = O(\frac{d^2}{c_f^2 \varepsilon})$. Here the hidden constant depends only on γ. If in addition f satisfies Assumption 1, Algorithm 4 in Section 3.2 over a time horizon T achieves a cumulative regret*

$$\mathfrak{R}_T^\star(f, d) = O\left(\min\left\{ \frac{C_f}{c_f} d\sqrt{T}, T \right\} \right).$$

Ignoring the constants (γ, c_f, C_f), the sample complexity $O(d^2/\varepsilon)$ and regret $O(d\sqrt{T})$ match the counterparts for linear bandits. Combining Theorems 1.1 and 1.6, we have the following characterization for the overall sample complexity and regret of general ridge bandits.

COROLLARY 1.1. *Suppose the link function f satisfies Assumptions 1 and 2 (with $\gamma = 2/3$). Then for $\varepsilon < 1/2$ and any fixed $\kappa \in (0, 1/4)$,*

$$T^\star(f, d, \varepsilon) \lesssim d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \leq y \leq x} \min_{z \in [(1-\kappa)y, (1+\kappa)y]} [f'(z)]^2} + \frac{d^2}{c_f^2 \varepsilon},$$

$$\mathfrak{R}_T^\star(f, d) \lesssim \min\left\{ d^2 \cdot \int_{1/\sqrt{d}}^{1/2} \frac{\mathrm{d}(x^2)}{\max_{1/\sqrt{d} \leq y \leq x} \min_{z \in [(1-\kappa)y, (1+\kappa)y]} [f'(z)]^2} + \frac{C_f}{c_f} d\sqrt{T}, T \right\}.$$

For the lower bounds in the learning phase, we need an additional assumption on f.

ASSUMPTION 3 (Lower bound regularity condition). The function f is L-Lipschitz on $[-1, 1]$, that is, $|f(x) - f(y)| \leq L|x - y|$.

Compared with Assumption 2, Assumption 3 additionally requires that $f'(x)$ is upper bounded for x close to zero as well. It turns out that this assumption is also necessary for the lower bound to hold, in the sense that a super small regret might be possible without this assumption. See Example 3 at the end of this section for details. The lower bounds on the sample complexity and regret are summarized in the following theorem.

THEOREM 1.7. *Suppose the link function f satisfies Assumptions 2 and 3. Then for every $\varepsilon < 1/2$, the following minimax lower bounds hold*:

$$T^\star(f, d, \varepsilon) \geq \frac{c d^2}{\varepsilon}, \mathfrak{R}_T^\star(f, d) \geq c \min\{d\sqrt{T}, T\},$$

where $c > 0$ is an absolute constant depending only on (γ, c_f, L).

Combining Theorems 1.2 and 1.7, we have the following immediate corollary on the overall lower bounds for ridge bandits.

COROLLARY 1.2. *Suppose Assumptions 1, 2 and 3 hold. Then for $\varepsilon < 1/2$,*

$$T^\star(f, d, \varepsilon) \gtrsim \max_{T \geq 1} \min\left\{ d \cdot \int_{\sqrt{c\log(T)/d}}^{1/2} \frac{\mathrm{d}(x^2)}{(g(x))^2}, T \right\} + \frac{d^2}{\varepsilon},$$

$$\mathfrak{R}_T^\star(f, d) \gtrsim \min\left\{ d \cdot \int_{\sqrt{c\log(T)/d}}^{1/2} \frac{\mathrm{d}(x^2)}{(g(x))^2} + d\sqrt{T}, T \right\},$$

where $g(x) := \max\{|f(x)|, |f(-x)|\}$, and the hidden constants depend only on (c_f, L).

EXAMPLE 3. This example illustrates the importance of Assumption 3 for the minimax lower bound. Consider an odd function whose restriction on $[0, 1]$ is

$$f(x) = (1-\gamma) \cdot \mathbb{1}(\varepsilon < x \leq 1-\gamma) + x \cdot \mathbb{1}(1-\gamma < x \leq 1).$$

This function satisfies both Assumptions 1 and 2. However, we show that when ε is very small, one can achieve an $o(d\sqrt{T})$ regret in this case. The key insight is that the function is 0 on $[0, \epsilon]$ and $\geq 1-\gamma$ on the rest of the domain. Note that for each action a, by playing it $\widetilde{O}(1)$ times, the learner learns with high probability whether or not $\langle \theta^\star, a\rangle \leq \varepsilon$, and whether or not $\langle \theta^\star, a\rangle \geq -\varepsilon$. Now choosing $a \in \{\lambda e_1, \ldots, \lambda e_d\}$ and performing bisection search over $\lambda \in [0, 1]$, $\widetilde{O}(d\log(1/\varepsilon))$ observations suffice to estimate every θ_j within an additive error ε. Committing to this estimate then leads to a regret upper bound $\widetilde{O}(d\log(1/\varepsilon) + \varepsilon\sqrt{d}T)$, which could be much smaller than $\Theta(d\sqrt{T})$ for small ε.

Note that in this example, the lower bound on the burn-in cost in Theorem 1.2 is still tight. Concretely, Theorem 1.2 gives a lower bound $\Omega(d)$ for the burn-in cost, which matches (up to logarithmic factors) the above upper bound $\widetilde{O}(d\log(1/\varepsilon))$.

1.5. *Related work.*

1.5.1. *Sequential estimation, testing, and experimental design.* Sequential decision making has a long history in the statistics literature. In sequential estimation [13, 56] or testing [70], in addition to designing the estimator/test, the learner also needs to decide when to stop collecting more observations. In sequential experimental design, the goal is to decide whether and which experiment to conduct given the outcomes of the past experiments [12, 20, 59]. Our framework falls broadly in the class of these problems.

1.5.2. *Stochastic bandits.* The stochastic bandit problem has received significant research effort dating back to [35, 47]. Under the most general scenario $f_{\theta^\star} \in \mathcal{F}$ without any structural assumption on $\mathcal{F}$, it is well known that the minimax regret scales as $\Theta(\sqrt{|\mathcal{A}|T\log|\mathcal{F}|})$ for a finite action set $\mathcal{A}$ [7]. Several algorithms have then been proposed to reduce the computational complexity, either with a strong classification oracle [4, 25], or under a realizability condition (i.e., $\mathbb{E}[r(a)] = f_{\theta^\star}(a)$) with regression oracles [2, 28, 63].

Specializing to ridge bandits, the most canonical examples are linear bandits [21, 23], with a link function $f(x) = x$, and "generalized" linear bandits, with $0 < c_1 \leq |f'(\cdot)| \leq c_2$ everywhere. For both examples, the minimax regret is $\widetilde{\Theta}(d\sqrt{T})$ [1, 26, 61]. There are only a few recent work beyond generalized linear bandits. For Lipschitz and concave f, the same regret bound $\widetilde{\Theta}(d\sqrt{T})$ holds via a duality argument without an explicit algorithm [49]. For convex f, the special cases of $f(x) = x^2$ and $f(x) = x^p$ with $p \geq 2$ were studied in [38, 52], where the optimal regret scales as $\widetilde{\Theta}(\sqrt{d^p T})$. Note that this is the case where the parameter set Θ is assumed to be $\mathbb{B}^d$, a setting we discuss in Section 4.3.

We discuss [38, 52] in greater detail as they are closest to ours. In [52], the burn-in cost is not the dominating factor in the minimax regret, but an algorithm is designed for the burn-in period and inspires ours. In [38], although the authors noticed a burn-in cost in the analysis, it does not appear in the final regret bound as Θ is assumed to be the unit ball $\mathbb{B}^d$ instead of the unit sphere. In our work, we identify a fundamental role of the burn-in period in ridge bandits, by providing a lower bound on the burn-in cost and a learning trajectory during the burn-in period. In addition, we identify a sphere parameter set Θ as a more fundamental model to illustrate the phase transition, with the ball assumption being the hardest problem over spheres with different radii (cf. Section 4.3). We also remark that [38] proposes algorithms that work beyond ridge bandits, and proves the suboptimality of a noiseless UCB algorithm in a special example; instead, we focus on a smaller but more general problem of ridge bandits, and additionally shows that the failure of UCB is general even in the noisy scenario, answering a question of [52].

1.5.3. *Complexity measures for interactive decision making.* Several structural conditions have been proposed to unify existing approaches and prove achievability results for interactive decision making, such as the Eluder dimension [60] for bandits, and various quantities [24, 40, 41, 64, 71] for reinforcement learning. These quantities essentially work for generalized linear models and are not necessary in general [74, 75].

A very recent line of research tries to characterize the statistical complexity of interactive decision making, with both upper and lower bounds, based on either the decision-estimation coefficient (DEC) and its variants [19, 30–33], or the generalized information ratio [50, 51]. Although these result typically lead to the right regret dependence on T for general bandit problems, the dependence on d could be loose in both their upper and lower bounds. For example, the DEC lower bounds are proved via a careful two-point argument, which cannot take into account the estimation complexity, a quantity depending on d; this quantity is indeed the last missing piece in the state-of-the-art lower bound in [30]. The DEC upper bounds are achieved under an online regression oracle model, which by Theorem 1.5 must be suboptimal in ridge bandits. Our work complements this line of research by providing an in-depth investigation of the role of estimation complexity in interactive decision making, through the special case of ridge bandits.

