# Optimization of Smooth Functions With Noisy Observations: Local Minimax Rates

Yining Wang, Sivaraman Balakrishnan, and Aarti Singh

***Abstract*—We consider the problem of *global optimization* of an unknown non-convex smooth function with noisy zeroth-order feedback. We propose a *local minimax* framework to study the fundamental difficulty of optimizing smooth functions with adaptive function evaluations. We show that for functions with fast growth around their global minima, carefully designed optimization algorithms can identify a near global minimizer with many fewer queries than worst-case global minimax theory predicts. For the special case of strongly convex and smooth functions, our implied convergence rates match the ones developed for zeroth-order *convex* optimization problems. On the other hand, we show that in the worst case no algorithm can converge faster than the minimax rate of estimating an unknown function in the $\ell_\infty$-norm. Finally, we show that non-adaptive algorithms, though optimal in a global minimax sense, do not attain the optimal local minimax rate.**

***Index Terms*—Optimization of smooth functions, nonparametric statistics, local minimax analysis.**

## I. Introduction

Global function optimization with stochastic (zeroth-order) query oracles is an important problem in optimization, machine learning and statistics. To optimize an unknown bounded function $f : \mathcal{X} \mapsto \mathbb{R}$ defined on a known compact $d$-dimensional domain $\mathcal{X} \subseteq \mathbb{R}^d$, the data analyst makes $n$ *active* queries $x_1, \ldots, x_n \in \mathcal{X}$ and observes

$$y_t = f(x_t) + w_t, \qquad w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1),^1 \quad t = 1, \ldots, n. \quad (1)$$

The queries $x_1, \ldots, x_t$ are *active* in the sense that the selection of $x_t$ can depend on the previous queries and their responses $x_1, y_1, \ldots, x_{t-1}, y_{t-1}$. After $n$ queries, an estimate $\widehat{x}_n \in \mathcal{X}$ is produced that approximately minimizes the unknown function $f$. Such "active query" models are relevant in a broad range of (noisy) global optimization applications, for instance in hyper-parameter tuning of machine learning algorithms [1] and sequential design in material synthesis experiments where the goal is to maximize the strength of the synthesized material as a function of experimental settings [2], [3]. We refer the readers to Section II-A for a rigorous formulation of the active query model and contrast it with the classical passive query model.

The error of the estimate $\widehat{x}_n$ is measured by the difference of $f(\widehat{x}_n)$ and the *global minimum* of $f$:

$$\mathfrak{L}(\widehat{x}_n; f) := f(\widehat{x}_n) - f^* \qquad \text{where } f^* := \inf_{x \in \mathcal{X}} f(x). \quad (2)$$

To simplify our presentation, throughout the paper we take the domain $\mathcal{X}$ to be the $d$-dimensional unit cube $[0, 1]^d$, while our results can be easily generalized to other compact domains satisfying minimal regularity conditions.

When $f$ belongs to a smoothness class, say the Hölder class with exponent $\alpha$, a straightforward global optimization method is to first sample $n$ points uniformly at random from $\mathcal{X}$ and then construct nonparametric estimates $\widehat{f}_n$ of $f$ using nonparametric regression methods such as kernel smoothing or local polynomial regression [4], [5]. Classical analysis shows that the sup-norm reconstruction error $\|\widehat{f}_n - f\|_\infty = \sup_{x \in \mathcal{X}} |\widehat{f}_n(x) - f(x)|$ can be upper bounded by $\widetilde{O}_{\mathbb{P}}(n^{-\alpha/(2\alpha+d)})$[2]. This global reconstruction guarantee then implies an $\widetilde{O}_{\mathbb{P}}(n^{-\alpha/(2\alpha+d)})$ upper bound on $\mathfrak{L}(\widehat{x}_n; f)$ by considering an estimate $\widehat{x}_n \in \mathcal{X}$ for which $\widehat{f}_n(\widehat{x}_n) = \inf_{x \in \mathcal{X}} \widehat{f}_n(x)$ (such an $\widehat{x}_n$ exists because $\mathcal{X}$ is closed and bounded). Formally, we have the following proposition (proved in the Appendix) that converts a global reconstruction guarantee into an upper bound on the optimization error:

**Proposition 1.** *Suppose* $\widehat{f}_n(\widehat{x}_n) = \inf_{x \in \mathcal{X}} \widehat{f}_n(x)$. *Then* $\mathfrak{L}(\widehat{x}_n; f) \leqslant 2\|\widehat{f}_n - f\|_\infty$.

Typically, fundamental limits on the optimal optimization error are understood through the lens of *minimax analysis* where the object of study is the (global) minimax risk:

$$\inf_{\widehat{x}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \mathfrak{L}(\widehat{x}_n, f), \quad (3)$$

where $\mathcal{F}$ is a certain class of smooth functions such as the Hölder class. Although optimization appears to be easier than global reconstruction, we show in this paper that the $n^{-\alpha/(2\alpha+d)}$ rate is *not* improvable in the global minimax sense in over Hölder classes. Such a surprising phenomenon was also noted in previous works [6]–[8] for related problems. On the

AQ:1 Manuscript received August 10, 2018; revised April 21, 2019; accepted May 5, 2019. S. Balakrishnan was supported in part by the NSF under Grant DMS-17130003. Y. Wang and A. Singh were supported in part by the NSF under Grant CCF-1563918 and in part by the AFRL under Grant FA8750-17-2-0212. This paper was presented in part at the 2018 NeurIPS Conference.

Y. Wang and A. Singh are with the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: yiningwa,aarti@cs.cmu.edu).

S. Balakrishnan is with the Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: siva@stat.cmu.edu).

Communicated by K. Chaudhuri, Associate Editor for Statistical Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2921985

[1]The exact Gaussianity of the independent noise variables $\varepsilon_t$ is not crucial and our results can be easily generalized to sub-Gaussian noise.

[2]In the $\widetilde{O}(\cdot)$ or $\widetilde{O}_{\mathbb{P}}(\cdot)$ notation we suppress constant factors and terms that depend poly-logarithmically on $n$.

other hand, extensive empirical evidence suggests that non-uniform/active allocations of query points can significantly reduce optimization error in practical global optimization of smooth, non-convex functions [1]. This raises the interesting question of understanding, from a theoretical perspective, the conditions under which the global optimization of smooth functions is *easier* than their reconstruction, and the power of *active/feedback-driven* queries that play important roles in global optimization.

In this paper, we propose a theoretical framework that partially answers the above questions. In contrast to classical *global* minimax analysis of nonparametric estimation problems, we adopt a *local analysis* which characterizes the optimal convergence rate of optimization error when the underlying function $f$ is within a neighborhood of a "reference" function $f_0$. (See Section II-B for the rigorous local minimax formulation considered in this paper.) Our main results are to characterize the local convergence rates $R_n(f_0)$ for a wide range of reference functions $f_0 \in \mathcal{F}$. Concretely, our contributions can be summarized as follows:

1) We design an iterative (active) algorithm whose optimization error $\mathfrak{L}(\widehat{x}_n; f)$ converges at a rate of $R_n(f_0)$ depending on the reference function $f_0$. When the level-sets of $f_0$ satisfy certain regularity and polynomial growth conditions, the local rate $R_n(f_0)$ can be upper bounded by $R_n(f_0) = \widetilde{O}(n^{-\alpha/(2\alpha+d-\alpha\beta)})$, where $\beta \in [0, d/\alpha]$ is a parameter depending on $f_0$ that characterizes the volume growth of the *level-sets* of the reference function $f_0$. (See assumption (A2), Proposition 2 and Theorem 1 for details). The rate matches the global minimax convergence rate $n^{-\alpha/(2\alpha+d)}$ for worst-case $f_0$ where $\beta = 0$, but can be much faster when $\beta > 0$. We emphasize that our algorithm has no knowledge of the reference function $f_0$ and achieves this rate adaptively.
2) We prove *local* minimax lower bounds that match the $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ upper bound, up to logarithmic factors in $n$. More specifically, we show that *even if $f_0$ is known*, no (active) algorithm can estimate $f$ in close neighborhoods of $f_0$ at a rate faster than $n^{-\alpha/(2\alpha+d-\alpha\beta)}$. We further show that, if active queries are not available and queries $x_1, \ldots, x_n$ are i.i.d. uniformly sampled from $\mathcal{X}$, then the $n^{-\alpha/(2\alpha+d)}$ global minimax rate also applies locally regardless of how large $\beta$ is. Thus, there is an explicit gap between local minimax rates in the active and uniform query models when $\beta$ is large.
3) In the special case when $f$ is *convex*, the global optimization problem is usually referred to as *zeroth-order convex optimization* and this problem has been widely studied [9]–[14]. Our results imply that, when $f_0$ is *strongly* convex and smooth, the local minimax rate $R_n(f_0)$ is on the order of $\widetilde{O}(n^{-1/2})$, which matches the convergence rates in [11]. Additionally, our negative results (Theorem 2) indicate that the $n^{-1/2}$ rate cannot be achieved if $f_0$ is merely convex, which seems to contradict $n^{-1/2}$ results in [13], [14] that do not require strong convexity of $f$. However, it should be noted that mere convexity of $f_0$ does *not* imply convexity of $f$ in a neighborhood of $f_0$ (e.g., $\|f - f_0\|_\infty \leqslant \varepsilon$). Our results show significant differences in the intrinsic difficulty of zeroth-order optimization of convex and near-convex functions.

## A. Related Work

*Global optimization*, known variously as *black-box optimization*, *Bayesian optimization* and the *continuum-armed bandit*, has a long history in the optimization research community [15], [16] and has also received a significant amount of recent interest in statistics and machine learning [1], [6], [8], [17]–[19]. Many previous works [17], [20] have derived rates for non-convex smooth payoffs in "continuum-armed" bandit problems.

The papers [21], [22] are closely related to our work. They studied the related problem of estimating the set of all optima of a smooth function in the Hausdorff distance. For Hölder smooth functions with polynomial growth, the paper [21] derives an $n^{-1/(2\alpha+d-\alpha\beta)}$ minimax rate for $\alpha < 1$ (subsequently improved to include $\alpha \geqslant 1$ in [23]). This result is similar to our Propositions 2 and 3. The papers [21], [22] also discussed adaptivity to unknown smoothness parameters. We however remark on several differences between our work and the papers [21], [22]. First, in [21], [22] only functions with polynomial growth are considered, while in our Theorems 1 and 2 functionals $\varepsilon_n^{\mathsf{U}}(f_0)$ and $\varepsilon_n^{\mathsf{L}}(f_0)$ are proposed for general reference functions $f_0$ satisfying mild regularity conditions, which include functions with polynomial growth as special cases. In addition, [21] considers the harder problem of estimating maxima sets in Hausdorff distance, as opposed to the problem of producing a single approximately optimal solution $\widehat{x}_T$. As a result, the minimax lower bounds in [21] do not apply to this latter setting. An algorithm, without distinguishing between two functions with different optima sets, can nevertheless produce a good approximate optimizer as long as the two functions under consideration have *overlapping* optima sets. New constructions and information-theoretic techniques are therefore required to prove lower bounds under the weaker (one-point) approximate optimization framework. Finally, we prove minimax lower bounds when only *uniform* query points are available and demonstrate a significant gap between algorithms having access to uniformly sampled or adaptively chosen data points.

The papers [18], [19] imposed additional assumptions on the level-sets of the underlying function to obtain an improved convergence rate. The level-set assumptions considered in the mentioned references are rather restrictive and essentially require the underlying function to be uni-modal, while our assumptions are much more flexible and apply to multi-modal functions as well. In addition, [18], [19] considered a *noiseless* setting in which exact function evaluations $f(x_t)$ can be obtained, while our paper studies the noise corrupted model in (1) for which vastly different convergence rates are derived. Finally, no matching lower bounds were proved in the papers [18], [19].

The (stochastic) global optimization problem is similar to *mode estimation* of either densities or regression functions, which has a rich literature [24]–[26]. An important difference between statistical mode estimation and global optimization is the way sample/query points $x_1, \ldots, x_n \in \mathcal{X}$ are distributed: in mode estimation it is customary to assume the samples are independently and identically distributed, while in global optimization sequential designs of samples/queries are typical. Furthermore, to estimate/locate the mode of an unknown density or regression function, such a mode has to be well-defined; on the other hand, producing an estimate $\widehat{x}_n$ with small $\mathfrak{L}(\widehat{x}_n, f)$ is easier and results in weaker conditions imposed on the underlying function.

Methodology-wise, our proposed algorithm is conceptually similar to the abstract *Pure Adaptive Search (PAS)* framework proposed and analyzed in [27]. The iterative procedure also resembles disagreement-based active learning methods [28]–[30] and the "successive rejection" algorithm in bandit problems [31]. The intermediate steps of candidate point elimination can also be viewed as level-set estimation problems [32]–[34] or cluster-tree estimation problems [35], [36] with active queries.

Another line of research has focused on *first-order* optimization of quasi-convex or non-convex functions [37]–[42], in which exact or unbiased evaluations of function *gradients* are available at query points $x \in \mathcal{X}$. The paper [42] considered a Cheeger's constant restriction on level-sets which is similar to our level-set regularity assumptions (A2 and A2'). The papers [43], [44] studied local minimax rates for the first-order optimization of convex functions. First-order optimization differs significantly from our setting because unbiased gradient estimation is generally impossible in the model of (1). Furthermore, most works on (first-order) non-convex optimization focus on obtaining stationary points or local minima, while we consider the problem of finding a (near) global minima.

### B. Comparison with the HOO Algorithm

The HOO algorithm [17], as well as similar algorithms such as Algorithm 2 in [45] and the POO algorithm in [22], are theoretically well-studied methods for global optimization. Below we summarize the differences of our results and the ones from these works.

(a) Weaker Smoothness Conditions I: In Algorithm 1, we use local polynomial estimation as a sub-routine to obtain local estimates of the objective function $f$. Compared to the sample average approach in HOO (e.g., Algorithm 2 in [45]), local polynomial estimates have the advantage of being unbiased for the estimation of low-degree polynomials. This translates to the improved (A1) Hölder-continuity condition that *only* restricts the $\lfloor\alpha\rfloor$-th order derivatives of objective functions. More specifically, the actual function values of $f(x)$ and $f(x')$ for $x, x'$ close to each other can be very different, as long as such differences can be perfectly modeled by low-degree polynomials. This is in contrast to the smoothness conditions imposed in [17], [45] which essentially require $f(x)$ to be close to $f(x^*)$ for $x$ close to $x^*$ the optima of $f$.