1.5.4. *Information-theoretic view of sequential decision making.* The sequential decision making is also related to the notion of feedback channel capacity in information theory [18, 67], where the target is to transmit $\theta^\star$ through multiple access of some noisy channel with feedback. Upper bounds on mutual information $I(\theta^\star; \mathcal{H}_T)$ are sometimes useful in other contexts. A typical example is the stochastic optimization literature, where the goal is to maximize a function given access to the function and/or its gradient through some noisy oracle. The work [5, 57] initiated the use of the mutual information to prove the oracle complexity for stochastic optimization, while the key is the reduction to hypothesis testing problems where the classical arguments of Le Cam, Assouad, and Fano could all be applied; see, for example, [39, 62]. Instead, our problem illustrates the difficulty of applying classical hypothesis testing arguments to the sequential case, and requires the understanding of the entire trajectory $t \mapsto I(\theta^\star; \mathcal{H}_t)$ for a suitable notion of "information."

We also note that our problem is similar to the zeroth-order stochastic optimization. A rich line of work studied the maximization of a concave function [3, 16, 17, 27, 43, 48], while instance-dependent bounds are also developed for general Lipschitz functions [8, 15, 36]. However, in ridge bandits, the latter bounds do not exploit the specific structure and give a complexity exponential in d.

2. Minimax lower bounds. In this section, we prove the minimax lower bounds for general nonlinear ridge bandits. We only prove the lower bound of the burn-in cost, which is the most challenging part and requires novel information-theoretic techniques to handle interactive decision making. In particular, by recursively upper bounding a proper notion of information, we are able to prove fundamental limits for the learning trajectory of any learner at every time step. The proof of the lower bounds in Theorem 1.7 is deferred to Appendix B of the Supplementary Material [58], where the high-level argument is similar to existing lower bounds for linear bandits, but we additionally require a delicate exploration-exploitation tradeoff to make sure that the unknown parameter $\theta^\star$ lies on the unit sphere. The main lower bound on the learning trajectory during the burn-in period is summarized in the following theorem.

THEOREM 2.1. *Suppose the link function f satisfies Assumption 1, and define $g(x) := \max\{|f(x)|, |f(-x)|\}$. Given $c > 0$, $\delta \in (0,1)$, let $\{\varepsilon_t\}_{t\ge 1}$ be a sequence of positive reals defined recursively as follows*:

$$\varepsilon_1 = \sqrt{\frac{c\log(1/\delta)}{d}}, \qquad \varepsilon_{t+1}^2 = \varepsilon_t^2 + \frac{c}{d} g(\varepsilon_t)^2, \quad \forall t \ge 1. \tag{2.1}$$

There exists a universal constant $c > 0$ such that for any $\delta \in (0,1)$, if $\theta^\star$ is uniform distributed on $\mathbb{S}^{d-1}$, then for the above sequence $\{\varepsilon_t\}_{t\ge 1}$ and all $t \ge 1$, any learner satisfies that

$$\mathbb{P}\Big(\bigcap_{s\le t}\{|\langle \theta^\star, a_s\rangle| \le \varepsilon_s\}\Big) \ge 1 - t\delta.$$

Note that Theorem 2.1 provides a pointwise Bayes lower bound of the learning trajectory for *every* function f and *every* time step t. In other words, the sequence $\{\varepsilon_t\}_{t\ge 1}$ determines an upper limit on the entire learning trajectory $\{\langle \theta^\star, a_t\rangle\}_{t\ge 1}$ for every possible learner. In particular, the sequence $\{\varepsilon_t\}_{t\ge 1}$ is determined in a recursive manner, which is an interesting consequence of the interactive decision making environment.

By monotonicity of f, it holds that $0 \le \varepsilon_t \le \varepsilon$ implies $g(\varepsilon_t) \le g(\varepsilon)$, and the following corollary on the sample complexity follows directly from Theorem 2.1.

COROLLARY 2.1. *Fix $\varepsilon < 1/2$. For a large enough constant $c > 0$, the sample complexity of achieving $\mathbb{P}(\langle \theta^\star, a_T\rangle \ge \varepsilon) \ge 1 - T\delta$ is at least*

$$T = \Omega\Big(d \cdot \int_{\sqrt{c\log(1/\delta)/d}}^{\varepsilon} \frac{\mathrm{d}(x^2)}{g(x)^2}\Big) = \Omega\Big(d \cdot \max_{2\sqrt{\frac{c\log(1/\delta)}{d}} \le x \le \varepsilon} \frac{x^2}{g(x)^2}\Big). \tag{2.2}$$

Note that (2.2) proves Theorem 1.2 and the lower bound part of Theorem 1.3. For $f(x) = |x|^p$, this gives a lower bound $\widetilde{\Omega}(d^{\max\{p,1\}})$ for the burn-in cost, which is tight for $p \ge 2$. For $p = 1$ which corresponds to the case of linear bandit, an improved lower bound in Theorem 4.4 shows that a tight lower bound $\widetilde{\Omega}(d^2)$ actually holds; we defer the discussions (including the existing results for linear bandits) to Section 4.4.

In the remainder of this section, we will provide an information-theoretic proof of Theorem 2.1. In Section 2.1, we provide the intuition behind the update of the sequence $\{\varepsilon_t\}_{t\ge 1}$ in (2.1), and discuss the failure of formalizing the above intuition using the classical mutual information. Then we introduce in Section 2.2 the notion of χ^2-informativity and how it could lower bound the probability of error. In Section 2.3, we upper bound the χ^2-informativity in a recursive way and complete the proof of Theorem 2.1. We also remark that although one might be tempted to apply hypothesis testing based arguments to prove Theorem 2.1, we find it difficult to even obtain the second (weaker) lower bound in (2.2). We refer to the discussions below Theorems 4.1 and 4.2 for some insights.

2.1. *Information-theoretic insights.* In this section, we provide some intuition behind the sequence $\{\varepsilon_t\}_{t\geq 1}$ in (2.1). Let $\theta^\star \sim \mathsf{Unif}(\mathbb{S}^{d-1})$, and $I_t \triangleq I(\theta^\star; \mathcal{H}_t)$ be the mutual information between $\theta^\star$ and the learner's history $\mathcal{H}_t = \{(a_s, r_s)\}_{s\leq t}$ up to time t. It holds that

$$\begin{aligned} I_{t+1} &= I(\theta^\star; \mathcal{H}_{t+1}) \\ &\overset{\text{(a)}}{=} I(\theta^\star; \mathcal{H}_t) + I(\theta^\star; a_{t+1}, r_{t+1}|\mathcal{H}_t) \\ &\overset{\text{(b)}}{=} I_t + I(\theta^\star; r_{t+1}|\mathcal{H}_t, a_{t+1}) \\ &\overset{\text{(c)}}{\leq} I_t + \frac{1}{2}\log\big(1 + \mathbb{E}\big[f\big(\langle\theta^\star, a_{t+1}\rangle\big)^2\big]\big) \\ &\leq I_t + \frac{1}{2}\mathbb{E}\big[f\big(\langle\theta^\star, a_{t+1}\rangle\big)^2\big], \end{aligned} \tag{2.3}$$

where (a) follows from the chain rule of mutual information, (b) is due to the conditional independence of $\theta^\star$ and a_{t+1} conditioned on $\mathcal{H}_t$ that $I(\theta^\star; a_{t+1}|\mathcal{H}_t) = 0$, and (c) is the capacity upper bound for Gaussian channels. The above inequality shows that the mutual information increment $I_{t+1} - I_t$ is upper bounded by the second moment of $f(\langle\theta^\star, a_{t+1}\rangle)$, which is intuitive as larger correlation $\langle\theta^\star, a_{t+1}\rangle$ should lead to a larger information gain. For the lower bound purposes, we aim to show that $\langle\theta^\star, a_{t+1}\rangle$ should not be too large. The only thing we know about a_{t+1} is that it is constrained in information: by the data-processing inequality of mutual information, $I(\theta^\star; a_{t+1}) \leq I(\theta^\star; \mathcal{H}_t) = I_t$. Here comes our key insight behind the update (2.1): $I(\theta^\star; a) \leq d\varepsilon^2$ implies that $|\langle\theta^\star, a\rangle| \leq \varepsilon$ with high probability. Plugging this insight back into the recursion (2.3) of mutual information leads to (recall that $g(x) = \max_{|z|\leq x}|f(z)|$)

$$d\varepsilon_{t+1}^2 \leq d\varepsilon_t^2 + \frac{g(\varepsilon_t)^2}{2},$$

which takes the same form as the update (2.1).

This insight is motivated by the following geometric calculation: if a is uniformly distributed on the spherical cap $\{a \in \mathbb{S}^{d-1} : \langle\theta^\star, a\rangle \geq \varepsilon\}$, then $I(\theta^\star; a) \asymp d\varepsilon^2$. However, the classical notion of the mutual information does not guarantee that this insight holds with a sufficiently high probability: the celebrated Fano's inequality (cf. Lemma A.5 of the Supplementary Material [58]) tells that

$$\mathbb{P}\big(|\langle\theta^\star, a\rangle| \leq \varepsilon\big) \geq 1 - \frac{I(\theta^\star; a) + \log 2}{c_0 d\varepsilon^2}, \tag{2.4}$$

for some absolute constant $c_0 > 0$. In other words, the probability of failure (i.e., $\langle\theta^\star, a\rangle > \varepsilon$) could be as large as $I(\theta^\star; a)/(d\varepsilon^2)$, which is insufficient as T could be much larger than d. Fano's inequality (2.4) is also tight in the worst case: conditioned on $\theta^\star \sim \mathsf{Unif}(\mathbb{S}^{d-1})$, take $a \sim \mathrm{Unif}(\{a \in \mathbb{S}^{d-1} : \langle\theta^\star, a\rangle \geq \varepsilon\})$ with probability $p \asymp I/(d\varepsilon^2)$ and $a \sim \mathsf{Unif}(\mathbb{S}^{d-1})$ with probability $1 - p$. Here $\mathbb{P}(|\langle\theta^\star, a\rangle| > \varepsilon) \asymp p$ and $I(\theta^\star; a) \asymp p \cdot (d\varepsilon^2) \asymp I$, and (2.4) is tight.

As a result, although the mutual information provides the correct intuition for the recursion in (2.1), the potentially large failure probability in Fano's inequality (2.4) prevents us from making the intuition formal. In the subsequent sections, we will find a proper notion of information such that:

1. it leads to a much smaller (e.g., exponential in d) failure probability in (2.4);
2. it satisfies an approximate chain rule such that the information recursion (2.3) still holds.

2.2. *χ^2-informativity.* In this section, we introduce a new notion of information which satisfies the above two properties. For a pair of random variables (X, Y) with a joint distribution P_{XY}, the *χ^2-informativity* [22] between X and Y is defined as

$$I_{\chi^2}(X;Y) \triangleq \inf_{Q_Y} \chi^2(P_{XY} \| P_X \times Q_Y), \tag{2.5}$$

where $\chi^2(P\|Q) = \int (\mathrm{d}P)^2/\mathrm{d}Q - 1$ is the χ^2-divergence. Note that when the χ^2-divergence is replaced by the Kullback-Leibler (KL) divergence, the expression in (2.5) exactly becomes the classical mutual information. Moreover, note that $I_{\chi^2}(X;Y) \neq I_{\chi^2}(Y;X)$ in general.

In the sequel, we shall also need the following notion of *conditional χ^2-informativity*: for any measurable subset $E \subseteq \mathcal{X} \times \mathcal{Y}$, the χ^2-informativity conditioned on E is defined as

$$I_{\chi^2}(X;Y|E) \triangleq \inf_{Q_Y} \chi^2(P_{XY|E} \| P_X \times Q_Y). \tag{2.6}$$

The main advantage of the (conditional) χ^2-informativity lies in the following lemma, which is reminiscent of the Fano's inequality in (2.4).

LEMMA 2.1. *Let the random vector $(\theta^\star, a)$ satisfy that $\theta^\star \sim \mathsf{Unif}(\mathbb{S}^{d-1})$, and a is supported on $\mathbb{B}^d$. For every $\varepsilon > 0$ and every event E of $(\theta^\star, a)$, it holds that*

$$\mathbb{P}\big(|\langle \theta^\star, a\rangle| \le \varepsilon | E\big) \ge 1 - c_1 e^{-c_0 d\varepsilon^2} \sqrt{I_{\chi^2}(\theta^\star; a|E) + 1},$$

where $c_0, c_1 > 0$ are absolute constants.

A proof of this lemma is discussed in Appendix B.1 of the Supplementary Material [58].

Compared with Fano's inequality in (2.4), the probability of error in Lemma 2.1 depends exponentially in $d\varepsilon^2$ and is thus sufficiently small, which enables us to apply a union bound argument. However, the χ^2-informativity does not satisfy the chain rule or subadditivity (i.e., $I_{\chi^2}(X;Y,Z) \le I_{\chi^2}(X;Y) + I_{\chi^2}(X;Z|Y)$ may not hold), which makes it difficult to upper bound $I_{\chi^2}(\theta^\star; \mathcal{H}_t)$ in the same manner as (2.3). This is the place where conditioning on a suitable event E helps, and is the main theme of the next section.

2.3. *Upper bounding the χ^2-informativity.* As we have discussed in the previous section, the χ^2-informativity does not satisfy the chain rule or subadditivity. In this section, we establish a key lemma which upper bounds the χ^2-informativity in a recursive manner via a proper conditioning.

Let $\mathcal{H}_t = \{(a_s, r_s)\}_{s \le t}$ be the learner's history up to time t, and $E_t = \bigcap_{s \le t} \{|\langle \theta^\star, a_s\rangle| \le \varepsilon_s\}$ be the target event with $\{\varepsilon_t\}_{t \ge 1}$ defined in (2.1). The following lemma establishes a recursive relationship between the conditional χ^2-informativity.

LEMMA 2.2. *For $t \ge 1$ and any prior distribution of $\theta^\star$, it holds that*

$$I_{\chi^2}(\theta^\star; \mathcal{H}_t | E_t) + 1 \le \frac{\exp(g(\varepsilon_t)^2)}{\mathbb{P}(E_t|E_{t-1})^2} \big(I_{\chi^2}(\theta^\star; \mathcal{H}_{t-1}|E_{t-1}) + 1\big).$$

The proof of this result is discussed in Appendix B.2 in the Supplementary Material [58].

The desired lower bound in Theorem 2.1 follows by combining Lemma 2.2 and Lemma 2.1 together to result on a bound on $\mathbb{P}(|\langle \theta^\star, a_{t+1}\rangle| \le \varepsilon_{t+1} | E_t)$ in terms of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{t+1}$. By choosing the ε_t's to scale as in (2.1), we arrive at the bound

$$\mathbb{P}(E_{t+1}) \ge \mathbb{P}(E_t) - c_1 \exp\big(-c_0 d\varepsilon_1^2\big) = \mathbb{P}(E_t) - c_1 \exp\big(-cc_0 \log(1/\delta)\big) \ge \mathbb{P}(E_t) - \delta,$$

and therefore by union bound, $\mathbb{P}(E_t) \ge 1 - t\delta$. A formal argument is discussed in Appendix B.4 in the Supplementary Material [58].

3. Algorithm design. In this section, we propose an algorithm for the ridge bandit problem and prove the upper bounds in Theorems 1.1 and 1.6. The algorithm consists of two stages. First, in Section 3.1, we introduce an algorithm based on iterative direction search which finds a good initial action a_0 with $\langle \theta^\star, a_0 \rangle \geq x_0$ for a given target level $x_0 \in (0, 1)$; this algorithm could even be made agnostic to the knowledge of f. Based on this action, we proceed with a different regression-based algorithm in Section 3.2 for the learning phase. For the ease of presentation, we assume that f in monotone on $[-1, 1]$ in the above sections, and the case where f is even is deferred to Appendix C of the Supplementary Material [58] with slight algorithmic changes.

3.1. *Algorithm for the burn-in period.* Recall that the ultimate target in the burn-in period is to find an action a_0 which satisfies $\langle \theta^\star, a_0 \rangle \geq x_0$ with high probability. Our algorithmic idea is simple: if we could find $m := \lceil x_0^2 d \rceil$ orthonormal vectors $v_1, v_2, \ldots, v_m$ with $\langle \theta^\star, v_i \rangle \geq 1/\sqrt{d}$ for all $i \in [m]$, then $a_0 := m^{-1/2} \sum_{i=1}^m v_i$ is a unit vector with $\langle \theta^\star, a_0 \rangle \geq \sqrt{m/d} = x_0$. Finding these actions are not hard: Lemma A.1 in the Supplementary Material [58] shows that if $v_i \sim \mathsf{Unif}(\mathbb{S}^{d-1} \cap \mathrm{span}(v_1, \ldots, v_{i-1})^\perp)$, then with a constant probability it holds that $\langle \theta^\star, v_i \rangle \geq 1/\sqrt{d}$. The main difficulty lies in the *certification* of $\langle \theta^\star, v_i \rangle \geq 1/\sqrt{d}$, where we aim to make both Type I and Type II errors negligible for the following hypothesis testing problem:

$$H_0 : \langle \theta^\star, v_i \rangle < \frac{1}{\sqrt{d}} \quad \text{v.s.} \quad H_1 : \langle \theta^\star, v_i \rangle \geq \frac{1+\kappa_1}{\sqrt{d}}.$$

Here $\kappa_1 > 0$ is a small constant to be chosen later in Theorem 3.1. The key ingredient of our algorithm is to find such a test which makes good use of the historic progress $(v_1, \ldots, v_{i-1})$.