(b) Weaker Smoothness Conditions II: Our results in Section IV-C hold on functions that are only assumed to be smooth in regions close to its global minimum, in contrast to Definition 1 in [45] and many other existing works that place smoothness assumptions on the entire domain of the objective function $f$.

(c) Spatially Restricted Queries: Our proposed algorithm is "grid" based, and can be run on any sufficiently dense finite grid $G_n$ in $\mathcal{X}$ and does not need to have the capacity to query arbitrary points in $\mathcal{X}$. As a result, our algorithm can be run in experimental settings where queries are restricted to belong to a large pool of a-priori chosen points.

(d) Results for any Smooth Function: Our algorithm and lower bounds yield essentially tight results for the complexity of optimization of arbitrary smooth functions. While these rates are most interpretable under the level-set growth conditions (also studied in [45]) our results also yield nearly matching guarantees for other (arbitrary, smooth) functions $f_0$.

## II. Background and Notation

We first review standard asymptotic notation that will be used throughout this paper. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $\limsup_{n\to\infty} |a_n|/|b_n| < \infty$, or equivalently $b_n = \Omega(a_n)$ or $b_n \gtrsim a_n$. Denote $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We also write $a_n = o(b_n)$ or equivalently $b_n = \omega(a_n)$ if $\lim_{n\to\infty} |a_n|/|b_n| = 0$. For two sequences of random variables $\{A_n\}_{n=1}^{\infty}$ and $\{B_n\}_{n=1}^{\infty}$, denote $A_n = O_{\mathbb{P}}(B_n)$ if for every $\epsilon > 0$, there exists $C > 0$ such that $\limsup_{n\to\infty} \Pr[|A_n| > C|B_n|] \leqslant \epsilon$. For $r > 0$, $1 \leqslant p \leqslant \infty$ and $x \in \mathbb{R}^d$, we denote by $B_r^p(x) := \{z \in \mathbb{R}^d : \|z - x\|_p \leqslant r\}$ the $d$-dimensional $\ell_p$-ball of radius $r$ centered at $x$, where the vector $\ell_p$ norm is defined as $\|x\|_p := (\sum_{j=1}^d |x_j|^p)^{1/p}$ for $1 \leqslant p < \infty$ and $\|x\|_\infty := \max_{1\leqslant j\leqslant d} |x_j|$. For any subset $S \subseteq \mathbb{R}^d$ we denote by $B_r^p(x; S)$ the set $B_r^p(x) \cap S$.

### A. Passive and Active Query Models

Let $U$ be a known random quantity defined on a probability space $\mathcal{U}$. The following definitions characterize all passive and active optimization algorithms:

**Definition 1** (The passive query model)**.** *Let $x_1, \ldots, x_n$ be i.i.d. points uniformly sampled on $\mathcal{X}$ and $y_1, \ldots, y_n$ be observations from the model* (1). *A* passive optimization algorithm *$\mathcal{A}$ with $n$ queries is parameterized by a mapping $\phi_n : (x_1, y_1, \ldots, x_n, y_n, U) \mapsto \widehat{x}_n$ that maps the i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$ to an estimated optimum $\widehat{x}_n \in \mathcal{X}$, potentially randomized by $U$.*

**Definition 2** (The active query model)**.** *An* active optimization algorithm *can be parameterized by mappings $(\chi_1, \ldots, \chi_n, \phi_n)$, where for $t = 1, \ldots, n$,*

$$\chi_t : (x_1, y_1, \ldots, x_{t-1}, y_{t-1}, U) \mapsto x_t$$

*produces a query point $x_t \in \mathcal{X}$ based on previous observations* $\{(x_i, t_i)\}_{i=1}^{t-1}$, *and*

$$\phi_n : (x_1, y_1, \ldots, x_n, y_n, U) \mapsto \widehat{x}_n$$

*produces the final estimate. All mappings* $(\chi_1, \ldots, \chi_n, \phi_n)$ *can be randomized by* $U$.

### B. Local Minimax Rates

We use a classical *local minimax analysis* [46] to understand the fundamental information-theoretic limits of noisy global optimization of smooth functions. On the upper bound side, we seek (active) estimators $\widehat{x}_n$ such that

$$\sup_{f_0 \in \Theta} \sup_{f \in \Theta', \|f - f_0\|_\infty \leqslant \varepsilon_n(f_0)} \Pr_f \left[ \mathfrak{L}(\widehat{x}_n; f) \geqslant C_1 \cdot R_n(f_0) \right] \leqslant 1/4, \tag{4}$$

where $C_1 > 0$ is a positive constant. Here $f_0 \in \Theta$ is referred to as the *reference function*, and $f \in \Theta'$ is the true underlying function to be optimized, which is assumed to be "near" $f_0$ (in the $\ell_\infty$ norm). The minimax convergence rate of $\mathfrak{L}(\widehat{x}_n; f)$ is then characterized *locally* by $R_n(f_0)$ which depends on the reference function $f_0$. The constant of 1/4 is chosen arbitrarily and any small constant leads to similar conclusions. To establish negative results (i.e., local minimax lower bounds), in contrast to the upper bound formulation, we assume the potential active optimization estimator $\widehat{x}_n$ has *perfect knowledge* about the reference function $f_0 \in \Theta$. We then prove local minimax lower bounds of the form

$$\inf_{\widehat{x}_n} \sup_{f \in \Theta', \|f - f_0\|_\infty \leqslant \varepsilon_n(f_0)} \Pr_f \left[ \mathfrak{L}(\widehat{x}_n; f) \geqslant C_2 \cdot R_n(f_0) \right] \geqslant 1/3, \tag{5}$$

where $C_2 > 0$ is another positive constant and $\varepsilon_n(f_0)$, $R_n(f_0)$ are desired local convergence rates for functions near the reference $f_0$.

Although in some sense classical, the local minimax definition we propose warrants further discussion:

1) **Roles of $\Theta$ and $\Theta'$:** The reference function $f_0$ and the true functions $f$ are assumed to belong to different but closely related function classes $\Theta$ and $\Theta'$. In particular, in our paper $\Theta \subseteq \Theta'$, meaning that less restrictive assumptions are imposed on the true underlying function $f$ compared to those imposed on the reference function $f_0$ on which $R_n$ and $\varepsilon_n$ are based.
2) **Upper Bounds:** It is worth emphasizing that the estimator $\widehat{x}_n$ has no knowledge of the reference function $f_0$. From the perspective of upper bounds, we can consider the simpler task of producing $f_0$-dependent bounds (eliminating the second supremum) to instead study the (already interesting) quantity:
$$\sup_{f_0 \in \Theta} \Pr_{f_0} \left[ \mathfrak{L}(\widehat{x}_n; f_0) \geqslant C_1 R_n(f_0) \right] \leqslant 1/4.$$
As indicated above we maintain the double-supremum in the definition because fewer assumptions are imposed directly on the true underlying function $f$, and further because it allows to more directly compare our upper and lower bounds.
3) **Lower Bounds and the choice of the "localization radius" $\varepsilon_n(f_0)$:** Our lower bounds allow the estimator knowledge of the reference function (this makes establishing the lower bound more challenging). The lower bound in (5) implies that no estimator $\widehat{x}_n$ can effectively optimize a function $f$ close to $f_0$ beyond the convergence rate of $R_n(f_0)$, even if perfect knowledge of the reference function $f_0$ is available a priori. The $\varepsilon_n(f_0)$ parameter that decides the "range" in which local minimax rates apply is taken to be on the same order as the actual local rate $R_n(f_0)$ in this paper. This is (up to constants) the smallest radius for which we can hope to obtain non-trivial lower-bounds: if we consider a much smaller radius than $R_n(f_0)$ then the trivial estimator which outputs the minimizer of the reference function would achieve a faster rate than $R_n(f_0)$. On the other hand selecting the smallest possible radius makes establishing the lower bound most challenging but provides a refined picture of the complexity of zeroth-order optimization.

We remark that our primary motivation for the local-minimax analysis stems from the fact that for natural function classes the global-minimax rate for the optimization complexity is excessively pessimistic, while the local minimax analysis provides a more refined picture. In machine learning applications, there are several cases where the population risk is well-behaved (smooth, potentially non-convex) but we are only able to access/query the empirical risk which we want to minimize. Using standard concentration bounds the empirical risk and population risk are close, and the resulting problem is then to minimize the approximate-smooth empirical risk (see for instance [42], [47] for a more detailed discussion).

## III. Main Results

With this background in place we now turn our attention to our main results. We begin by collecting our assumptions about the true underlying function and the reference function in Section III-A. We state and discuss the consequences of our upper and lower bounds in Sections III-B and III-C respectively. We defer most technical proofs to Section V and turn our attention to our optimization algorithm in Section IV.

### A. Assumptions

We first state and motivate assumptions that will be used. The first assumption states that $f$ is locally Hölder smooth on its level-sets.

(A1) There exist constants $\kappa, \alpha, M, \zeta > 0$ such that $f$ restricted to $\mathcal{X}_{f,\kappa,\zeta} := \{x \in \mathcal{X} : \inf_{z \in \mathcal{X}, \|z - x\|_\infty \leqslant \zeta} f(z) \leqslant f^* + \kappa\}$ belongs to the Hölder class $\Sigma^\alpha(M)$, meaning that $f$ is $k$-times differentiable on $\mathcal{X}_{f,\kappa,\zeta}$ and furthermore for any $x, x' \in \mathcal{X}_{f,\kappa,\zeta}$[3],

$$\sum_{\alpha_1 + \ldots + \alpha_d = k} \frac{|f^{(\boldsymbol{\alpha},k)}(x) - f^{(\boldsymbol{\alpha},k)}(x')|}{\|x - x'\|_\infty^{\alpha - k}} \leqslant M. \tag{6}$$

[3]We use the $\ell_\infty$-norm for convenience and it can be replaced by any equivalent vector norm.

Here $k = \lfloor \alpha \rfloor$ is the largest integer lower bounding $\alpha$ and $f^{(\boldsymbol{\alpha},j)}(x) := \partial^j f(x)/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}$.

We use $\Sigma_\kappa^\alpha(M)$ to denote the class of all functions satisfying (A1). We remark that (A1) is weaker than the usual Hölder assumption in two ways. First, (6) only imposes stability conditions on the $\lfloor\alpha\rfloor$-th order derivatives of the function $f$, in contrast to conditions involving all orders of derivatives in previous works [17], [45]. Second, (A1) only imposes the Hölder smoothness assumption on certain regions of $\mathcal{X}$, because regions with function values larger than $f^* + \kappa$ can be easily detected and removed by a pre-processing step, highlighting an important difference between optimization and $\ell_\infty$-norm estimation. We give further details of the pre-processing step in Section IV-C.

Our next assumption concerns the "regularity" of the *level-sets* of the "reference" function $f_0$. Define $L_{f_0}(\epsilon) := \{x \in \mathcal{X} : f_0(x) \leqslant f_0^* + \epsilon\}$ as the $\epsilon$-level-set of $f_0$, and $\mu_{f_0}(\epsilon) := \lambda(L_{f_0}(\epsilon))$ as the Lebesgue measure of $L_{f_0}(\epsilon)$, which we refer to as the *distribution function*. Define, $N(L_{f_0}(\epsilon), \delta)$ as the smallest number of $\ell_2$-balls of radius $\delta$ that cover $L_{f_0}(\epsilon)$. Then we make the following assumption:

(A2) There exist constants $c_0 > 0$ and $C_0 > 0$ such that $N(L_{f_0}(\epsilon), \delta) \leqslant C_0[1 + \mu_{f_0}(\epsilon)\delta^{-d}]$ for all $\epsilon, \delta \in (0, c_0]$.

We use $\Theta_\mathbf{C}$ to denote all functions that satisfy (A2) with respect to parameters $\mathbf{C} = (c_0, C_0)$.

At a high-level, the regularity condition (A2) assumes that the level-sets are sufficiently "regular" such that covering them with small-radius balls does not require significantly larger total volume. For example, consider the perfectly regular case when $L_{f_0}(\epsilon)$ is the $d$-dimensional $\ell_2$ ball of radius $r$: $L_{f_0}(\epsilon) = \{x \in \mathcal{X} : \|x - x^*\|_2 \leqslant r\}$. Clearly, $\mu_{f_0}(\epsilon) \asymp r^d$. In addition, the $\delta$-covering number in $\ell_2$ of $L_{f_0}(\epsilon)$ is on the order of $1 + (r/\delta)^d \asymp 1 + \mu_{f_0}(\epsilon)\delta^{-d}$, which satisfies the scaling in (A2).

When (A2) holds, uniform confidence intervals for $f$ on its level-sets are easier to construct because little statistical efficiency is lost by slightly enlarging the level-sets so that complete (sufficiently small) $d$-dimensional cubes are contained in the enlarged level-sets. On the other hand, when regularity of level-sets fails to hold such nonparametric estimation can be very difficult or even impossible. As an extreme example, suppose the level-set $L_{f_0}(\epsilon)$ consists of $\mathfrak{n}$ standalone and well-spaced points in $\mathcal{X}$: the Lebesgue measure of $L_{f_0}(\epsilon)$ would be zero, but at least $\Omega(\mathfrak{n})$ queries are necessary to construct uniform confidence intervals on $L_{f_0}(\epsilon)$. It is clear that such $L_{f_0}(\epsilon)$ violates (A2), because $N(L_{f_0}(\epsilon), \delta) \geqslant \mathfrak{n}$ as $\delta \to 0^+$ but $\mu_{f_0}(\epsilon) = 0$.

### B. Upper Bound

The following theorem is our main result that provides an upper bound on the local minimax rate of noisy global optimization with active queries.