The detailed algorithm is summarized in Algorithm 1. The algorithm runs in two stages, and calls two certification algorithms InitialActionHypTest and GoodActionHypTest as subroutines for the respective stages. In each of the m epochs at both stages, a uniformly random direction $v_i \sim \mathsf{Unif}(\mathbb{S}^{d-1} \cap \mathrm{span}(v_1, \ldots, v_{i-1})^\perp)$ orthogonal to the past directions is chosen, and the difference lies in how we decide whether to accept v_i or not, that is, the certification of v_i. Concretely, each epoch aims to achieve the following two targets:

- With a constant probability, the certification algorithm accepts a random v_i. This leads to a small number of trials in each loop and a small overall sample complexity.
- Whenever the certification algorithm accepts v_i, then with high probability we have $\langle \theta^\star, v_i \rangle \in [1/\sqrt{d}, (1+\kappa_1)/\sqrt{d}]$. This leads to the correctness of the algorithm. (In principle, we only need the lower bound, and the upper bound is mainly for technical convenience.)

The initial stage consists of the first d_0 epochs, and uses a simple certification algorithm InitialActionHypTest displayed in Algorithm 2. The recursive stage consists of the rest of the epochs, and the certification algorithm GoodActionHypTest in Algorithm 3 exploits the current progress, including a good direction v_{pre} and an estimate x_{pre} of the inner product: we will show that $\langle \theta^\star, v_{\mathrm{pre}} \rangle \in [x_{\mathrm{pre}}, (1+\kappa_1) x_{\mathrm{pre}}]$ in every epoch. The certification algorithms will be detailed in the next few subsections, and they aim to collect as few as samples to reliably solve the hypothesis testing problem.

The performance of Algorithm 1 is summarized in the following theorem.

THEOREM 3.1. *Let $\delta \in (0, 1/2)$. Suppose that $\kappa_1 \in (0, (x_0^{-1} - 1)/2)$, $\kappa_2 \in (0, 1/4)$, and*

$$d_0 = \left\lceil \frac{(2\kappa_1 + 4)^2 \kappa_2 (2 - \kappa_2)}{\kappa_1^2 (1 - \kappa_2)^2} \right\rceil + 1, \qquad c_0 = c\left(1 + \frac{\kappa_1}{4}, 1 + \frac{\kappa_1}{2}, 1 - x_0^2, \frac{1 - x_0}{2}\right),$$

Algorithm 2: INITIALACTIONHYPTEST$(v; f, d, \mathcal{O}, \delta, \kappa_1)$

Input: link function f, dimensionality d, a noisy oracle $\mathcal{O} : a \in \mathbb{B}^d \mapsto \mathcal{N}(f(\langle\theta^\star, a\rangle), 1)$, error probability δ, accuracy parameter κ_1, test direction v.

Output: with probablity $\geq 1-\delta$, True if $\langle\theta^\star, v\rangle \in [(1+\kappa_1)/\sqrt{d}, (1+2\kappa_1)/\sqrt{d}]$, False if $\langle\theta^\star, v\rangle \notin [1/\sqrt{d}, (1+3\kappa_1)/\sqrt{d}]$.

Define

$$\varepsilon := \frac{1}{2} \min_{z\in[1,1+2\kappa_1]} \left| f\left(\frac{z+\kappa_1}{\sqrt{d}}\right) - f\left(\frac{z}{\sqrt{d}}\right)\right|. \tag{3.1}$$

Query the test action $2\log(2/\delta)/\varepsilon^2$ times and compute the sample average $\overline{r}$;

Return True if $\exists x \in [(1+\kappa_1)/\sqrt{d}, (1+2\kappa_1)/\sqrt{d}]$ such that $|\overline{r} - f(x)| \leq \varepsilon$, and False otherwise.

where the function $c(\cdot)$ appears in Lemma A.1 *of the Supplementary Material* [58]. *Let $\{\varepsilon_i\}_{i\geq 0}$ be a set of positive reals defined by*

$$\varepsilon_i = \begin{cases} \dfrac{1}{2} \displaystyle\min_{z\in[1/\sqrt{d},(1+\kappa_1/2)/\sqrt{d}]} \left| f\left(z + \frac{c_1}{\sqrt{d}}\right) - f(z)\right| & \text{if } 1 \leq i \leq d_0, \\ \dfrac{1}{2} \displaystyle\max_{c_2/\sqrt{d}\leq y\leq(1-\kappa_2)\sqrt{(i-1)/d}} \min_{z\in[(1-\kappa_1)y,(1+\kappa_1)y]} \left| f\left(z + \frac{c_1}{\sqrt{d}}\right) - f(z)\right| & \text{if } d_0+1 \leq i \leq m, \end{cases}$$

where $c_1 = \kappa_1\sqrt{1-(1-\kappa_2)^2}/4$, $c_2 = (2\kappa_1+4)\sqrt{1-(1-\kappa_2)^2}/\kappa_1$ are numerical constants determined by (κ_1, κ_2), and $m = \lceil x_0^2 d\rceil$.

Algorithm 3: GOODACTIONHYPTEST$(v; f, d, \mathcal{O}, \delta, \kappa_1, \kappa_2, v_{\text{pre}}, x_{\text{pre}})$

Input: link function f, dimensionality d, a noisy oracle $\mathcal{O} : a \in \mathbb{B}^d \mapsto \mathcal{N}(f(\langle\theta^\star, a\rangle), 1)$, error probability δ, accuracy parameters (κ_1, κ_2), test direction v, previous action v_{pre}, previous inner product x_{pre}.

Output: with probability $\geq 1-\delta$, True if $\langle\theta^\star, v\rangle \in [(1+\kappa_1)/\sqrt{d}, (1+2\kappa_1)/\sqrt{d}]$, False if $\langle\theta^\star, v\rangle \notin [1/\sqrt{d}, (1+3\kappa_1)/\sqrt{d}]$.

Define $\kappa_2^\perp := \sqrt{1-(1-\kappa_2)^2}$, $\kappa_3 := \kappa_1\kappa_2^\perp$, $\kappa_4 := (\kappa_1^{-1}+2)\kappa_2^\perp$, and

$$\varepsilon := \frac{1}{2} \max_{\kappa_4/\sqrt{d}\leq y\leq(1-\kappa_2)x_{\text{pre}}} \min_{z\in[(1-4\kappa_1)y,(1+4\kappa_1)y]} \left| f\left(z + \frac{\kappa_3}{\sqrt{d}}\right) - f(z)\right|. \tag{3.2}$$

Let $y^\star$ be the maximizer of (3.2), and define $\lambda := y^\star/[(1-\kappa_2)x_{\text{pre}}] \in [0,1]$.

Query both actions $a_- = \lambda(1-\kappa_2)v_{\text{pre}} - \kappa_2^\perp v$ and $a_+ = \lambda(1-\kappa_2)v_{\text{pre}} + \kappa_2^\perp v$ for $2\log(4/\delta)/\varepsilon^2$ times, and compute the sample averages $\overline{r}_-$ and $\overline{r}_+$.

Return True if $\exists z \in [y^\star, (1+3\kappa_1)y^\star]$ and $x \in [(1+\kappa_1)\kappa_2^\perp/\sqrt{d}, (1+2\kappa_1)\kappa_2^\perp/\sqrt{d}]$ such that

$$|\overline{r}_- - f(z-x)| \leq \varepsilon \text{ and } |\overline{r}_+ - f(z+x)| \leq \varepsilon, \tag{3.3}$$

and False otherwise.

If f is monotone on $[-1, 1]$, then with probability at least $1 - \delta$, Algorithm 1 outputs an action a_0 with $\langle \theta^\star, a_0 \rangle \geq x_0$ using at most

$$O\left(\log^2\left(\frac{d}{\delta}\right)\sum_{i=1}^{m}\frac{1}{\varepsilon_i^2}\right)$$

queries, where the hidden constant depends only on $(x_0, \kappa_1, \kappa_2)$.

By Lemma C.1 of the Supplementary Material [58], Theorem 3.1 implies the integral form of Theorem 1.1. Moreover, an inspection of the proof reveals that using $\widetilde{O}(\sum_{i=1}^{k}\varepsilon_i^{-2})$ samples gives an action a_k with $\langle \theta^\star, a_k \rangle \geq \sqrt{k/d}$, and this implies the learning trajectory upper bound in Theorem 1.3.

The remainder of this section is organized as follows. In Sections 3.1.1 and 3.1.2, we detail the certification algorithms INITIALACTIONHYPTEST and GOODACTIONHYPTEST, and analyze their performances. Section 3.1.3 modifies Algorithm 1 to make it *agnostic* to the knowledge of f, such that the same upper bound in Theorem 3.1 could be achieved by an algorithm without the knowledge of f. The proofs of the correctness and the sample complexity upper bounds for these algorithms are deferred to the Supplementary Material [58].

3.1.1. *Certifying initial directions.* The INITIALACTIONHYPTEST algorithm certifying the quality of the initial directions v is displayed in Algorithm 2. The idea is simple and requires nothing from the past: we query the test action v multiple times to obtain an accurate estimate of $f(\langle \theta^\star, v \rangle)$, and apply a projection based test to see if the inner product $\langle \theta^\star, v \rangle$ lies in the target interval. Here the parameter κ_1 represents the target accuracy for certification. The performance of this test is summarized in the following lemma.

LEMMA 3.1. *Suppose f is monotone on $[-1, 1]$. Then with probability at least $1 - \delta$, the* INITIALACTIONHYPTEST *algorithm outputs*:

- True *if* $\langle \theta^\star, v \rangle \in [(1 + \kappa_1)/\sqrt{d}, (1 + 2\kappa_1)/\sqrt{d}]$;
- False *if* $\langle \theta^\star, v \rangle \notin [1/\sqrt{d}, (1 + 3\kappa_1)/\sqrt{d}]$.

3.1.2. *Certifying subsequent directions.* In principle, certifying subsequent directions can also use the INITIALACTIONHYPTEST algorithm, but this may lead to a suboptimal sample complexity. The central question we answer in this section is as follows: *Given an action v_{pre} with a known estimate x_{pre} for the inner product $\langle \theta^\star, v_{\text{pre}} \rangle \approx x_{\text{pre}}$, can we certify the test direction v with a smaller sample complexity?*

Recall the simple idea of the INITIALACTIONHYPTEST algorithm: by querying the action v, we estimate the value of $f(\langle \theta^\star, v \rangle)$, and then certify the value of the inner product $\langle \theta^\star, v \rangle$. Our new observation is that, if we could estimate the value of $f(x + \langle \theta^\star, v \rangle)$ for a known x, then we could certify the value of $\langle \theta^\star, v \rangle$ as well. Since the propagation from the estimation error of $f(\langle \theta^\star, v \rangle)$ to that of $\langle \theta^\star, v \rangle$ depends on the derivative of f, such a translation by x could lead to a better derivative and benefit the certification step. This intuition leads to the following GOODACTIONHYPTEST algorithm displayed in Algorithm 3.

In Algorithm 3, instead of directly querying the test direction v, we query two actions based on the current progress: for some $\lambda \in [0, 1]$ to be chosen later, pick

$$a_- = \lambda(1 - \kappa_2)v_{\text{pre}} - \kappa_2^{\perp} v, \qquad a_+ = \lambda(1 - \kappa_2)v_{\text{pre}} + \kappa_2^{\perp} v.$$

Here, the parameter $\kappa_2 \in (0, 1)$ controls the range of the center x, and $\kappa_2^{\perp} := \sqrt{1 - (1 - \kappa_2)^2}$. Since $v \perp v_{\text{pre}}$ and $\lambda \in [0, 1]$, both actions lie in $\mathbb{B}^d$. By querying these actions for multiple

times, we obtain accurate estimates of $f(\lambda(1-\kappa_2)\langle\theta^\star, v_{\text{pre}}\rangle \pm \kappa_2^{\perp}\langle\theta^\star, v\rangle)$. As $\langle\theta^\star, v_{\text{pre}}\rangle \approx x_{\text{pre}}$, this corresponds to the shift of the center as outlined above, and the tuning of $\lambda \in [0,1]$ gives us the flexibility of centering anywhere below the current progress x_{pre}. Roughly speaking, we will choose $\lambda \in [0,1]$ to maximize the derivative $f'(\lambda x_{\text{pre}})$. Finally, as the relation $\langle\theta^\star, v_{\text{pre}}\rangle \approx x_{\text{pre}}$ is only approximate, we use a more complicated projection based test (3.3) for the certification of $\langle\theta^\star, v\rangle$.

The performance of Algorithm 3 is summarized in the following lemma.

LEMMA 3.2. *Suppose the link function f is monotone on $[-1,1]$, $\langle\theta^\star, v_{\text{pre}}\rangle \in [x_{\text{pre}}, (1+3\kappa_1)x_{\text{pre}}]$, and $x_{\text{pre}} \geq \kappa_4/[(1-\kappa_2)\sqrt{d}]$. Then with probability at least $1-\delta$, the* GOODACTIONHYPTEST *algorithm outputs*:

- True *if* $\langle\theta^\star, v\rangle \in [(1+\kappa_1)/\sqrt{d}, (1+2\kappa_1)/\sqrt{d}]$;
- False *if* $\langle\theta^\star, v\rangle \notin [1/\sqrt{d}, (1+3\kappa_1)/\sqrt{d}]$.

3.1.3. *An algorithm without the knowledge of f.* Recall that Algorithm 1 crucially relies on the knowledge of f, as it is used in both projection based tests in the INITIALACTIONHYPTEST and GOODACTIONHYPTEST algorithms. The main result of this section is summarized in the following theorem, showing that the knowledge of f is not required for the burn-in period.

THEOREM 3.2. *Consider the same setting of Theorem 3.1, and assume that f is continuous and strictly increasing on $[-1,1]$. Then there is an algorithm without the knowledge of f such that, with probability at least $1-\delta$, it outputs an action a_0 with $\langle\theta^\star, a_0\rangle \geq x_0$ using*

$$O\left(\log^3\left(\frac{d}{\delta}\right)\sum_{i=1}^{m}\frac{1}{\varepsilon_i^2}\right)$$

queries, where the hidden constant depends only on $(x_0, \kappa_1, \kappa_2)$.

Up to logarithmic factors in $(d, 1/\delta)$, the sample complexity in Theorem 3.2 matches the result in Theorem 3.1. The main algorithmic idea is to solve the following *infinite-armed bandit problem*. Let F be an unknown, continuous, and strictly increasing CDF, so that $F^{-1}(t)$ is well-defined for every $t \in (0,1)$. Let $X_1, X_2, \ldots \sim F$ be an (unobserved) infinite i.i.d. sequence (treat the index set $\mathbb{N}$ as *arms*). At each time t, the learner chooses an arm $i_t \in \mathbb{N}$ and observes $Y_t \sim \mathcal{N}(X_{i_t}, 1)$; the learner could either pull a new arm for exploration, or pull an existing arm to refine the knowledge of X. We assume that the noises at different rounds are independent. Given two values $p, q \in [0,1]$ with $p < q$, the learner's target is to find some $i \in \mathbb{N}$ such that $F(X_i) \in [p,q]$. A line of work [10, 14, 72, 73] considered similar settings, but typically focused on different targets such as best arm identification or functional estimation.

The following lemma presents a simple algorithm based on upper and lower confidence bounds, together with a high-probability guarantee on the sample complexity.

LEMMA 3.3. *Fix any $\varepsilon \in (0, (q-p)/4)$, and a failure probability $\delta \in (0, 1/2)$. There is a learning algorithm such that with probability at least $1-\delta$, it outputs some $i \in \mathbb{N}$ with $F(X_i) \in [p,q]$ using*

$$O_{q-p,\varepsilon}\left(\frac{\log^2(1/\delta)}{(F^{-1}(p+2\varepsilon) - F^{-1}(p+\varepsilon))^2} + \frac{\log^2(1/\delta)}{(F^{-1}(q-\varepsilon) - F^{-1}(q-2\varepsilon))^2}\right)$$

queries, where both the algorithm and the hidden constant are independent of F.

To see how Lemma 3.3 is related to our problem, consider the initial certification steps in Algorithm 1. Let F be the CDF of $f(\langle\theta^\star, v\rangle)$ for $v \sim \mathsf{Unif}(\mathbb{S}^{d-1})$, which is unknown due to the unknown f. If we sample $v_1, v_2, \ldots \sim \mathsf{Unif}(\mathbb{S}^{d-1})$, then each direction v_i is an arm in the infinite-armed bandit problem, with corresponding $X_i = f(\langle\theta^\star, v_i\rangle)$, and the reward $r_t \sim \mathcal{N}(X_{i_t}, 1)$ is the observation Y_t when the direction v_{i_t} is chosen. The crucial observation here is that both

$$p = F^{-1}\big(f(1/\sqrt{d})\big) = \mathbb{P}\big(|\langle\theta^\star, v\rangle| \le 1/\sqrt{d}\big),$$
$$q = F^{-1}\big(f((1+\kappa_1)/\sqrt{d})\big) = \mathbb{P}\big(|\langle\theta^\star, v\rangle| \le (1+\kappa_1)/\sqrt{d}\big)$$

are known thanks to the strict monotonicity of f, and Lemma A.1 in the Supplementary Material [58] tells that $q - p = \Omega(1)$. Therefore, we can apply Lemma 3.3 to find a direction (arm) v_i such that $\langle\theta^\star, v_i\rangle \in [1/\sqrt{d}, (1+\kappa_1)/\sqrt{d}]$, with sample complexity essentially $\widetilde{O}(1/\varepsilon_1^2)$ in Theorem 3.1. In summary, instead of certifying each direction one after one using the knowledge of f in Algorithm 1, the agnostic algorithm makes use of the empirical CDF based on the comparisons between different actions.

The same idea could also be applied to recursive certification steps, with two additional caveats. First, the CDF of $\langle\theta^\star, v\rangle$ with $v \sim \mathsf{Unif}(\mathbb{S}^{d-1} \cap V^\perp)$ involves an unknown magnitude $\|\mathrm{Proj}_{V^\perp}(\theta^\star)\|_2$; in the algorithm we estimate it and apply an induction in the analysis. Second, the optimal value of λ in Algorithm 3 is unknown; we overcome it by searching over a geometric grid on λ. The detailed algorithms, as well as the proofs of Lemma 3.3 and Theorem 3.2, are deferred to the Supplementary Material [58].