**Theorem 1.** *For any $\alpha, M, \kappa, c_0, C_0 > 0$ and $f_0 \in \Sigma_\kappa^\alpha(M) \cap \Theta_\mathbf{C}$, where $\mathbf{C} = (c_0, C_0)$, define*

$$\varepsilon_n^{\mathsf{U}}(f_0) := \sup\left\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n/\log^\omega n\right\}, \quad (7)$$

*where $\omega > 5 + d/\alpha$ is a large constant. Suppose also that $\varepsilon_n^{\mathsf{U}}(f_0) \to 0$ as $n \to \infty$. Then for sufficiently large $n$, there exists an estimator $\widehat{x}_n$ with access to $n$ active queries $x_1, \ldots, x_n \in \mathcal{X}$, a constant $C_R > 0$ depending only on $\alpha, M, \kappa, c, c_0, C_0$ and a constant $\gamma > 0$ depending only on $\alpha$ and $d$ such that*

$$\sup_{f_0 \in \Sigma_\kappa^\alpha(M) \cap \Theta_\mathbf{C}} \; \sup_{\substack{f \in \Sigma_\kappa^\alpha(M), \\ \|f - f_0\|_\infty \leqslant \varepsilon_n^{\mathsf{U}}(f_0)}} \Pr_f\Big[\mathfrak{L}(\widehat{x}_n, f) > C_R \log^\gamma n \cdot (\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2})\Big] \leqslant 1/4. \quad (8)$$

**Remark 1.** *Unlike the (local) smoothness class $\Sigma_\kappa^\alpha(M)$, the additional function class $\Theta_\mathbf{C}$ that encapsulates (A2) is imposed only on the "reference" function $f_0$ but not the true function $f$ to be estimated. This makes the assumptions considerably weaker because the true function $f$ may violate (A2) while our results remain valid.*

**Remark 2.** *The estimator $\widehat{x}_n$ does* not *require knowledge of parameters $\kappa, c_0, C_0$ or $\varepsilon_n^{\mathsf{U}}(f_0)$, and automatically adapts to them, as shown in the next section. While the knowledge of smoothness parameters $\alpha$ and $M$ is in general unavoidable in non-parametric regression (see [48]), in the zeroth-order optimization problem it is possible to adapt to $\alpha$ and $M$ by running $O(\log^2 n)$ parallel sessions of $\widehat{x}_n$ on $O(\log n)$ grids of $\alpha$ and $M$ values, and then using $\Omega(n/\log^2 n)$ single-point queries to decide on the location with the smallest function value. This adaptive strategy was suggested in [22] to remove an additional condition in [21], and also applies to our setting.*

**Remark 3.** *When the distribution function $\mu_{f_0}(\epsilon)$ does not change abruptly with $\epsilon$ the expression of $\varepsilon_n^{\mathsf{U}}(f_0)$ can be significantly simplified. In particular, if for all $\epsilon \in (0, c_0]$ it holds that*

$$\mu_{f_0}(\epsilon/\log n) \geqslant \mu_{f_0}(\epsilon)/[\log n]^{O(1)}, \quad (9)$$

*then $\varepsilon_n^{\mathsf{U}}(f_0)$ can be upper bounded as*

$$\varepsilon_n^{\mathsf{U}}(f_0) \leqslant [\log n]^{O(1)} \cdot \sup\left\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n\right\}. \quad (10)$$

*If $\mu_{f_0}(\epsilon)$ scales polynomially with $\epsilon$, i.e. $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ for some constant $\beta \geqslant 0$, then* (9) *and* (10) *are both satisfied.*

The quantity $\varepsilon_n^{\mathsf{U}}(f_0) = \sup\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n/\log^\omega n\}$ is crucial in determining the convergence rate of optimization error of $\widehat{x}_n$ *locally* around the reference function $f_0$. While the definition of $\varepsilon_n^{\mathsf{U}}(f_0)$ is mostly implicit and involves solving an inequality involving the distribution function $\mu_{f_0}(\cdot)$, we remark that it admits a simple form when $\mu_{f_0}$ has a polynomial growth rate similar to a local Tsybakov noise condition [4], [49], as shown in the following proposition:

**Proposition 2.** *Suppose $\mu_{f_0}(\epsilon) \lesssim \epsilon^\beta$ for some constant $\beta \in [0, 2 + d/\alpha)$. Then $\varepsilon_n^{\mathsf{U}}(f_0) = \widetilde{O}(n^{-\alpha/(2\alpha+d-\alpha\beta)})$. In addition, if $\beta \in [0, d/\alpha]$ then $\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2} \lesssim \varepsilon_n^{\mathsf{U}}(f_0) = \widetilde{O}(n^{-\alpha/(2\alpha+d-\alpha\beta)})$.*

We remark that, following Proposition 1 of [45], $\alpha$, $\beta$ and $d$ must satisfy the relationship that $\beta \leqslant d/\alpha$. Proposition 2 can be easily verified by solving the system $\varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n/\log^{\omega} n$ with the condition $\mu_{f_0}(\epsilon) \lesssim \epsilon^{\beta}$. We therefore omit its proof. The following two examples give some simple reference functions $f_0$ that satisfy the $\mu_{f_0}(\epsilon) \lesssim \epsilon^{\beta}$ condition in Proposition 2 with particular values of $\beta$.

**Example 1.** *The constant function $f_0 \equiv 0$ satisfies (A1) through (A3) with $\beta = 0$.*

**Example 2.** *$f_0 \in \Sigma_{\kappa}^{2}(M)$ that is* strongly convex[4] *satisfies (A1) through (A3) with $\beta = d/2$.*

Example 1 is simple to verify, as the volume of level-sets of the constant function $f_0 \equiv 0$ exhibit a phase transition at $\epsilon = 0$ and $\epsilon > 0$. Consequently, $\beta = 0$ is the only parameter for which $\mu_{f_0}(\epsilon) \lesssim \epsilon^{\beta}$. Example 2 is more involved, and holds because the strong convexity of $f_0$ *lower bounds* the growth rate of $f_0$ when moving away from its minimum. We give a rigorous proof for Example 2 in the appendix. We also remark that $f_0$ does *not* need to be exactly strongly convex for $\beta = d/2$ to hold, and the example is valid for, e.g., piecewise strongly convex functions with a constant number of pieces too.

To best interpret the results in Theorem 1 and Proposition 2, it is instructive to compare the "local" rate $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ with the baseline rate $n^{-\alpha/(2\alpha+d)}$, which can be attained by reconstructing $f$ in sup-norm and applying Proposition 1. Since $\beta \geqslant 0$, the local convergence rate established in Theorem 1 is never slower, and the improvement compared to the baseline rate $n^{-\alpha/(2\alpha+d)}$ is dictated by $\beta$, which governs the growth rate of volume of level-sets of the reference function $f_0$. In particular, for functions that grows fast when moving away from its minimum, the parameter $\beta$ is large and therefore the local convergence rate around $f_0$ could be much faster than $n^{-\alpha/(2\alpha+d)}$.

Theorem 1 also implies concrete convergence rates for special functions considered in Examples 1 and 2. For the constant reference function $f_0 \equiv 0$, Example 1 and Theorem 1 yield that $R_n(f_0) \asymp n^{-\alpha/(2\alpha+d)}$, which matches the baseline rate $n^{-\alpha/(2\alpha+d)}$ and suggests that $f_0 \equiv 0$ is the worst-case reference function. This is intuitive, because $f_0 \equiv 0$ has a drastic level-set change at $\epsilon \to 0^+$ and therefore small perturbations of $f_0$ result in changes to the optimal location. On the other hand, if $f_0$ is strongly smooth and convex as in Example 2, Theorem 1 leads to the bound of $R_n(f_0) \asymp n^{-1/2}$, which is significantly better than the $n^{-2/(4+d)}$ baseline rate[5] and also matches existing works on zeroth-order optimization of convex functions [11]. The faster rate holds intuitively because strongly convex functions grow quickly when moving away from the minimum. An active query algorithm can focus most of its queries on the small level-sets of the underlying function, resulting in more accurate local function reconstruction and faster optimization error rate.

Our proof of Theorem 1 is constructive, by upper bounding the local minimax optimization error of an explicit algorithm. Roughly, our algorithm partitions the $n$ active queries evenly into $\log n$ epochs, and level-sets of $f$ are estimated at the end of each epoch by comparing (uniform) confidence intervals on a dense grid on $\mathcal{X}$. It is then proved that the volume of the estimated level-sets contracts *geometrically*, until the target convergence rate $R_n(f_0)$ is attained. The algorithm is described in more detail in Section IV and the complete proof of Theorem 1 is in Section V-B.

[4]A twice differentiable function $f_0$ is strongly convex if there exists $\sigma > 0$ such that $\nabla^2 \ f_0(x) \succeq \sigma I, \forall x \in \mathcal{X}$.

[5]Note that $f_0$ being strongly smooth corresponds to $\alpha = 2$ in the local smoothness assumption.

### C. Lower Bounds

We prove local minimax lower bounds that match the upper bounds in Theorem 1 up to logarithmic terms. As we remarked in Section II-B, in the local minimax lower bound formulation we assume the data analyst has full knowledge of the reference function $f_0$, which makes the lower bounds stronger as more information is available a priori.

To facilitate such local minimax lower bounds, the following additional condition is imposed on the reference function $f_0$ of which the data analyst has perfect information.

(A2') There exist constants $c_0', C_0' > 0$ such that $M(L_{f_0}(\epsilon), \delta) \geqslant C_0' \mu_{f_0}(\epsilon)\delta^{-d}$ for all $\epsilon, \delta \in (0, c_0']$, where $M(L_{f_0}(\epsilon), \delta)$ is the maximum number of disjoint $\ell_2$ balls of radius $\delta$ that can be packed into $L_{f_0}(\epsilon)$.

We denote $\Theta'_{\mathbf{C}'}$ as the class of functions that satisfy (A2') with respect to parameters $\mathbf{C}' = (c_0', C_0') > 0$. Intuitively, (A2') can be regarded as a converse of (A2).

We are now ready to state our main negative result, which shows, from an information-theoretic perspective, that the upper bound in Theorem 1 is not improvable.

**Theorem 2.** *Suppose $\alpha, c_0, C_0, c_0', C_0' > 0$ and $\kappa = \infty$. Denote $\mathbf{C} = (c_0, C_0)$ and $\mathbf{C}' = (c_0', C_0')$. For any $f_0 \in \Theta_{\mathbf{C}} \cap \Theta'_{\mathbf{C}'}$, define*

$$\varepsilon_n^{\mathsf{L}}(f_0) := \sup\left\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n\right\}. \tag{11}$$

*Then there exists a constant $M > 0$ depending on $\alpha, d, \mathbf{C}$ and $\mathbf{C}'$ such that, for any $f_0 \in \Sigma_{\kappa}^{\alpha}(M/2) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$,*

$$\inf_{\widehat{x}_n} \sup_{\substack{f \in \Sigma_{\kappa}^{\alpha}(M), \\ \|f - f_0\|_\infty \leqslant 2\varepsilon_n^{\mathsf{L}}(f_0)}} \Pr_f\left[\mathfrak{L}(\widehat{x}_n; f) \geqslant \varepsilon_n^{\mathsf{L}}(f_0)\right] \geqslant \frac{1}{3}. \tag{12}$$

**Remark 4.** *We note in passing that for any $f_0$ and n it always holds that $\varepsilon_n^{\mathsf{L}}(f_0) \leqslant \varepsilon_n^{\mathsf{U}}(f_0)$.*

**Remark 5.** *If the distribution function $\mu_{f_0}(\epsilon)$ satisfies* (9) *(i.e. it does not change too abruptly) in Remark 3, then $\varepsilon_n^{\mathsf{L}}(f_0) \geqslant \varepsilon_n^{\mathsf{U}}(f_0)/[\log n]^{O(1)}$. Consequently, the upper and lower bounds for these functions match up to logarithmic factors.*

The following proposition derives an explicit expression for $\varepsilon_n^{\mathsf{L}}(f_0)$ for reference functions whose distribution functions have a polynomial growth, which matches the upper bound in Proposition 2 up to $\log n$ factors. The proof of this Proposition is straightforward and is omitted.

**Proposition 3.** *Suppose $\mu_{f_0}(\epsilon) \gtrsim \epsilon^{\beta}$ for some $\beta \in [0, 2 + d/\alpha)$. Then $\varepsilon_n^{\mathsf{L}}(f_0) = \Omega(n^{-\alpha/(2\alpha+d-\alpha\beta)})$.*

The following proposition additionally shows the existence of $f_0 \in \Sigma_\infty^\alpha(M) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$ that satisfies $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ for any values of $\alpha > 0$ and $\beta \in [0, d/\alpha]$. Its proof is given in the Appendix.

**Proposition 4.** *Fix arbitrary $\alpha, M > 0$ and $\beta \in [0, d/\alpha]$. There exists $f_0 \in \Sigma_\kappa^\alpha(M) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$ for $\kappa = \infty$ and constants $\mathbf{C} = (c_0, C_0)$, $\mathbf{C}' = (c_0', C_0')$ that depend only on $\alpha, \beta, M$ and $d$ such that $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$.*

Theorem 2 and Proposition 3 show that the $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ upper bound on local minimax convergence rate established in Theorem 1 is not improvable up to logarithmic factors of $n$. Such information-theoretic lower bounds on the convergence rates hold *even if the data analyst has perfect information of* $f_0$, the reference function on which the $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ local rate is based. Our results also imply an $n^{-\alpha/(2\alpha+d)}$ minimax lower bound over all $\alpha$-Hölder smooth functions, showing that without additional assumptions, noisy optimization of smooth functions is as difficult as reconstructing the unknown function in sup-norm.

Our proof of Theorem 2 also differs from those of existing minimax lower bounds for active nonparametric models [50]. The classical approach is to invoke Fano's inequality and to upper bound the KL divergence between different underlying functions $f$ and $g$ using $\|f - g\|_\infty$, corresponding to the point $x \in \mathcal{X}$ that leads to the largest KL divergence. Such an approach, however, does not produce tight lower bounds for our problem. To overcome such difficulties, we borrow the lower bound analysis for bandit pure exploration problems in [51]. In particular, our analysis considers the query distribution of any active query algorithm $\mathcal{A} = (\varphi_1, \ldots, \varphi_n, \phi_n)$ under the reference function $f_0$ and bounds the perturbation in query distributions between $f_0$ and $f$ using Le Cam's lemma. Afterwards, an adversarial function choice $f$ can be made based on the query distributions of the considered algorithm $\mathcal{A}$. We defer the complete proof of Theorem 2 to Section V-C.

Theorem 2 applies to any global optimization method that makes *active* queries, corresponding to the query model in Definition 2. The following theorem, on the other hand, shows that for passive algorithms (Definition 1) the $n^{-\alpha/(2\alpha+d)}$ optimization rate is not improvable even with additional level-set assumptions imposed on $f_0$. This demonstrates an explicit gap between passive and adaptive query models in global optimization problems.