3.2. *Algorithm for the learning phase.* In this section, we design an algorithm after a good action a_0 with $\langle\theta^\star, a_0\rangle \ge 1 - 3\gamma/4$ is found, and prove the upper bound in Theorem 1.6. The algorithm is based on a simple idea of explore-then-commit (ETC) shown in Algorithm 4. In the first m rounds, we cyclically explore all directions *around* a_0 in a nonadaptive manner:

$$a_t \in \left\{\left(1 - \frac{\gamma}{8}\right)a_0 \pm \frac{\gamma}{8}e_i : i \in [d]\right\} \subseteq \mathbb{B}^d.$$

Here e_i is the ith canonical vector of $\mathbb{R}^d$. We center these actions around a_0 to ensure that

$$\left\langle\theta^\star, \left(1 - \frac{\gamma}{8}\right)a_0 \pm \frac{\gamma}{8}e_i\right\rangle \ge \left(1 - \frac{3\gamma}{4}\right)\left(1 - \frac{\gamma}{8}\right) - \frac{\gamma}{8} > 1 - \frac{3\gamma}{4} - \frac{\gamma}{8} - \frac{\gamma}{8} = 1 - \gamma$$

Algorithm 4: Regression-based explore-then-commit algorithm

Input: link function f, dimensionality d, time horizon T, action a_0 with $\langle a_0, \theta^\star\rangle \ge 1 - \gamma$.

Output: final estimator $\widehat{\theta}_T$, or a sequence of actions $(a_1, \ldots, a_T)$.

Set $m \leftarrow T$ for estimation, and $m \leftarrow \min\{T, d\sqrt{T}/c_f\}$ for regret minimization.

for $t = 1, 2, \ldots, m$ **do**

  Play action $a_t = (1 - \frac{\gamma}{8})a_0 + (-1)^{\lceil t/d\rceil} \cdot \frac{\gamma}{8}e_{((t-1) \bmod d)+1}$;

  Receive reward $r_t \sim \mathcal{N}(f(\langle\theta^\star, a_t\rangle), 1)$.

Compute the constrained least squares estimator:

$$\widehat{\theta}^{\mathrm{LS}} = \mathop{\arg\min}_{\theta \in \mathbb{S}^{d-1} : \langle\theta, a_0\rangle \ge 1-\gamma} \sum_{t=1}^{m} \big(f(\langle\theta, a_t\rangle) - r_t\big)^2. \tag{3.4}$$

for $t = m+1, \ldots, T$ **do:** commit to the action $a_t = \widehat{\theta}^{\mathrm{LS}}$.

Return $\widehat{\theta}_T = \widehat{\theta}^{\mathrm{LS}}$ or $(a_1, \ldots, a_T)$.

for all $i \in [d]$ and therefore we are operating in the locally linear regime in Assumption 2. After the exploration rounds, we compute the constrained least squares estimator $\widehat{\theta}^{\mathrm{LS}}$ for $\theta^\star$ in (3.4). If our target is the estimation of $\theta^\star$, we just set $m = T$ and use $\widehat{\theta}^{\mathrm{LS}}$ as the final estimator. If our target is to minimize the regret, we commit to the action $a_t = \widehat{\theta}^{\mathrm{LS}}$ after $t > m$, and choose m appropriately to balance the errors in the exploration and commit rounds. Further details are discussed in Appendix C.8 of the Supplementary Material [58].

4. Additional discussions.

4.1. *Nonadaptive sampling.* In this section, we show that the upper bound on the burn-in cost in Theorem 1.1 cannot be attained by any nonadaptive sampling approaches in general. Here under nonadaptive sampling, the actions $a_1, \dots, a_T \in \mathbb{B}^d$ are chosen in advance without knowing the history. This result reveals a gap between adaptive and nonadaptive samplings, and emphasizes the importance of the sequential nature in our decision making problem.

THEOREM 4.1. *Let the link function f satisfy Assumption 1 in the ridge bandit, and $\theta^\star \sim \mathsf{Unif}(\mathbb{S}^{d-1})$. Then any nonadaptive learner cannot find $\widehat{\theta}_T$ with $\mathbb{E}[\langle \theta^\star, \widehat{\theta}_T \rangle] > 1/2$ if*

$$T < \max_{K \geq 1} \frac{cd}{g(\sqrt{(\log K)/d})^2 + K^{-1}},$$

where $c > 0$ is an absolute constant, and $g(x) := \max\{|f(x)|, |f(-x)|\}$.

If $f(x) = |x|^p$ with $p > 0$, Theorem 4.1 shows that the burn-in cost for all nonadaptive algorithms is at least $\widetilde{\Omega}(d^{p+1})$, which is suboptimal compared with Example 1 when $p > 1$. Thanks to the nonadaptive nature where a_t is independent of $\theta^\star$, Theorem 4.1 could be proven via the classical Fano's inequality. Without this independence in the adaptive setting, we need a recursive relationship for the mutual information $I(\theta^\star; a_t)$ in the proof of Theorem 2.1.

4.2. *Finitely many actions.* In this section we consider the case where the action space $\mathcal{A}$ is not continuous and is a finite subset of $\mathbb{B}^d$, with $|\mathcal{A}| = K$. For linear bandits, a finite set of actions helps reduce the minimax regret from $\Theta(d\sqrt{T})$ to $\Theta(\sqrt{dT \log K})$, essentially due to the reason that it becomes less expensive to maintain a confidence bound for each action (see, e.g., [53], Chapter 22). However, to achieve the optimal burn-in cost for general ridge bandits, we already know that it is necessary to go beyond confidence bounds. In this case, does a finite number of actions help to reduce the burn-in cost as well? The next theorem shows that for many link functions, a smaller set of actions does not essentially help.

THEOREM 4.2. *Let the link function f satisfy Assumption 1 in the ridge bandit problem. For every $K = \exp(o(d))$, there exists a finite action set $\mathcal{A}$ with $|\mathcal{A}| = K$ such that any learner cannot find $\widehat{\theta}_T$ with $\inf_{\theta^\star \in \mathbb{S}^{d-1}} \mathbb{E}_{\theta^\star}[\langle \theta^\star, \widehat{\theta}_T \rangle] \geq 4/5$ if*

$$T < \frac{c}{g(\sqrt{(c' \log K)/d})^2 + K^{-1}},$$

where $c, c' > 0$ are absolute constants, and $g(x) := \max\{|f(x)|, |f(-x)|\}$.

For $f(x) = |x|^p$ with $p > 0$, Theorem 4.2 shows that the burn-in cost with appropriately chosen K actions is at least $\widetilde{\Omega}(d^p)$ as long as $K \gtrsim d^p$. If $p \geq 2$, this is no smaller than the optimal burn-in cost with a continuous set of actions, showing that a smaller action set is essentially not beneficial. From the algorithmic perspective, this is because that Algorithm 1 for the burn-in period crucially requires that every direction, and in particular every convex

combination of actions, could be explored—a structure that may break down for finitely many actions. Under a given discrete action set, it is an interesting future direction to understand both the burn-in cost and the appropriate algorithm for the burn-in period.

We also point out some technical aspects in the proof of Theorem 4.2. First, a proof based on the χ^2-informativity argument in Section 2 still works, but the proof we present uses the classical two-point method with an additional change-of-measure trick to a common distribution. Second, this trick does not suffice to give Corollary 2.1: when passing through the common distribution to exchange the order of expectations, the inner product is always of the scale $\widetilde{\Theta}(1/\sqrt{d})$ but no other intermediate scales $1/\sqrt{d} \ll \varepsilon \ll 1$ as in Corollary 2.1. See Appendix D of the Supplementary Material [58] for more details.

4.3. *Unit sphere vs unit ball.* In this section, we relax the assumption $\theta^\star \in \mathbb{S}^{d-1}$ and investigate the statistical complexity of ridge bandits when $\theta^\star \in \mathbb{B}^d$. The following theorem shows that it is equivalent to think of the unit ball as a union of spheres with different radii.

THEOREM 4.3. *Suppose the link function f satisfies the monotonicity condition in Assumption 1, and $f'(x)/f'(y) \le C$ as long as $1/c \le x/y \le c$ for some constants $c, C > 1$. Then the following upper and lower bounds hold for the minimax regret over $\theta^\star \in \mathbb{B}^d$:*

$$\begin{aligned}
&\max_{r\in[0,1]} \min\left\{ \frac{f(r)}{r^2} d \int_{r/\sqrt{d}}^{r/2} \frac{\mathrm{d}(x^2)}{\max\{f(x)^2, f(-x)^2\}} + d\sqrt{T}, Tf(r) \right\} \\
&\lesssim \mathfrak{R}_T^\star(f,d) \\
&\lesssim \max_{r\in[0,1]} \min\left\{ \frac{f(r)}{r^4} d^2 \int_{r/\sqrt{d}}^{r/2} \frac{\mathrm{d}(x^2)}{\max_{r/\sqrt{d}\le y\le x} \min_{z\in[(1-\kappa)y,(1+\kappa)y]}[f'(z)]^2} + d\sqrt{T}, Tf(r) \right\},
\end{aligned}$$

where $\kappa \in (0, 1/4)$ is any fixed parameter, and the hidden factors depend only on (c, C, κ).

The sample complexity for estimation could be obtained in a similar manner, and we omit the details. For $f(x) = |x|^p$ with $p > 0$, the above theorem shows that $\mathfrak{R}_T^\star(f,d) \asymp \min\{\sqrt{d^{\max\{2,p\}}T}, T\}$, matching the result in [38]. Note that because of an additional maximum over $r \in [0,1]$, the minimax regret over the unit ball only exhibits one elbow at $T \asymp d^{\max\{2,p\}}$, in contrast to two elbows in Figure 1 over the unit sphere. The assumption in Theorem 4.3 is also stronger than Assumption 2, for we need Assumption 2 to hold for every function $x \in [0,1] \mapsto f(rx)$ with $r > 0$.

If $r := \|\theta\|_2 \in [0,1]$ is known, the proof of Theorem 4.3 adapts from our upper and lower bounds for the unit sphere after proper scaling, and we simply take the worst case radius $r \in [0,1]$. It then remains to find an estimate $\widehat{r}$ of r such that $r \in [\widehat{r}/4, \widehat{r}]$ with high probability. This step is deferred to Appendix D of the Supplementary Material [58], with an additional sample complexity which is negligible compared to Theorem 4.3.

4.4. *Closing the gap between upper and lower bounds.* There is a gap in Theorems 1.1 and 1.2: the upper bound is in terms of the derivative of f, but the lower bound is only in terms of the function value of f. We conjecture that the lower bound could be strengthened, due to the following intuition. In the proof of Lemma 2.2, the distribution $Q_{\mathcal{H}_t}$ is constructed so that $r_t \sim \mathcal{N}(0,1)$. In principle, the mean of r_t could be any function $\mu(a_1, r_1, \ldots, a_{t-1}, r_{t-1}, a_t)$ of the available history, and a natural choice is $r_t \sim \mathcal{N}(\mathbb{E}[f(\langle\theta^\star, a_t\rangle)|\mathcal{H}_{t-1}], 1)$. Under this choice, the information gain in the recursion is $\mathsf{Var}(f(\langle\theta^\star, a_t\rangle)|\mathcal{H}_{t-1})$, with expected value

$$\mathbb{E}\big[\mathsf{Var}\big(f\big(\langle\theta^\star, a_t\rangle\big)|\mathcal{H}_{t-1}\big)\big] \lesssim \max_{y\le\varepsilon_t}\big[f'(y)\big]^2 \cdot \mathbb{E}\big[\mathsf{Var}\big(\langle\theta^\star, a_t\rangle|\mathcal{H}_{t-1}\big)\big] \le \frac{1}{d}\max_{y\le\varepsilon_t}\big[f'(y)\big]^2.$$

Proceeding with this intuition will give a lower bound of a similar form to Theorem 1.1. However, a formal argument will require that the above upper bound holds with high probability rather than in expectation, a challenging claim that involves a complicated posterior distribution of $\theta^\star$. We leave it as an open direction, but give a special example where the high probability argument is feasible using the Brascamp–Lieb inequality on manifolds [45].

THEOREM 4.4. *For the linear bandit $f(x) = \mathrm{id}(x) = x$ with dimension d, it holds that $T^\star_{\textit{burn-in}}(\mathrm{id}, d) \gtrsim d^2$.*

Note that the lower bound $T^\star(\mathrm{id}, d, \varepsilon) \gtrsim d^2$ shown in [69] only works for a small error ε (say $\varepsilon \le 0.1$), due to an intrinsic limitation of the hypercube structure used in Assouad's lemma. In contrast, Theorem 4.4 shows the same $\widetilde{\Omega}(d^2)$ lower bound for $\varepsilon = 1/2$ (or any fixed $\varepsilon < 1$), which improves over the lower bound $\widetilde{\Omega}(d)$ in Theorem 1.2 for linear bandits.

Funding. Nived Rajaraman and Jiantao Jiao were partially supported by NSF Grants IIS-1901252 and CIF-2211209. Yanjun Han was supported by the Simons-Berkeley research fellowship and the Norbert Wiener postdoctoral fellowship.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical complexity and optimal algorithms for nonlinear ridge bandits" (DOI: 10.1214/24-AOS2395SUPP; .pdf). We provide auxiliary lemmas used in this paper and proofs of minimax lower bound (Theorem 1.7), Theorems 1.1, 1.4, 1.5, 1.6, 2.1, 3.1, 3.2, 4.1, 4.2, 4.3, 4.4, and Lemmas 2.1, 2.2, 3.2, 3.3.

REFERENCES

[1] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Adv. Neural Inf. Process. Syst.* **24**.

[2] AGARWAL, A., DUDÍK, M., KALE, S., LANGFORD, J. and SCHAPIRE, R. (2012). Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics* 19–26. PMLR.

[3] AGARWAL, A., FOSTER, D. P., HSU, D. J., KAKADE, S. M. and RAKHLIN, A. (2011). Stochastic convex optimization with bandit feedback. *Adv. Neural Inf. Process. Syst.* **24**.

[4] AGARWAL, A., HSU, D., KALE, S., LANGFORD, J., LI, L. and SCHAPIRE, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning* 1638–1646. PMLR.

[5] AGARWAL, A., WAINWRIGHT, M. J., BARTLETT, P. and RAVIKUMAR, P. (2009). Information-theoretic lower bounds on the oracle complexity of convex optimization. *Adv. Neural Inf. Process. Syst.* **22**.

[6] AUER, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3** 397–422. MR1984023 https://doi.org/10.1162/153244303321897663

[7] AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002/03). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* **32** 48–77. MR1954855 https://doi.org/10.1137/S0097539701398375

[8] BACHOC, F., CESARI, T. and GERCHINOVITZ, S. (2021). Instance-dependent bounds for zeroth-order Lipschitz optimization with error certificates. *Adv. Neural Inf. Process. Syst.* **34** 24180–24192.

[9] BARTROFF, J., LAI, T. L. and SHIH, M.-C. (2012). *Sequential Experimentation in Clinical Trials: Design and Analysis* **298**. Springer, Berlin.

[10] BERRY, D. A., CHEN, R. W., ZAME, A., HEATH, D. C. and SHEPP, L. A. (1997). Bandit problems with infinitely many arms. *Ann. Statist.* **25** 2103–2116. MR1474085 https://doi.org/10.1214/aos/1069362389

[11] BILLARD, A. and KRAGIC, D. (2019). Trends and challenges in robot manipulation. *Science* **364**. https://doi.org/10.1126/science.aat8414

[12] BLOT, W. J. and MEETER, D. A. (1973). Sequential experimental design procedures. *J. Amer. Statist. Assoc.* **68** 586–593. MR0359209

[13] BLUMENTHAL, S. (1976). Sequential estimation of the largest normal mean when the variance is known. *Ann. Statist.* **4** 1077–1087. MR0431479