**Theorem 3.** *Suppose $\alpha, c_0, C_0, c_0', C_0' > 0$ and $\kappa = \infty$. Denote $\mathbf{C} = (c_0, C_0)$ and $\mathbf{C}' = (c_0', C_0')$. Then there exist constants $M > 0$ depending on $\alpha, d, \mathbf{C}, \mathbf{C}'$ and $N$ depending on $M$ such that, for any $f_0 \in \Sigma_\kappa^\alpha(M/2) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$ satisfying $\varepsilon_n^{\mathsf{L}}(f_0) \leqslant \widetilde{\varepsilon}_n^{\mathsf{L}} =: [\log n/n]^{\alpha/(2\alpha+d)}$,*

$$\inf_{\check{x}_n} \sup_{\substack{f \in \Sigma_\kappa^\alpha(M), \\ \|f - f_0\|_\infty \leqslant 2\widetilde{\varepsilon}_n^{\mathsf{L}}}} \Pr_f\left[\mathfrak{L}(\widehat{x}_n; f) \geqslant \widetilde{\varepsilon}_n^{\mathsf{L}}\right] \geqslant \frac{1}{3} \quad \text{for all } n \geqslant N. \tag{13}$$

Intuitively, the apparent gap demonstrated by Theorems 2 and 3 between the active and passive query models stems from the observation that, a passive algorithm $\mathcal{A}$ only has access to uniformly sampled query points $x_1, \ldots, x_n$ and therefore cannot focus on a small level-set of $f$ in order to improve query efficiency. In addition, for functions that grow faster when moving away from their minima (implying a larger value of $\beta$), the gap between passive and active query models becomes bigger as active queries can more effectively exploit the restricted level-sets of such functions.

## IV. Our Algorithm

In this section we describe a concrete algorithm that attains the upper bound in Theorem 1. We start with a cleaner algorithm that operates under the slightly stronger condition that $\kappa = \infty$ in (A1), meaning that $f$ is $\alpha$-Hölder smooth on the entire domain $\mathcal{X}$. The generalization to $\kappa > 0$ being a constant is given in Section IV-C with an additional pre-processing step.

Let $G_n \in \mathcal{X}$ be a *finite* grid of points in $\mathcal{X}$. We assume the finite grid $G_n$ satisfies the following two mild conditions:

(B1) Points in $G_n$ are sampled i.i.d. from an unknown distribution $P_X$ on $\mathcal{X}$; furthermore, the density $p_X$ associated with $P_X$ satisfies $\underline{p}_0 \leqslant p_X(x) \leqslant \overline{p}_0$ for all $x \in \mathcal{X}$, where $0 < \underline{p}_0 \leqslant \overline{p}_0 < \infty$ are universal constants;

(B2) $|G_n| \gtrsim n^3$ and $\log|G_n| = O(\log n)$.

**Remark 6.** *Although typically the choices of the grid points $G_n$ belong to the data analyst, in some applications the choices of design points are not completely unconstrained. For example, in material synthesis experiments described previously some environmental parameter settings (e.g., temperature and pressure) might not be allowed due to budget or physical constraints. Thus, we choose to consider less restrictive conditions imposed on the design grid $G_n$, allowing it to be more flexible in real-world applications.*

**Remark 7.** *Condition (B2) ensures that the grid $G_n$ is sufficiently dense, such that even with the smallest bandwidth our algorithm possibly uses ($h_t(x) = 1/n^2$, see (18)), each $x \in G_n$ has abundant neighboring points in $G_n$, so that the local polynomial estimates in* (15) *are well-defined.*

For any subset $S \subseteq G_n$ and a "weight" function $\varrho : G_n \to \mathbb{R}^+$, define the *extension* $S^\circ(\varrho)$ of $S$ with respect to $\varrho$ as

$$S^\circ(\varrho) := \bigcup_{x \in S} B_{\varrho(x)}^\infty(x; G_n) \quad \text{where}$$
$$B_{\varrho(x)}^\infty(x; G_n) = \{z \in G_n : \|z - x\|_\infty \leqslant \varrho(x)\}. \tag{14}$$

The algorithm can then be formulated as two levels of iterations, with the outer loop shrinking the "active set" $S_\tau$ and the inner loop collecting data in order to reduce the lengths of the confidence intervals on the points in the active set. A pseudocode description of our proposed algorithm is given in Figure 1.

### A. Local Polynomial Regression

We use local polynomial regression [5] to obtain the estimate $\widehat{f}$. In particular, for any $x \in G_n$ and a bandwidth

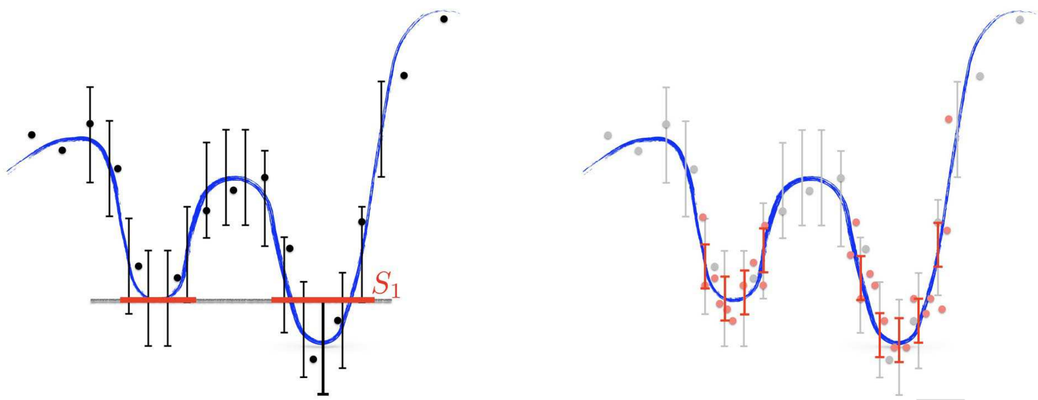


Fig. 1. An informal illustration of Algorithm 1. Solid blue curves depict the underlying function $f$ to be optimized, black and red solid dots denote the query points and their responses $\{(x_t, y_t)\}$, and black and red vertical line segments correspond to uniform confidence intervals on function evaluations constructed using the current batch of data observed. The left figure illustrates the first epoch of our algorithm, where query points are uniformly sampled from the entire domain $\mathcal{X}$. Afterwards, sub-optimal locations based on the constructed confidence intervals are removed, and a shrunken "candidate set" $S_1$ is obtained. The algorithm then proceeds to the second epoch, illustrated in the right figure, where query points (in red) are sampled only from the restricted candidate set and shorter confidence intervals (also in red) are constructed and updated. The procedure is repeated until $O(\log n)$ epochs are completed.

**Parameters**: $\alpha,\ M,\ \delta,\ n$
**Output**: $\widehat{x}_n$, the final prediction
Initialization: $S_0 = G_n$, $\varrho_0(x) \equiv \infty$, $T = \lfloor \log_2\ n \rfloor$, $n_0 = \lfloor n/T \rfloor$;
**for** $\tau = 1, 2, \ldots, T$ **do**
  Compute "extended" sample set $S^{\circ}_{\tau-1}(\varrho_{\tau-1})$ defined in (14);
  **for** $t = (\tau-1)n_0 + 1$ *to* $\tau n_0$ **do**
    Sample $x_t$ uniformly at random from $S^{\circ}_{\tau-1}(\varrho_{\tau-1})$ and observe $y_t = f(x_t) + w_t$;
  **end**
  For every $x \in S_{\tau-1}$, compute bandwidth $h_\tau(x)$ using (18) and build the confidence interval $[\ell_\tau(x), u_\tau(x)]$ in (19);
  $S_\tau := \{x \in S_{\tau-1} : \ell_\tau(x) \leqslant \min_{x' \in S_{\tau-1}} u_\tau(x')\}$,
  $\varrho_\tau(x) := \min\{\varrho_{\tau-1}(x), h_\tau(x)\}$,
**end**
Final processing: for every $x \in S_T$ let $\widehat{f}_{h_T,x}(\cdot)$ be the local polynomial estimates constructed in (15) at $x$.
Output $\widehat{x}_n = \arg\min_{x \in S_T} \min_{x' \in B^{\infty}_{h_T(x)}(x;\mathcal{X})} \widehat{f}_{h_T,x}(x')$.

**Algorithm 1** The Main Algorithm

parameter $h > 0$, consider the least squares polynomial estimate

$$\widehat{f}_h \in \operatorname*{argmin}_{g \in \mathcal{P}_k} \sum_{t'=1}^{t} \mathbb{I}[x_{t'} \in B^{\infty}_h(x)] \cdot (y_{t'} - g(x_{t'}))^2, \qquad (15)$$

where $B^{\infty}_h(x) := \{x' \in \mathcal{X} : \|x' - x\|_\infty \leqslant h\}$ and $\mathcal{P}_k$ denotes all polynomials of degree $k$ on $\mathcal{X}$.

To analyze the performance of $\widehat{f}_h$ evaluated at a certain point $x \in \mathcal{X}$, define the mapping

$$\psi_{x,h} : z \mapsto (1, \psi^1_{x,h}(z), \ldots, \psi^k_{x,h}(z))$$

where $\psi^j_{x,h} : z \mapsto [\prod_{\ell=1}^{j} h^{-1}(z_{i_\ell} - x_{i_\ell})]^d_{i_1,\ldots,i_j=1}$ is the degree-$j$ polynomial mapping from $\mathbb{R}^d$ to $\mathbb{R}^{d^j}$. Also define $\Psi_{t,h} := (\psi_{x,h}(x_{t'}))_{1 \leqslant t' \leqslant t, x_{t'} \in B_h(x)}$ as the $m \times D$ aggregated design matrix, where $m = \sum_{t'=1}^{t} \mathbb{I}[x_{t'} \in B^{\infty}_h(x)]$ and $D = 1 + d + \ldots + d^k$, $k = \lfloor \alpha \rfloor$. The estimate $\widehat{f}_h$ defined in (15) then admits the following closed-form expression:

$$\widehat{f}_h(z) \equiv \psi_{x,h}(z)^\top (\Psi_{t,h}^\top \Psi_{t,h})^\dagger \Psi_{t,h}^\top Y_{t,h}, \qquad (16)$$

where $Y_{t,h} = (y_{t'})_{1 \leqslant t' \leqslant t, x_{t'} \in B^{\infty}_h(x)}$ and $A^\dagger$ is the Moore-Penrose pseudo-inverse of $A$.

The following lemma gives a finite-sample analysis of the error of $\widehat{f}_h(x)$:

**Lemma 1.** *Suppose that $f$ satisfies (6) on $B^{\infty}_h(x;\mathcal{X})$, $\max_{z \in B^{\infty}_h(x;\mathcal{X})} \|\psi_{x,h}(z)\|_2 \leqslant b$ and $\frac{1}{m}\Psi_{t,h}^\top \Psi_{t,h} \succeq \sigma I_{D \times D}$ for some $\sigma > 0$. Then for any $\delta \in (0, 1/2)$, with probability $1-\delta$*

$$\left|\widehat{f}_h(x') - f(x')\right| \leqslant \underbrace{\frac{b^2}{\sigma} M d^k h^\alpha}_{\mathfrak{b}_{h,\delta}(x)} + \underbrace{b\sqrt{\frac{5D\ln(1/\delta)}{\sigma m}}}_{\mathfrak{s}_{h,\delta}(x)} =: \eta_{h,\delta}(x), \quad \forall x' \in B^{\infty}_h(x;\mathcal{X}). \qquad (17)$$

**Remark 8.** *$\mathfrak{b}_{h,\delta}(x)$, $\mathfrak{s}_{h,\delta}(x)$ and $\eta_{h,\delta}(x)$ depend on $x$ because $\sigma$ depends on $\Psi_{t,h}$, which further depends on the sample points in the neighborhood $B^{\infty}_h(x;\mathcal{X})$ of $x$.*

In the rest of the paper we define $\mathfrak{b}_{h,\delta}(x) := (b^2/\sigma) M d^k h^\alpha$ and $\mathfrak{s}_{h,\delta}(x) := b\sqrt{5D\ln(1/\delta)/\sigma m}$ as the bias and standard deviation terms in the error of $\widehat{f}_h(x)$, respectively. We also denote $\eta_{h,\delta}(x) := \mathfrak{b}_{h,\delta}(x) + \mathfrak{s}_{h,\delta}(x)$ as the overall error in $\widehat{f}_h(x)$.

Notice that when bandwidth $h$ increases, the bias term $\mathfrak{b}_{h,\delta}(x)$ increases because of the $h^\alpha$ term; on the other hand, with $h$ increasing the local neighborhood $B^{\infty}_h(x;\mathcal{X})$ grows and would potentially contain more samples, implying a larger $m$ and smaller standard deviation term $\mathfrak{s}_{h,\delta}(x)$. A careful selection of the bandwidth $h$ balances $\mathfrak{b}_{h,\delta}(x)$ and $\mathfrak{s}_{h,\delta}(x)$ and yields appropriate confidence intervals on $f(x)$, and we turn our attention to this in the next section.

### B. Bandwidth Selection and Confidence Intervals

Given the expressions of bias $\mathfrak{b}_{h,\delta}(x)$ and standard deviation $\mathfrak{s}_{h,\delta}(x)$ in (17), the bandwidth $h_\tau(x) > 0$ at epoch $\tau$ and point $x$ is selected as

$$h_\tau(x) := \frac{j_\tau(x)}{n^2} \quad \text{where} \quad j_\tau(x) := \arg\max\big\{j \in \mathbb{N}^+, \; j \leqslant n^2 : \mathfrak{b}_{j/n^2,\delta}(x) \leqslant \mathfrak{s}_{j/n^2,\delta}(x)\big\}. \tag{18}$$

More specifically, $h_\tau(x)$ is the largest positive value in an evenly spaced grid $\{j/n^2\}$ such that the bias of $\widehat{f}_{h_\tau}(x)$ is smaller than its standard deviation. This bandwidth selection is in principle similar to the Lepski's method [52], with the exception that an upper bound on the bias for any bandwidth parameter is known and does not need to be estimated from data.

With the selection of bandwidth $h_\tau(x)$ at epoch $\tau$ and query point $x$, a confidence interval on $f(x')$ for all $x' \in B^\infty_{h_\tau(x)}(x; \mathcal{X})$ is constructed as

$$\ell_\tau(x) := \max_{1\leqslant t'\leqslant \tau} \sup_{x'\in B^\infty_{h_\tau(x)}(x;\mathcal{X})} \left\{\widehat{f}_{h_{t'}(x)}(x') - \eta_{h_{t'}(x),\delta}(x)\right\};$$
$$u_\tau(x) := \min_{1\leqslant t'\leqslant \tau} \inf_{x'\in B^\infty_{h_\tau(x)}(x;\mathcal{X})} \left\{\widehat{f}_{h_{t'}(x)}(x') + \eta_{h_{t'}(x),\delta}(x)\right\}. \tag{19}$$

Note that for any $x \in \mathcal{X}$, the lower confidence edge $\ell_\tau(x)$ is a non-decreasing function in $\tau$ and the upper confidence edge $u_\tau(x)$ is a non-increasing function in $\tau$.

### C. Pre-processing

We describe a pre-processing step that relaxes the smoothness condition from $\kappa = \infty$ to $\kappa = \Omega(1)$, meaning that only local smoothness of $f$ around its minimum values is required. Let $n_0 = \lfloor n/\log n \rfloor$, $x_1, \ldots, x_{n_0}$ be points i.i.d. uniformly sampled from $\mathcal{X}$ and $y_1, \ldots, y_{n_0}$ be their corresponding responses. For every grid point $x \in G_n$, perform the following:

1) Compute $\check{f}_x(\cdot)$ as the local polynomial fits of all $y_i$ corresponding to $\|x_i - x\|_\infty \leqslant n_0^{-1/2d} \log^3 n =: h_0$;
2) Compute $\overline{f}(x)$ as the sample average of all $y_i$ corresponding to $\|x_i - x\|_\infty \leqslant h_0$;
3) Remove all $x \in G_n$ from $S_0$ if $\overline{f}(x) \geqslant \min_{z\in G_n} \inf_{z'\in B^\infty_{h_0}(z;\mathcal{X})} \check{f}_z(z') + 1/\log n$.

**Remark 9.** *The $1/\log n$ term in the removal condition $\check{f}(x) \geqslant \min_{z\in G_n} \check{f}(z) + 1/\log n$ is not important, and can be replaced with any sequence $\{\omega_n\}$ such that $\lim_{n\to\infty} \omega_n = 0$ and $\lim_{n\to\infty} \omega_n n^t = \infty$ for any $t > 0$. The readers are referred to the proof of Proposition 5 in the appendix for the motivation of this term as well as the selection of the pre-processing bandwidth $h_0$.*

To analyze the pre-processing step, we state the following proposition:

**Proposition 5.** *Assume $f \in \Sigma^\alpha_\kappa(M)$ and let $S_0'$ be the screened grid after step 2 of the pre-processing procedure. Then for sufficiently large $n$, with probability $1 - O(n^{-1})$ we have*

$$B^\infty_{h_0}(x; \mathcal{X}) \cap L_f(\kappa/2) \neq \varnothing, \qquad \forall x \in S_0', \tag{20}$$

*where $L_f(\kappa/2) = \{x \in \mathcal{X} : f(x) \leqslant f^* + \kappa/2\}$.*

To interpret Proposition 5, note that for sufficiently large $n$, $f \in \Sigma^\alpha_\kappa(M)$ implies $f$ being $\alpha$-Hölder smooth (i.e., $f$ satisfies (6)) on $\bigcup_{x\in L_f(\kappa/2)} B^\infty_{h_0}(x; \mathcal{X})$, because $\kappa > 0$ is a constant and $h_0 \to 0$ as $n \to \infty$. Subsequently, the proposition shows that with high probability, the pre-processing step will remove all grid points in $G_n$ in non-smooth regions of $f$, while maintaining the global optimal solution. This justifies the pre-processing step for $f \in \Sigma^\alpha_\kappa(M)$, because $f$ is smooth on the grid and its close neighborhood after pre-processing.

The proof of Proposition 5 uses the fact that the local mean estimation is large provided that all data points in the local mean estimator are large, regardless of their underlying smoothness. The complete proof of Proposition 5 is deferred to the Appendix.

## V. Proofs of Main Theorems

### A. Proof of Lemma 1

Our proof closely follows the analysis of asymptotic convergence rates for series estimators in [53]. We further work out all constants in the error bounds to arrive at a completely finite-sample result, which is then used to construct finite-sample confidence intervals.

We start with as polynomial interpolation results for all Hölder smooth functions in $B^\infty_{h_t}(x; \mathcal{X})$.

**Lemma 2.** *Suppose $f$ satisfies (6) on $B^\infty_h(x; \mathcal{X})$. Then there exists $\widetilde{f}_x \in \mathcal{P}_k$ such that*

$$\sup_{z\in B^\infty_h(x;\mathcal{X})} \left|f(z) - \widetilde{f}_x(z)\right| \leqslant M d^k h^\alpha. \tag{21}$$

*Proof.* Consider

$$\widetilde{f}_x(z) := f(x) + \sum_{j=1}^{k} \sum_{\alpha_1+\ldots+\alpha_d=j} \frac{\partial^j f(x)}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \prod_{\ell=1}^{d} (z_\ell - x_\ell)^{\alpha_\ell}. \tag{22}$$

By Taylor expansion with Lagrangian remainders, there exists $\xi \in (0, 1)$ such that

$$\left|\widetilde{f}_x(z) - f(z)\right| \leqslant \sum_{\alpha_1+\ldots+\alpha_d=k} \left|f^{(\boldsymbol{\alpha})}(x + \xi(z-x)) - f^{(\boldsymbol{\alpha})}(x)\right| \cdot \prod_{\ell=1}^{d} |z_\ell - x_\ell|^{\alpha_\ell}.$$

Because $f$ satisfies (6) on $B^\infty_h(x; \mathcal{X})$, we have that $|f^{(\boldsymbol{\alpha})}(x + \xi(z-x)) - f^{(\boldsymbol{\alpha})}(x)| \leqslant M \cdot \|z-x\|_\infty^{\alpha-k}$. Also note that $|z_\ell - x_\ell| \leqslant \|z - x\|_\infty \leqslant h$ for all $z \in B^\infty_h(x; \mathcal{X})$. The lemma is thus proved. $\square$

Using (16), the local polynomial estimate $\widehat{f}_h$ can be written as $\widehat{f}_h(z) \equiv \psi_{x,h}(z)^\top \widehat{\theta}_h$, where

$$\widehat{\theta}_h = (\Psi_{t,h}^\top \Psi_{t,h})^{-1} \Psi_{t,h}^\top Y_{t,h}. \tag{23}$$

In addition, because $\widetilde{f}_x \in \mathcal{P}_k$, there exists $\widetilde{\theta} \in \mathbb{R}^D$ such that $\widetilde{f}_x(z) \equiv \psi_{x,h}(z)^\top \widetilde{\theta}$. Denote also that $F_{t,h} := (f(x_{t'}))_{1\leqslant t'\leqslant t, x_{t'}\in B^\infty_h(x)}$, $\Delta_{t,h} := (f(x_{t'}) - \widetilde{f}_x(x_{t'}))$

$_{1\leqslant t'\leqslant t, x_{t'}\in B_h^\infty(x)}$ and $W_{t,h} := (w_{t'})_{1\leqslant t'\leqslant t, x_{t'}\in B_h^\infty(x)}$. (23) can then be re-formulated as

$$\widehat{\theta}_h = (\Psi_{t,h}^\top\Psi_{t,h})^{-1}\Psi_{t,h}^\top\left[\Psi_{t,h}\widetilde{\theta} + \Delta_{t,h} + W_{t,h}\right] \tag{24}$$

$$= \widetilde{\theta} + \left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\left[\frac{1}{m}\Psi_{t,h}^\top(\Delta_{t,h} + W_{t,h})\right]. \tag{25}$$

Because $\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h} \succeq \sigma I_{D\times D}$ and $\sup_{z\in B_h^\infty(x)}\|\psi_{x,h}(z)\|_2 \leqslant b$, we have that

$$\|\widehat{\theta}_h - \widetilde{\theta}\|_2 \leqslant \frac{b}{\sigma}\|\Delta_{t,h}\|_\infty + \left\|\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\frac{1}{m}\Psi_{t,h}^\top W_t\right\|_2. \tag{26}$$

Invoking Lemma 2 we have $\|\Delta_{t,h}\|_\infty \leqslant Md^kh^\alpha$. In addition, because $W_t \sim \mathcal{N}_m(0, I_{m\times n})$, we have that

$$\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\frac{1}{m}\Psi_{t,h}^\top W_t \sim \mathcal{N}_D\left(0, \frac{1}{m}\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\right). \tag{27}$$

Applying concentration inequalities for quadratic forms of Gaussian random vectors (Lemma 10), with probability $1-\delta$ it holds that

$$\left\|\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\frac{1}{m}\Psi_{t,h}^\top W_t\right\|_2 \leqslant \sqrt{\frac{5D\log(1/\delta)}{\sigma m}}. \tag{28}$$

We then have that with probability $1-\delta$ that

$$\|\widehat{\theta}_h - \widetilde{\theta}\|_2 \leqslant \frac{b}{\sigma_h}Md^kh_t^\alpha + \sqrt{\frac{5D\log(1/\delta)}{\sigma m}}. \tag{29}$$

Finally, noting that for all $x' \in B_h^\infty(x;\mathcal{X})$, $\|\psi_{x,h}(x')\|_2 \leqslant b$ by definition, we have that

$$|\widehat{f}_h(x') - f(x')| = |\widehat{f}_h('x) - \widetilde{f}_x(x')|$$
$$= |\psi_{x,h}(x')^\top(\widehat{\theta}_h - \widetilde{\theta})| \leqslant b\|\widehat{\theta}_h - \widetilde{\theta}\|_2,$$

which completes the proof of Lemma 1.

### *B. Proof of Theorem 1*

In this section we prove Theorem 1. We prove the theorem by considering every reference function $f_0 \in \Sigma_\kappa^\alpha(M)\cap\Theta_{\mathbf{C}}$ separately. For simplicity, we assume $\kappa = \infty$ throughout the proof. The $0 < \kappa < \infty$ can be handled by replacing $\mathcal{X}$ with $S_0$ which is the grid after the pre-processing step described in Section IV-C. We also suppress dependency on $d, \alpha, M, \mathbf{C}, \underline{p}_0, \overline{p}_0$ in $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, $\gtrsim$, $\lesssim$ and $\asymp$ notations. We further suppress logarithmic terms of $n$ in $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ notations.

The following lemma is our main lemma, which shows that the active set $S_\tau$ in our proposed algorithm shrinks geometrically before it reaches a certain level. To simplify notations, denote $\widetilde{c}_0 := 10c_0$ and (A2) then hold for all $\epsilon,\delta\in[0,\widetilde{c}_0]$ for all $f_0\in\Theta_{\mathbf{C}}$.

**Lemma 3.** *For $\tau = 1,\ldots,T$ define $\varepsilon_\tau := \max\{\widetilde{c}_0\cdot 2^{-\tau}, C_3[\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n\}$, where $C_3 > 0$ is a constant depending only on $d,\alpha,M,\underline{p}_0,\overline{p}_0$ and $\mathbf{C}$. Denote also $\rho_\tau^* := \max_{x\in S_\tau}\varrho_\tau(x)$. Then for sufficiently large $n$, with probability $1-O(n^{-1})$ the following holds uniformly for all outer iterations $\tau = 1,\ldots,T$:*

$$B_{\rho_\tau^*}^\infty(x;\mathcal{X}) \cap L_f(\varepsilon_\tau) \neq \varnothing. \tag{30}$$

Lemma 3 shows that the level $\varepsilon_\tau$ in $L_f(\varepsilon_\tau)$ that contains $S_{\tau-1}$ shrinks *geometrically*, until the condition $\varepsilon_\tau \geqslant C_3[\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n$ is violated. If the condition is never violated, then at the end of the last epoch $\tau^*$ we have $\varepsilon_{\tau^*} = O(n^{-1})$ because $\tau^* = \log n$. On the other hand, because $S_\tau \subseteq S_{\tau-1}$ always holds, we have $\varepsilon_{\tau^*} \lesssim [\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n$. Combining both cases we have that $\varepsilon_{\tau^*} \lesssim [\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n + n^{-1}$. Theorem 1 is thus proved.

In the rest of this section we prove Lemma 3. We need several technical lemmas and propositions. Except for Proposition 6 that is straightforward, the proofs of the other technical lemmas are deferred to the end of this section.

Denote $x_n^* := \mathrm{argmin}_{x\in G_n} f(x)$ as the point on the grid $G_n$ with the smallest objective value The following proposition shows that with high probability, the confidence intervals constructed in the algorithm are truthful and the successive rejection procedure will never exclude the true optimizer of $f$ on $G_n$.

**Proposition 6.** *Suppose $\delta = 1/n^4|G_n|$. Then with probability $1-O(n^{-1})$ the following hold:*

1) *$f(x') \in [\ell_t(x), u_t(x)]$ for all $1\leqslant t\leqslant n$ and $x\in G_n$, $x'\in B_{h_t(x)}^\infty(x;\mathcal{X})$;*
2) *$x_n^* \in S_\tau$ for all $0\leqslant\tau\leqslant n$.*

*Proof.* The first property is true by applying the union bound over all $t = 1,\ldots,n$ and $x\in G_n$. The second property then follows, because $\ell_t(x_n^*) \leqslant f(x_n^*)$ and $\min_{x\in S_{\tau-1}} u_t(x) \geqslant f(x_n^*)$ for all $\tau$. □

The following lemma shows that every small box centered around a certain sample point $x\in G_n$ contains a sufficient number of sample points whose least eigenvalue can be bounded with high probability under the polynomial mapping $\psi_{x,h}$ defined in Section III-B.