[14] BONALD, T. and PROUTIERE, A. (2013). Two-target algorithms for infinite-armed bandits with Bernoulli rewards. *Adv. Neural Inf. Process. Syst.* **26**.
[15] BOUTTIER, C., CESARI, T., DUCOFFE, M. and GERCHINOVITZ, S. (2020). Regret analysis of the Piyavskii-Shubert algorithm for global Lipschitz optimization. ArXiv preprint arXiv:2002.02390.
[16] BUBECK, S. and ELDAN, R. (2016). Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory* 583–589. PMLR.
[17] BUBECK, S., LEE, Y. T. and ELDAN, R. (2017). Kernel-based methods for bandit convex optimization. In *STOC'17—Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing* 72–85. ACM, New York. MR3678172 https://doi.org/10.1145/3055399.3055403
[18] BURNAŠEV, M. (1980). Sequential discrimination of hypotheses with control of observations. *Math. USSR, Izv.* **15** 419.
[19] CHEN, F., MEI, S. and BAI, Y. (2022). Unified algorithms for RL with decision-estimation coefficients: No-regret, PAC, and reward-free learning. ArXiv preprint arXiv:2209.11745.
[20] CHERNOFF, H. (1959). Sequential design of experiments. *Ann. Math. Stat.* **30** 755–770. MR0108874 https://doi.org/10.1214/aoms/1177706205
[21] 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.
[22] CSISZÁR, I. (1972). A class of measures of informativity of observation channels. *Period. Math. Hungar.* **2** 191–213. MR0335152 https://doi.org/10.1007/BF02018661
[23] DANI, V., HAYES, T. P. and KAKADE, S. M. (2008). Stochastic linear optimization under bandit feedback. *Conf. Learn. Theory* 355–366.
[24] DU, S., KAKADE, S., LEE, J., LOVETT, S., MAHAJAN, G., SUN, W. and WANG, R. (2021). Bilinear classes: A structural framework for provable generalization in RL. In *International Conference on Machine Learning* 2826–2836. PMLR.
[25] DUDIK, M., HSU, D., KALE, S., KARAMPATZIAKIS, N., LANGFORD, J., REYZIN, L. and ZHANG, T. (2011). Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* 169–178.
[26] FILIPPI, S., CAPPE, O., GARIVIER, A. and SZEPESVÁRI, C. (2010). Parametric bandits: The generalized linear case. *Adv. Neural Inf. Process. Syst.* **23**.
[27] FLAXMAN, A. D., KALAI, A. T. and MCMAHAN, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 385–394. ACM, New York. MR2298287
[28] FOSTER, D. and RAKHLIN, A. (2020). Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning* 3199–3210. PMLR.
[29] FOSTER, D. J., GENTILE, C., MOHRI, M. and ZIMMERT, J. (2020). Adapting to misspecification in contextual bandits. *Adv. Neural Inf. Process. Syst.* **33** 11478–11489.
[30] FOSTER, D. J., GOLOWICH, N. and HAN, Y. (2023). Tight guarantees for interactive decision making with the decision-estimation coefficient. In *The Thirty Sixth Annual Conference on Learning Theory* 3969–4043. PMLR.
[31] FOSTER, D. J., GOLOWICH, N., QIAN, J., RAKHLIN, A. and SEKHARI, A. (2022). A note on model-free reinforcement learning with the decision-estimation coefficient. ArXiv preprint arXiv:2211.14250.
[32] FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. ArXiv preprint arXiv:2112.13487.
[33] FOSTER, D. J., RAKHLIN, A., SEKHARI, A. and SRIDHARAN, K. (2022). On the complexity of adversarial decision making. *Adv. Neural Inf. Process. Syst.* **35** 35404–35417.
[34] GARIVIER, A., MÉNARD, P. and STOLTZ, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.* **44** 377–399. MR3959077 https://doi.org/10.1287/moor.2017.0928
[35] GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B* **41** 148–164.
[36] HANSEN, P., JAUMARD, B. and LU, S.-H. (1991). On the number of iterations of Piyavskiĭ's global optimization algorithm. *Math. Oper. Res.* **16** 334–350. MR1106805 https://doi.org/10.1287/moor.16.2.334
[37] HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004). *Nonparametric and Semiparametric Models. Springer Series in Statistics*. Springer, New York. MR2061786 https://doi.org/10.1007/978-3-642-17146-8
[38] HUANG, B., HUANG, K., KAKADE, S., LEE, J. D., LEI, Q., WANG, R. and YANG, J. (2021). Optimal gradient-based algorithms for non-concave bandit optimization. *Adv. Neural Inf. Process. Syst.* **34** 29101–29115.

[39] JAMIESON, K. G., NOWAK, R. and RECHT, B. (2012). Query complexity of derivative-free optimization. *Adv. Neural Inf. Process. Syst.* **25**.
[40] JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning* 1704–1713. PMLR.
[41] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman Eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Adv. Neural Inf. Process. Syst.* **34** 13406–13418.
[42] KALASHNIKOV, D., IRPAN, A., PASTOR, P., IBARZ, J., HERZOG, A., JANG, E., QUILLEN, D., HOLLY, E., KALAKRISHNAN, M. et al. (2018). Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. ArXiv preprint arXiv:1806.10293.
[43] KLEINBERG, R. (2004). Nearly tight bounds for the continuum-armed bandit problem. *Adv. Neural Inf. Process. Syst.* **17**.
[44] KOBER, J., BAGNELL, J. A. and PETERS, J. (2013). Reinforcement learning in robotics: A survey. *Int. J. Robot. Res.* **32** 1238–1274.
[45] KOLESNIKOV, A. V. and MILMAN, E. (2016). Riemannian metrics on convex sets with applications to Poincaré and log-Sobolev inequalities. *Calc. Var. Partial Differential Equations* **55** 77. MR3514409 https://doi.org/10.1007/s00526-016-1018-3
[46] KRISHNAMURTHY, S. K., HADAD, V. and ATHEY, S. (2021). Adapting to misspecification in contextual bandits with offline regression oracles. In *International Conference on Machine Learning* 5805–5814. PMLR.
[47] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8
[48] LATTIMORE, T. (2019). Improved regret for zeroth-order adversarial bandit convex optimisation. *Math. Stat. Learn.* **2** 311–334. MR4165267 https://doi.org/10.4171/msl/17
[49] LATTIMORE, T. (2021). Minimax regret for bandit convex optimisation of ridge functions. ArXiv preprint arXiv:2106.00444.
[50] LATTIMORE, T. (2022). Minimax regret for partial monitoring: Infinite outcomes and rustichini's regret. In *Conference on Learning Theory* 1547–1575. PMLR.
[51] LATTIMORE, T. and GYORGY, A. (2021). Mirror descent and the information ratio. In *Conference on Learning Theory* 2965–2992. PMLR.
[52] LATTIMORE, T. and HAO, B. (2021). Bandit phase retrieval. *Adv. Neural Inf. Process. Syst.* **34** 18801–18811.
[53] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.
[54] LOGAN, B. F. and SHEPP, L. A. (1975). Optimal reconstruction of a function from its projections. *Duke Math. J.* **42** 645–659. MR0397240
[55] MINSKY, M. (1961). Steps toward artificial intelligence. *Proc. IRE* **49** 8–30. MR0134428
[56] OKAMOTO, I., AMARI, S. and TAKEUCHI, K. (1991). Asymptotic theory of sequential estimation: Differential geometrical approach. *Ann. Statist.* **19** 961–981. MR1105855 https://doi.org/10.1214/aos/1176348131
[57] RAGINSKY, M. and RAKHLIN, A. (2011). Information-based complexity, feedback and dynamics in convex programming. *IEEE Trans. Inf. Theory* **57** 7036–7056. MR2882278 https://doi.org/10.1109/TIT.2011.2154375
[58] RAJARAMAN, N., HAN, Y., JIAO, J. and RAMCHANDRAN, K. (2024). Supplement to "Statistical Complexity and Optimal Algorithms for Non-linear Ridge Bandits." https://doi.org/10.1214/24-AOS2395SUPP
[59] ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246 https://doi.org/10.1090/S0002-9904-1952-09620-8
[60] RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Adv. Neural Inf. Process. Syst.* **26**.
[61] RUSSO, D. and VAN ROY, B. (2014). Learning to optimize via information-directed sampling. *Adv. Neural Inf. Process. Syst.* **27**.
[62] SHAMIR, O. (2013). On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory* 3–24. PMLR.
[63] SIMCHI-LEVI, D. and XU, Y. (2022). Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Math. Oper. Res.* **47** 1904–1931. MR4506358
[64] SUN, W., JIANG, N., KRISHNAMURTHY, A., AGARWAL, A. and LANGFORD, J. (2019). Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory* 2898–2933. PMLR.
[65] SUTTON, R. S. (1984). *Temporal Credit Assignment in Reinforcement Learning*. Univ. Massachusetts Amherst, Amherst.
[66] SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement Learning*: *An Introduction*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3889951

[67] TATIKONDA, S. and MITTER, S. (2009). The capacity of channels with feedback. *IEEE Trans. Inf. Theory* **55** 323–349. MR2589700 https://doi.org/10.1109/TIT.2008.2008147

[68] VILLAR, S. S., BOWDEN, J. and WASON, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statist. Sci.* **30** 199–215. MR3353103 https://doi.org/10.1214/14-STS504

[69] WAGENMAKER, A. J., CHEN, Y., SIMCHOWITZ, M., DU, S. and JAMIESON, K. (2022). Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *International Conference on Machine Learning* 22430–22456. PMLR.

[70] WALD, A. and WOLFOWITZ, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Stat.* **19** 326–339. MR0026779 https://doi.org/10.1214/aoms/1177730197

[71] WANG, R., SALAKHUTDINOV, R. R. and YANG, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Adv. Neural Inf. Process. Syst.* **33** 6123–6135.

[72] WANG, Y., AUDIBERT, J.-Y. and MUNOS, R. (2008). Algorithms for infinitely many-armed bandits. *Adv. Neural Inf. Process. Syst.* **21**.

[73] WANG, Y., BAHARAV, T., HAN, Y., JIAO, J. and TSE, D. (2022). Beyond the best: Estimating distribution functionals in infinite-armed bandits. *Adv. Neural Inf. Process. Syst.* **35**.

[74] WANG, Y., WANG, R. and KAKADE, S. (2021). An exponential lower bound for linearly realizable MDP with constant suboptimality gap. *Adv. Neural Inf. Process. Syst.* **34** 9521–9533.

[75] WEISZ, G., AMORTILA, P. and SZEPESVÁRI, C. (2021). Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory* 1237–1264. PMLR.

[76] ZHU, H., GUPTA, A., RAJESWARAN, A., LEVINE, S. and KUMAR, V. (2019). Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost. In 2019 *International Conference on Robotics and Automation* (*ICRA*) 3651–3657. IEEE, Los Alamitos.