**Lemma 4.** *For any $x\in G_n$, $1\leqslant m\leqslant n$ and $h>0$, let $K_{h,m}^1(x),\ldots,K_{h,m}^n(x)$ be $n$ independent point sets, where each point set consists of $m$ points sampled i.i.d. uniformly at random from $B_h^\infty(x;G_n) = G_n\cap B_h^\infty(x;\mathcal{X})$. With probability $1-O(n^{-1})$ the following holds true uniformly for all $x\in G_n$, $h\in\{j/n^2 : j\in\mathbb{N}, j\leqslant n^2\}$ and $K_{h,m}^\ell(x)$, $\ell\in[n]$ as $n\to\infty$:*

1) *$\sup_{h>0}\sup_{z\in B_h^\infty(x)}\|\psi_{x,h}(z)\|_2 \asymp \Theta(1)$;*
2) *$|B_h^\infty(x;G_n)| \asymp h^d|G_n|$;*
3) *$\sigma_{\min}(K_{h,m}^\ell(x)) \asymp \Theta(1)$ for all $m\geqslant\Omega(\log^2 n)$ and $m\leqslant|G_n|$, where $\sigma_{\min}(K_{h,m}^\ell(x))$ is the least eigenvalue of $\frac{1}{m}\sum_{z\in K_{h,m}^\ell(x)}\psi_{x,h}(z)\psi_{x,h}(z)^\top$.*

**Remark 10.** *It is possible to improve the concentration result in (48) using the strategies adopted in [35] based on sharper Bernstein type concentration inequalities. Such improvements are, however, not important in establishing the main results of this paper.*

The next lemma shows that, the bandwidth $h_t$ selected at the end of each outer iteration $\tau$ is near-optimal, being sandwiched between two quantities determined by the size of the active sample grid $\widetilde{S}_{\tau-1} := S^{\circ}_{\tau-1}(\varrho_{\tau-1})$.

**Lemma 5.** *There exist constants $C_1, C_2 > 0$ depending only on $d, \alpha, M, \underline{p}_0, \overline{p}_0$ and $\mathbf{C}$ such that with probability $1 - O(n^{-1})$, the following holds for every outer iteration $\tau \in \{1, \ldots, T\}$ and all $x \in S_{\tau-1}$:*

$$\begin{aligned} C_1[\widetilde{\nu}_{\tau-1} n_0]^{-1/(2\alpha+d)} - \tau/n &\leqslant \varrho_\tau(x) \\ &\leqslant h_t(x) \leqslant C_2[\widetilde{\nu}_{\tau-1} n_0]^{-1/(2\alpha+d)} \log n + \tau/n, \quad (31) \end{aligned}$$

*where $\widetilde{\nu}_{\tau-1} := |G_n|/|\widetilde{S}_{\tau-1}|$.*

We are now ready to state the proof of Lemma 3, which is based on an inductive argument over the epochts $\tau = 1, \ldots, T$.

*Proof.* We use induction to prove this lemma. For the base case $\tau = 1$, because $\|f - f_0\|_\infty \leqslant \varepsilon_n^{\mathsf{U}}(f_0)$ and $\varepsilon_n^{\mathsf{U}}(f_0) \to 0$ as $n \to \infty$, it suffices to prove that $B^{\infty}_{\rho_1^*}(x; \mathcal{X}) \cap L_{f_0}(\widetilde{c}_0/4) \neq \varnothing$ for all $x \in S_1$ and sufficiently large $n$. Because $\widetilde{S}_0 = S_0 = G_n$, invoking Lemmas 5 and 1 we have that $|\eta_{h_t(x),\delta}(x)| = \widetilde{O}(n^{-\alpha/(2\alpha+d)})$ for all $x \in G_n$ with high probability at the end of the first outer iteration $\tau = 1$. Therefore, for sufficiently large $n$ we conclude that $\sup_{x \in G_n} |\eta_{h_t(x),\delta}(x)| \leqslant c_0/16$ and hence $B^{\infty}_{\rho_1^*}(x; \mathcal{X}) \cap L_{f_0}(\widetilde{c}_0/4) \neq \varnothing$ for all $x \in S_1$.

We now prove the lemma for $\tau \geqslant 2$, assuming it holds for $\tau - 1$. We also assume that $n$ (and hence $n_0$) is sufficiently large, such that the maximum CI length $\max_{x \in G} |\eta_{h_t(x),\delta}(x)|$ after the first outer iteration $\tau = 1$ is smaller than $c_0/2$.

Because $\|f - f_0\|_\infty \leqslant \varepsilon_n^{\mathsf{U}}(f_0)$ and $\varepsilon_{\tau-1} \geqslant C_3 \varepsilon_n^{\mathsf{U}}(f_0) \log^2 n$, for appropriately chosen constant $C_3$ that is not too small, we have that $\|f - f_0\|_\infty \leqslant \varepsilon_{\tau-1}$. By the inductive hypothesis we have

$$\forall x \in S_{\tau-1}, \quad B^{\infty}_{\rho^*_{\tau-1}}(x; \mathcal{X}) \cap L_f(\varepsilon_{\tau-1}) \neq \varnothing;$$

Equivalently,

$$\begin{aligned} S_{\tau-1} \subseteq L^{\circ}_f(\varepsilon_{\tau-1}, \rho^*_{\tau-1}) &\subseteq L^{\circ}_{f_0}(\varepsilon_{\tau-1} + \|f - f_0\|_\infty, \rho^*_{\tau-1}) \\ &\subseteq L^{\circ}_{f_0}(2\varepsilon_{\tau-1}, \rho^*_{\tau-1}). \quad (32) \end{aligned}$$

Subsequently,

$$\widetilde{S}_{\tau-1} = S^{\circ}_{\tau-1} \subseteq L^{\circ}_{f_0}(2\varepsilon_{\tau-1}, 2\rho^*_{\tau-1}). \quad (33)$$

Let $\bigcup_{x \in H_n} B^2_{2\rho^*_{\tau-1}}(x)$ be the smallest covering set of $L_{f_0}(2\varepsilon_{\tau-1})$, meaning that $L_{f_0}(2\varepsilon_{\tau-1}) \subseteq \bigcup_{x \in H_n} B^2_{2\rho^*_{\tau-1}}(x)$, where $B^2_{2\rho^*_{\tau-1}}(x) = \{z \in \mathcal{X} : \|z - x\|_2 \leqslant 2\rho^*_{\tau-1}\}$ is the $\ell_2$ ball of radius $2\rho^*_{\tau-1}$ centered at $x$. By (A2), we know that $|H_n| \lesssim 1 + [\rho^*_{\tau-1}]^{-d} \mu_{f_0}(2\varepsilon_{\tau-1})$. In addition, the enlarged level-set satisfies $L^{\circ}_{f_0}(2\varepsilon_{\tau-1}, 2\rho^*_{\tau-1}) \subseteq \bigcup_{x \in H_n} B^{\infty}_{4\rho^*_{\tau-1}}(x)$. Subsequently,

$$\mu^{\circ}_{f_0}(2\varepsilon_{\tau-1}, \rho^*_{\tau-1}) \lesssim |H_n| \cdot [\rho^*_{\tau-1}]^d \lesssim \mu_{f_0}(2\varepsilon_{\tau-1}) + [\rho^*_{\tau-1}]^d. \quad (34)$$

By Lemma 5, the monotonicity of $|\widetilde{S}_{\tau-1}|$ and the fact that $\underline{p}_0 \leqslant p_X(z) \leqslant \overline{p}_0$ for all $z \in \mathcal{X}$, we have

$$\begin{aligned} \rho^*_{\tau-1} &\lesssim [\mu^{\circ}_f(\varepsilon_{\tau-1}, \rho^*_{\tau-1})]^{1/(2\alpha+d)} n_0^{-1/(2\alpha+d)} \log n \quad (35) \\ &\leqslant [\mu^{\circ}_{f_0}(2\varepsilon_{\tau-1}, \rho^*_{\tau-1})]^{1/(2\alpha+d)} n_0^{-1/(2\alpha+d)} \log n \quad (36) \\ &\lesssim \left(\mu_{f_0}(2\varepsilon_{\tau-1}) + [\rho^*_{\tau-1}]^d\right)^{1/(2\alpha+d)} n_0^{-1/(2\alpha+d)} \log n. \quad (37) \end{aligned}$$

Re-arranging terms on both sides of (37) we have

$$\rho^*_{\tau-1} \lesssim \max\left\{[\mu_{f_0}(2\varepsilon_{\tau-1})]^{\frac{1}{2\alpha+d}} n_0^{-\frac{1}{2\alpha+d}} \log n,\ n_0^{-\frac{1}{2\alpha}} \log n\right\}. \quad (38)$$

On the other hand, according to the selection procedure of the bandwidth $h_t(x)$, we have that $\eta_{h_t(x),\delta}(x) \lesssim \mathfrak{b}_{h_t(x),\delta}(x)$. Invoking Lemma 5 we have for all $x \in S_{\tau-1}$ that

$$\begin{aligned} \eta_{h_t(x),\delta}(x) \lesssim \mathfrak{b}_{h_t(x),\delta}(x) &\lesssim [h_t(x)]^\alpha \quad (39) \\ &\lesssim [\widetilde{\nu}_{\tau-1} n_0]^{-\alpha/(2\alpha+d)} \log n \quad (40) \\ &\lesssim [\widetilde{\nu}_{\tau-2} n_0]^{-\alpha/(2\alpha+d)} \log n \quad (41) \\ &\lesssim [\rho^*_{\tau-1}]^\alpha \log n. \quad (42) \end{aligned}$$

Here (40) holds by invoking the upper bound on $h_t(x)$ in Lemma 5, (41) holds because $\widetilde{\nu}_{\tau-1} \geqslant \widetilde{\nu}_{\tau-2}$, and (42) holds by again invoking the lower bound on $\varrho_{\tau-1}(x)$ in Lemma 5. Combining Eqs. (38,42) we have

$$\begin{aligned} &\max_{x \in S_{\tau-1}} \eta_{h_t(x),\delta}(x) \quad (43) \\ &\lesssim \max\left\{[\mu_{f_0}(2\varepsilon_{\tau-1})]^{\frac{\alpha}{2\alpha+d}} n_0^{-\frac{\alpha}{2\alpha+d}} \log^2 n,\ n_0^{-\frac{1}{2}} \log n\right\}. \quad (44) \end{aligned}$$

Recall that $n_0 = n/\log n$ and $\varepsilon_n^{\mathsf{U}}(f_0) \leqslant \varepsilon_{\tau-1}$, provided that $C_3$ is not too small. By definition, every $\varepsilon \geqslant \varepsilon_n^{\mathsf{U}}(f_0)$ satisfies $\varepsilon^{-(2+d/\alpha)} \mu_{f_0}(\varepsilon) \leqslant n/\log^\omega n$ for some large constant $\omega > 5 + d/\alpha$. Subsequently,

$$\begin{aligned} &[\mu_{f_0}(2\varepsilon_{\tau-1})]^{\frac{\alpha}{2\alpha+d}} n_0^{-\frac{\alpha}{2\alpha+d}} \log^2 \\ &\lesssim 2\varepsilon_{\tau-1} n^{\frac{\alpha}{2\alpha+d}} \log^{-\frac{\omega\alpha}{2\alpha+d}} n \cdot n_0^{-\frac{\alpha}{2\alpha+d}} \log^2 n \quad (45) \\ &\lesssim \varepsilon_{\tau-1}/[\log n]^{\frac{(\omega-5-d/\alpha)\alpha}{2\alpha+d}}. \quad (46) \end{aligned}$$

Because $\omega > 5 + d/\alpha$, the right-hand side of (46) is asymptotically dominated [6] by $\varepsilon_{\tau-1}$. In addition, $n_0^{-1/2} \log n$ is also asymptotically dominated by $\varepsilon_{\tau-1}$ because $\varepsilon_{\tau-1} \geqslant C_3 n^{-1/2} \log^\omega n$. Therefore, for sufficiently large $n$ we have

$$\max_{x \in S_{\tau-1}} \eta_{h_t(x),\delta}(x) \leqslant \varepsilon_{\tau-1}/4. \quad (47)$$

Lemma 3 is thus proved. □

[6] We say $\{a_n\}$ is asymptotically dominated by $\{b_n\}$ if $\lim_{n\to\infty} |a_n|/|b_n| = 0$.

*1) Proof of Lemma 4:*

*Proof.* We first show that the first property holds almost surely. Recall the definition of $\psi_{x,h}$, we have that $1 \leqslant \|\psi_{x,h}(z)\|_2 \leqslant D \cdot [\max_{1\leqslant j\leqslant d} h^{-1}|z_j - x_j|]^k$. Because $\|z-x\|_\infty \leqslant h$ for all $z \in B_h^\infty(x)$, $\sup_{z\in B_h^\infty(x)} \|\psi_{x,h}(z)\|_2 \lesssim O(1)$ for all $h > 0$. Thus, $\sup_{h>0}\sup_{z\in B_h^\infty(x)} \|\psi_{x,h}(z)\|_2 \asymp \Theta(1)$ for all $x \in G_n$.

For the second property, by Hoeffding's inequality (Lemma 9) and the union bound, with probability $1-O(n^{-1})$ we have that

$$\max_{x,h}\left|\frac{|B_h^\infty(x;G_n)|}{|G_n|} - P_X(z\in B_h^\infty(x))\right| \lesssim \sqrt{\frac{\log n}{|G_n|}}. \quad (48)$$

In addition, note that $P_X(z \in B_h^\infty(x;\mathcal{X})) \geqslant \underline{p}_0\lambda(B_h^\infty(x;\mathcal{X})) \gtrsim h^d$ and $P_X(z \in B_h^\infty(x;\mathcal{X})) \leqslant \overline{p}_0\lambda(B_h^\infty(x;\mathcal{X})) \lesssim h^d$, where $\lambda(\cdot)$ denotes the Lebesgue measure on $\mathcal{X}$. Subsequently, $|B_h^\infty(x;G_n)|$ is lower bounded by $\Omega(h^d|G_n| - \sqrt{|G_n|\log n})$ and upper bounded by $O(h^d|G_n| + \sqrt{|G_n|\log n})$. The second property is then proved by noting that $h_d \gtrsim n^{-d}$ and $|G_n| \gtrsim n^{3d/\min(\alpha,1)}$.

We next prove the third property. Because $\underline{p}_0 \leqslant p_X(z) \in \overline{p}_0$ for all $z \in \mathcal{X}$, we have that

$$\begin{aligned}
&\underline{p}_0 \int_{B_h^\infty(x;\mathcal{X})} \psi_{x,h}(z)\psi_{x,h}(z)^\top \mathrm{d}U_{x,h}(z)\\
&\preceq \mathbb{E}\left[\frac{1}{m}\sum_{z\in K_{h,m}^\ell}\psi_{x,h}(z)\psi_{x,h}(z)^\top\right] \quad (49)\\
&\preceq \overline{p}_0 \int_{B_h^\infty(x;\mathcal{X})} \psi_{x,h}(z)\psi_{x,h}(z)^\top \mathrm{d}U_{x,h}(z), \quad (50)
\end{aligned}$$

where $U_{x,h}$ is the uniform distribution on $B_h^\infty(x;\mathcal{X})$. Note also that

$$\begin{aligned}
&\int_{\mathcal{X}} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z)\\
&\preceq \int_{B_h^\infty(x;\mathcal{X})} \psi_{x,h}(z)\psi_{x,h}(z)^\top \mathrm{d}U_{x,h}(z) \quad (51)\\
&\preceq 2^d \int_{\mathcal{X}} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z) \quad (52)
\end{aligned}$$

where $U$ is the uniform distribution on $\mathcal{X} = [0,1]^d$. The following proposition upper and lower bounds the eigenvalues of $\int_{\mathcal{X}} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z)$, which is proved in the appendix.

**Proposition 7.** *There exist constants* $0 < \psi_0 \leqslant \Psi_0 < \infty$ *depending only on* $d, D$ *such that*

$$\psi_0\, I_{D\times D} \preceq \int_{\mathcal{X}} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z) \preceq \Psi_0 I_{D\times D}. \quad (53)$$

Using Proposition 7 and Eqs. (51,52), we conclude that

$$\Omega(1)\cdot I_{D\times D} \preceq \mathbb{E}\left[\frac{1}{m}\sum_{z\in K_{h,m}^\ell}\psi_{x,h}(z)\psi_{x,h}(z)^\top\right] \preceq O(1)\cdot I_{D\times D}. \quad (54)$$

Applying matrix Chernoff bound (Lemma 11) and the union bound, we have that with probability $1-O(n^{-1})$,

$$\begin{aligned}
\max_{x,h,m,\ell}\Bigg\| &\frac{1}{m}\sum_{z\in K_{h,m}^\ell(x)}\psi_{x,h}(z)\psi_{x,h}(z)^\top\\
&-\mathbb{E}\left[\psi_{x,h}(z)\psi_{x,h}(z)^\top | z\in B_h^\infty(x)\right]\Bigg\|_{\mathrm{op}} \lesssim \sqrt{\frac{\log n}{m}}.
\end{aligned}$$

Combining Eqs. (54,55) and applying Weyl's inequality (Lemma 12) we have

$$\begin{aligned}
\Omega(1) - O(\sqrt{\log n/m}) &\lesssim \sigma_{\min}(K_{h,m}^\ell(x))\\
&\lesssim O(1) - O(\sqrt{\log n/m}). \quad (55)
\end{aligned}$$

The third property is therefore proved. □

*2) Proof of Lemma 5: Proof.* We use induction to prove this lemma. For the base case of $\tau = 1$, we have $\widetilde{S}_0 = S_0 = G_n$ and therefore $\widetilde{\nu}_{\tau-1} = 1$. Furthermore, applying Lemma 4 we have that for all $h = j/n^2$,

$$\mathfrak{b}_{h,\delta}(x) \asymp h^\alpha, \quad \mathfrak{s}_{h,\delta}(x) \asymp \sqrt{\frac{\log n}{h^d n_0}}. \quad (56)$$

Thus, for $h$ selected according to (18) as the largest bandwidth of the form $j/n^2$, $j \in \mathbb{N}$ such that $\mathfrak{b}_{h,\delta}(x) \leqslant \mathfrak{s}_{h,\delta}(x)$, both $\mathfrak{b}_{h,\delta}(x), \mathfrak{s}_{h,\delta}(x)$ are on the order of $n_0^{-1/(2\alpha+d)}$ up to logarithmic terms of $n$, and therefore one can pick appropriate constants $C_1, C_2 > 0$ such that $C_1 n_0^{-1/(2\alpha+d)} \leqslant \varrho_1(x) \leqslant C_2 n_0^{-1/(2\alpha+d)}\log n$ holds for all $x \in G_n$.

We next prove the lemma for $\tau > 1$, assuming it holds for $\tau - 1$. We first establish the lower bound part. Define $\rho_{\tau-1}^* := \min_{z\in S_{\tau-1}} \varrho_{\tau-1}(z)$. By inductive hypothesis, $\rho_{\tau-1}^* \geqslant C_1[\widetilde{\nu}_{\tau-2}n_0]^{-1/(2\alpha+d)} - (\tau-1)/n$. Note also that $\widetilde{\nu}_{\tau-1} \geqslant \widetilde{\nu}_{\tau-2}$ because $\widetilde{S}_{\tau-1} \subseteq \widetilde{S}_{\tau-2}$, which holds because $S_{\tau-1} \subseteq S_{\tau-2}$ and $\varrho_{\tau-1}(z) \leqslant \varrho_{\tau-2}(z)$ for all $z$. Let $h_t^*$ be the smallest number of the form $j_t^*/n^2$, $j_t^* \in [n^2]$ such that $h_t^* \geqslant C_1[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} - \tau/n$. We then have $h_t^* \leqslant \rho_{\tau-1}^*$ and therefore query points in epoch $\tau$ are uniformly distributed in $B_{h_t^*}^\infty(x;G_n)$. Subsequently, applying Lemma 4 we have with probability $1-O(n^{-1})$ that

$$\mathfrak{b}_{h_t^*,\delta}(x) \leqslant C'[h_t^*]^\alpha, \quad \mathfrak{s}_{h_t^*,\delta}(x) \geqslant C''\sqrt{\frac{\log n}{[h_t^*]^d\widetilde{\nu}_{\tau-1}n}}, \quad (57)$$

where $C', C'' > 0$ are constants that depend on $d, \alpha, M, \underline{p}_0, \overline{p}_0$ and $\mathbf{C}$, but *not* $C_1$, $C_2$, $\tau$ or $h_t^*$. By choosing $C_1$ appropriately (depending on $C'$ and $C''$) we can make $\mathfrak{b}_{h_t^*,\delta}(x) \leqslant \mathfrak{s}_{h_t^*,\delta}(x)$ holds for all $x \in S_{\tau-1}$, thus establishing $\varrho_\tau(x) \geqslant \min\{\varrho_{\tau-1}(x), h_t^*\} \geqslant C_1[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} - \tau/n$.

We next prove the upper bound part. For any $h_t = j_t/n^2$ where $j_t \in [n^2]$, invoking Lemma 4 we have that

$$\mathfrak{b}_{h,\delta}(x) \geqslant \widetilde{C}'h^\alpha, \ \mathfrak{s}_{h,\delta}(x) \leqslant \widetilde{C}''\sqrt{\frac{\log n}{\min\{h, \rho_{\tau-1}^*\}^d\cdot\widetilde{\nu}_{\tau-1}n_0}}, \quad (58)$$

where $\widetilde{C}'$ and $\widetilde{C}''$ are again constants depending on $d, \alpha, M, \underline{p}_0, \overline{p}_0$ and $\mathbf{C}$, but *not* $C_1, C_2$. Note also that $\rho^*_{\tau-1} \geqslant C_1[\widetilde{\nu}_{\tau-2}n_0]^{-1/(2\alpha+d)} - (\tau-1)/n \geqslant C_1[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} - \tau/n$, because $\widetilde{\nu}_{\tau-1} \geqslant \widetilde{\nu}_{\tau-2}$. By selecting constant $C_2 > 0$ carefully (depending on $\widetilde{C}', \widetilde{C}''$ and $C_1$), we can ensure $\mathfrak{b}_{h,\delta}(x) > \mathfrak{s}_{h,\delta}(x)$ for all $h \geqslant C_2[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} + \tau/n$. Therefore, $\varrho_\tau(x) \leqslant h_t(x) \leqslant C_2[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} + \tau/n$. $\square$

### C. Proof of Theorem 2

In this section we prove the main negative result in Theorem 2. To simplify presentation, we suppress dependency on $\alpha, d, c_0$ and $C_0$ in $\lesssim, \gtrsim, \asymp, O(\cdot)$ and $\Omega(\cdot)$ notations. However, we do *not* suppress dependency on $\underline{C}_R$ or $M$ in any of the above notations.

Let $\varphi_0 : [-2,2]^d \to \mathbb{R}^*$ be a non-negative function defined on $\mathcal{X}$ such that $\varphi_0 \in \Sigma_\kappa^{\lceil\alpha\rceil}(1)$ with $\kappa = \infty$, $\sup_{x\in\mathcal{X}} \varphi_0(x) = \Omega(1)$ and $\varphi_0(z) = 0$ for all $\|z\|_2 \geqslant 1$. Here $\lceil\alpha\rceil$ denotes the smallest integer that upper bounds $\alpha$. Such functions exist and are the cornerstones of the construction of information-theoretic lower bounds in nonparametric estimation problems [50]. One typical example is the "smoothstep" function (see for example [54])

$$S_N(x) := \frac{1}{Z} x^{N+1} \sum_{n=0}^{N} \binom{N+n}{n}\binom{2N+1}{N-n}(-x)^n, \qquad N = 0, 1, 2, \ldots,$$

where $Z > 0$ is a scaling parameter. The smoothstep function $S_N$ is defined on $[0,1]$ and satisfies the Hölder condition in (6) of order $\alpha = N$ on $[0,1]$. It can be easily extended to $\widetilde{S}_{N,d} : [-2,2]^d \to \mathbb{R}$ by considering $\widetilde{S}_{N,d}(x) := 1/Z - S_N(a\|x\|_1)$ where $\|x\|_1 = |x_1| + \ldots + |x_d|$ and $a = 1/(2d)$. It is easy to verify that, with $Z$ chosen appropriately, $\widetilde{S}_{N,d} \in \Sigma_\infty^N(1)$, $\sup_{x\in\mathcal{X}} \widetilde{S}_{N,d}(x) = 1/Z = \Omega(1)$ and $\widetilde{S}_{N,d}(z) = 0$ for all $\|z\|_2 \geqslant 1$, where $M > 0$ is a constant.

For any $x \in \mathcal{X}$ and $h > 0$, define $\varphi_{x,h} : \mathcal{X} \to \mathbb{R}^*$ as

$$\varphi_{x,h}(z) := \mathbb{I}[z \in B_h^\infty(x)] \cdot \frac{Mh^\alpha}{2}\varphi_0\left(\frac{z-x}{h}\right). \tag{59}$$

It is easy to verify that $\varphi_{x,h} \in \Sigma_\infty^\alpha(M/2)$, and furthermore $\sup_{z\in\mathcal{X}} \varphi_{x,h}(z) \asymp Mh^\alpha$ and $\varphi_{x,h}(z) = 0$ for all $z \notin B_h^\infty(x)$.

Let $L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0))$ be the level-set of $f_0$ at $\varepsilon_n^{\mathsf{L}}(f_0)$. Let $H_n \subseteq L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0))$ be the largest *packing* set such that $B_h^\infty(x)$ are disjoint for all $x \in H_n$, and $\bigcup_{x\in H_n} B_h^\infty(x) \subseteq L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0))$. By (A2') and the definition of $\varepsilon_n^{\mathsf{L}}(f_0)$, we have that

$$\begin{aligned} |H_n| &\geqslant M(L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0)), 2\sqrt{d}h) \\ &\gtrsim \mu_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0)) \cdot h^{-d} \geqslant [\varepsilon_n^{\mathsf{L}}(f_0)]^{2+d/\alpha} \cdot nh^{-d}. \end{aligned} \tag{60}$$

For any $x \in H_n$, construct $f_x : \mathcal{X} \to \mathbb{R}$ as

$$f_x(z) := f_0(z) - \varphi_{x,h}(z). \tag{61}$$

Let $\mathcal{F}_n := \{f_x : x \in H_n\}$ be the class of functions indexed by $x \in H_n$. Let also $h \asymp (\varepsilon_n^{\mathsf{L}}(f_0)/M)^{1/\alpha}$ such that $\|\varphi_{x,h}\|_\infty = 2\varepsilon_n^{\mathsf{L}}(f_0)$. We then have that $\|f_x - f_0\|_\infty \leqslant 2\varepsilon_n^{\mathsf{L}}(f_0)$ and $f_x \in \Sigma_\infty^\alpha(M)$, because $f_0, \varphi_{x,h} \in \Sigma_\infty^\alpha(M/2)$.

The next lemma shows that, with $n$ adaptive queries to the noisy zeroth-order oracle $y_t = f(x_t) + w_t$, it is information theoretically not possible to identify a certain $f_x$ in $\mathcal{F}_n$ with high probability.

**Lemma 6.** *Suppose $|\mathcal{F}_n| \geqslant 2$. Let $\mathcal{A}_n = (\chi_1, \ldots, \chi_n, \phi_n)$ be an active optimization algorithm operating with a sample budget n, which consists of* samplers $\chi_\ell : \{(x_i, y_i)\}_{i=1}^{\ell-1} \mapsto x_\ell$ *and an* estimator $\phi_n : \{(x_i, y_i)\}_{i=1}^n \mapsto \widehat{f}_x \in \mathcal{F}_n$, *both can be deterministic or randomized functions. Then*

$$\inf_{\mathcal{A}_n} \sup_{f_x\in\mathcal{F}_n} \Pr_{f_x}\left[\widehat{f}_x \neq f_x\right] \geqslant \frac{1}{2} - \sqrt{\frac{n \cdot \sup_{f_x\in\mathcal{F}_n}\|f_x - f_0\|_\infty^2}{2|\mathcal{F}_n|}}. \tag{62}$$

**Lemma 7.** *There exists constant $M > 0$ depending on $\alpha, d, c_0, C_0$ such that the right-hand side of (62) is lower bounded by* $1/3$.

Lemmas 6 and 7 are proved at the end of this section. Combining both lemmas and noting that for any distinct $f_x, f_{x'} \in \mathcal{F}_n$ and $z \in \mathcal{X}$, $\max\{\mathfrak{L}(z; f_x), \mathfrak{L}(z; f_{x'})\} \geqslant \varepsilon_n^{\mathsf{L}}(f_0)$, we proved the minimax lower bound formulated in Theorem 2.

*1) Proof of Lemma 6:* Our proof is inspired by the negative result of multi-arm bandit pure exploration problems established in [51].

*Proof.* For any $x \in H_n$, define

$$n_x := \mathbb{E}_{f_0}\left[\sum_{i=1}^n \mathbb{I}[x \in B_h^\infty(x)]\right]. \tag{63}$$

Because $B_h^\infty(x)$ are disjoint for $x \in H_n$, we have $\sum_{x\in H_n} n_x \leqslant n$. Also define, for every $x \in H_n$,

$$\wp_x := \Pr_{f_0}\left[\widehat{f}_x = f_x\right]. \tag{64}$$

Because $\sum_{x\in H_n} \wp_x = 1$, by pigeonhole principle there is at most one $x \in H_n$ such that $\wp_x > 1/2$. Let $x_1, x_2 \in H_n$ be the points that have the smallest and second smallest $n_x$. Then there exists $x \in \{x_1, x_2\}$ such that $\wp_x \leqslant 1/2$ and $n_x \leqslant 2n/|\mathcal{F}_n|$. By Le Cam's and Pinsker's inequality (see, for example, [4]) we have that

$$\begin{aligned} \Pr_{f_x}\left[\widehat{f}_x = f_x\right] &\leqslant \Pr_{f_0}\left[\widehat{f}_x = f_x\right] + d_{\mathrm{TV}}(P_{f_0}^{\mathcal{A}_n} \| P_{f_x}^{\mathcal{A}_n}) && (65) \\ &\leqslant \Pr_{f_0}\left[\widehat{f}_x = f_x\right] + \sqrt{\frac{1}{2}\mathrm{KL}(P_{f_0}^{\mathcal{A}_n} \| P_{f_x}^{\mathcal{A}_n})} && (66) \\ &= \wp_x + \sqrt{\frac{1}{2}\mathrm{KL}(P_{f_0}^{\mathcal{A}_n} \| P_{f_x}^{\mathcal{A}_n})} && (67) \\ &\leqslant \frac{1}{2} + \sqrt{\frac{1}{2}\mathrm{KL}(P_{f_0}^{\mathcal{A}_n} \| P_{f_x}^{\mathcal{A}_n})}. && (68) \end{aligned}$$

It remains to upper bound KL divergence of the active queries made by $\mathcal{A}_n$. Using the standard lower bound analysis for active learning algorithms [50], [55] and the fact that

$f_x \equiv f_0$ on $\mathcal{X}\backslash B_h^\infty(x)$, we have

$$\mathrm{KL}(P_{f_0}^{\mathcal{A}_n}\|P_{f_x}^{\mathcal{A}_n}) = \mathbb{E}_{f_0,\mathcal{A}_n}\left[\log\frac{P_{f_0,\mathcal{A}_n}(x_{1:n},y_{1:n})}{P_{f_x,\mathcal{A}_n}(x_{1:n},y_{1:n})}\right] \tag{69}$$

$$= \mathbb{E}_{f_0,\mathcal{A}_n}\left[\log\frac{\prod_{i=1}^n P_{f_0}(y_i|x_i)P_{\mathcal{A}_n}(x_i|x_{1:(i-1)},y_{1:(i-1)})}{\prod_{i=1}^n P_{f_x}(y_i|x_i)P_{\mathcal{A}_n}(x_i|x_{1:(i-1)},y_{1:(i-1)})}\right] \tag{70}$$

$$= \mathbb{E}_{f_0,\mathcal{A}_n}\left[\log\frac{\prod_{i=1}^n P_{f_0}(y_i|x_i)}{\prod_{i=1}^n P_{f_x}(y_i|x_i)}\right] \tag{71}$$

$$= \mathbb{E}_{f_0,\mathcal{A}_n}\left[\sum_{x_i\in B_h^\infty(x)}\log\frac{P_{f_0}(y_i|x_i)}{P_{f_x}(y_i|x_i)}\right] \tag{72}$$

$$\leqslant n_x\cdot\sup_{z\in B_h^\infty(x;\mathcal{X})}\mathrm{KL}(P_{f_0}(\cdot|z)\|P_{f_x}(\cdot|z)) \tag{73}$$

$$\leqslant n_x\cdot\|f_0-f_x\|_\infty^2. \tag{74}$$

Therefore,

$$\Pr_{f_x}\left[\widehat{f}_x = f_x\right] \leqslant \frac{1}{2}+\sqrt{\frac{1}{4}n_x\varepsilon_n^2} \leqslant \frac{1}{2}+\sqrt{\frac{n\|f_x-f_0\|_\infty^2}{2|\mathcal{F}_n|}}. \tag{75}$$

□

*2) Proof of Lemma 7:*

*Proof.* By construction, $n\sup_{f_x\in\mathcal{F}_x}\|f_x-f_0\|_\infty^2 \lesssim M^2\ nh^{2\alpha}$ and $|\mathcal{F}_n| = |H_n| \gtrsim [\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)]^{2+d/\alpha}nh^{-d}$. Note also that $h \asymp (\varepsilon/M)^{1/\alpha} \asymp (\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)/M)^{1/\alpha}$ because $\|f_x-f_0\|_\infty = \varepsilon = \underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)$. Subsequently,

$$\frac{n\sup_{f_x\in\mathcal{F}_x}\|f_x-f_0\|_\infty^2}{2|\mathcal{F}_n|} \lesssim \frac{n[\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)]^2}{n[\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)]^2\cdot M^{d/\alpha}} = M^{-d/\alpha}. \tag{76}$$

By choosing the constant $M>0$ to be sufficiently large, the right-hand side of the above inequality is upper bounded by $1/36$. The lemma is thus proved. □

### *D. Proof of Theorem 3*

The proof of Theorem 3 is similar to the proof of Theorem 2, but is much more standard by invoking the *Fano's inequality* [4]. In particular, adapting the Fano's inequality on any finite function class $\mathcal{F}_n$ constructed we have the following lemma:

**Lemma 8** (Fano's inequality)**.** *Suppose $|\mathcal{F}_n| \geqslant 2$, and $\{(x_i,y_i)\}_{i=1}^n$ are i.i.d. random variables. Then*

$$\inf_{\widehat{f}_x}\sup_{f_x\in\mathcal{F}_n}\Pr_{f_x}\left[\widehat{f}_x\neq f_x\right] \geqslant 1-\frac{\log 2+n\cdot\sup_{f_x,f_{x'}\in\mathcal{F}_n}\mathrm{KL}(P_{f_x}\|P_{f_{x'}})}{\log|\mathcal{F}_n|}, \tag{77}$$

*where $P_{f_x}$ denotes the distribution of $(x,y)$ under the law of $f_x$.*

Let $\mathcal{F}_n$ be the function class constructed in the previous proof of Theorem 2, corresponding to the largest packing set $H_n$ of $L_{f_0}(\widetilde{\varepsilon}_n^{\mathsf{L}})$ such that $B_h^\infty(x)$ for all $x\in H_n$ are disjoint, where $h\asymp(\widetilde{\varepsilon}_n^{\mathsf{L}}/M)^{1/\alpha}$ such that $\|\varphi_{x,h}\|_\infty = 2\widetilde{\varepsilon}_n^{\mathsf{L}}$ for all $x\in H_n$. Because $f_0$ satisfies (A2'), we have that $|\mathcal{F}_n| = |H_n| \gtrsim \mu_{f_0}(\widetilde{\varepsilon}_n^{\mathsf{L}})h^{-d}$. Under the condition that $\varepsilon_n^{\mathsf{U}}(f_0)\leqslant\widetilde{\varepsilon}_n^{\mathsf{L}}$, it holds that $\mu_{f_0}(\widetilde{\varepsilon}_n^{\mathsf{L}}) \geqslant [\widetilde{\varepsilon}_n^{\mathsf{L}}]^{2+d/\alpha}n$. Therefore,

$$|\mathcal{F}_n| \gtrsim [\widetilde{\varepsilon}_n^{\mathsf{L}}]^{2+d/\alpha}\cdot nh^{-d} \gtrsim [\widetilde{\varepsilon}_n^{\mathsf{L}}]^2\cdot nM^{d/\alpha}. \tag{78}$$

Because $\log(n/\widetilde{\varepsilon}_n^{\mathsf{L}}) \gtrsim \log n$ and $M>0$ is a constant, we have that $\log|\mathcal{F}_n| \geqslant c\log n$ for all $n\geqslant N$, where $c>0$ is a constant depending only on $\alpha, d$ and $N\in\mathbb{N}$ is a constand depending on $M$.

Let $U$ be the uniform distribution on $\mathcal{X}$. Because $x\sim U$ and $f_x\equiv f_{x'}$ on $\mathcal{X}\backslash B_h^\infty(x)$, we have that

$$\mathrm{KL}(P_{f_x}\|P_{f_{x'}}) = \frac{1}{2}\int_{\mathcal{X}}|f_x(z)-f_{x'}(z)|^2\mathrm{d}U(z) \tag{79}$$

$$\leqslant \frac{1}{2}\Pr_U\left[z\in B_h^\infty(x)\right]\cdot\|f_x-f_{x'}\|_\infty^2 \tag{80}$$

$$\leqslant \frac{1}{2}\lambda(B_h^\infty(x))\cdot[\varepsilon_n^{\mathsf{L}}]^2 \tag{81}$$

$$\lesssim h^d[\widetilde{\varepsilon}_n^{\mathsf{L}}]^2 \lesssim [\widetilde{\varepsilon}_n^{\mathsf{L}}]^{2+d/\alpha}/M^{d/\alpha}. \tag{82}$$

By choosing $M$ to be sufficiently large, the right-hand side of (77) can be lower bounded by an absolute constant. The theorem is then proved following the same argument as in the proof of Theorem 2.

## APPENDIX A
## SOME CONCENTRATION INEQUALITIES

In this section, to ease readability of our paper, we provide some concentration inequalities and other standard results that we use extensively.

**Lemma 9** ( [56])**.** *Suppose $X_1,\ldots,X_n$ are i.i.d. random variables such that $a\leqslant X_i\leqslant b$ almost surely. Then for any $t>0$,*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i-\mathbb{E}X\right|>t\right]\leqslant 2\exp\left\{-\frac{nt^2}{2(b-a)^2}\right\}.$$

**Lemma 10** ( [57])**.** *Suppose $x\sim\mathcal{N}_d(0,I_{d\times d})$ and let $A$ be a $d\times d$ positive semi-definite matrix. Then for all $t>0$,*

$$\Pr\left[x^\top Ax>\mathrm{tr}(A)+2\sqrt{\mathrm{tr}(A^2)t}+2\|A\|_{\mathrm{op}}t\right]\leqslant e^{-t}.$$

**Lemma 11** ( [58], simplified)**.** *Suppose $A_1,\ldots,A_n$ are i.i.d. positive semidefinite random matrices of dimension $d$ and $\|A_i\|_{\mathrm{op}}\leqslant R$ almost surely. Then for any $t>0$,*

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n A_i-\mathbb{E}A\right\|_{\mathrm{op}}>t\right]\leqslant 2\exp\left\{-\frac{nt^2}{8R^2}\right\}.$$

**Lemma 12** (Weyl's inequality)**.** *Let $A$ and $A+E$ be $d\times d$ matrices with $\sigma_1,\ldots,\sigma_d$ and $\sigma_1',\ldots,\sigma_d'$ be their singular values, sorted in descending order. Then* $\max_{1\leqslant i\leqslant d}|\sigma_i-\sigma_i'|\leqslant\|E\|_{\mathrm{op}}$.

## APPENDIX B
## ADDITIONAL PROOFS

*Proof of Proposition 1.* Consider arbitrary $x^* \in \mathcal{X}$ such that $f(x^*) = \inf_{x\in\mathcal{X}} f(x)$. Then we have that $\mathfrak{L}(\widehat{x}_n; f) = f(\widehat{x}_n) - f(x^*) \leqslant [\widehat{f}_n(\widehat{x}_n) + \|\widehat{f}_n - f\|_\infty] - [\widehat{f}_n(x^*) - \|\widehat{f}_n - f\|_\infty] \leqslant 2\|\widehat{f}_n - f\|_\infty$, where the last inequality holds because $\widehat{f}_n(\widehat{x}_n) \leqslant \widehat{f}_n(x^*)$ by optimality of $\widehat{x}_n$. □

*Proof of Example 2.* Because $f_0 \in \Sigma_\kappa^2(M)$ is strongly convex, there exists $\sigma > 0$ such that $\nabla^2 f_0(x) \succeq \sigma I$ for all $x \in \mathcal{X}_{f_0,\kappa}$, where $\mathcal{X}_{f_0,\kappa} := L_{f_0}(\kappa)$ is the $\kappa$-level-set of $f_0$. Let $x^* = \arg\min_{x\in\mathcal{X}} f_0(x)$, which is unique because $f_0$ is strongly convex. The smoothness and strong convexity of $f_0$ implies that

$$f_0^* + \frac{\sigma}{2}\|x - x^*\|_\infty^2 \leqslant f_0(x) \leqslant f_0^* + \frac{M}{2}\|x - x^*\|_\infty^2 \quad \forall x \in \mathcal{X}_{f_0,\kappa}. \tag{83}$$

Subsequently, there exist constants $c_0, C_1, C_2 > 0$ depending only on $\sigma, M, \kappa$ and $d$ such that for all $\epsilon \in (0, c_0]$,

$$B^\infty_{C_1\sqrt{\epsilon}}(x^*; \mathcal{X}) \subseteq L_{f_0}(\epsilon) \subseteq B^\infty_{C_2\sqrt{\epsilon}}(x^*; \mathcal{X}). \tag{84}$$

The property $\mu_{f_0}(\epsilon) \lesssim \epsilon^\beta$ holds because $\mu(L_{f_0}(\epsilon)) \leqslant \mu(B^\infty_{C_1\sqrt{\epsilon}}(x^*; \mathcal{X})) \lesssim \epsilon^{d/2}$. To prove (A2), note that $N(L_{f_0}(\epsilon), \delta) \leqslant N(B^\infty_{C_2\sqrt{\epsilon}}(x^*; \mathcal{X}), \delta) \lesssim 1 + (\sqrt{\epsilon}/\delta)^d$. Because $\epsilon^{d/2} \lesssim \mu(L_{f_0}(\epsilon)) = \mu_{f_0}(\epsilon)$, we conclude that $N(L_{f_0}(\epsilon), \delta) \lesssim 1 + \delta^{-d}\mu_{f_0}(\epsilon)$ and (A2) is thus proved. □

*Proof of Proposition 4.* Consider $f_0 \equiv 0$ if $\beta = 0$ and $f_0(z) := a_0\left[z_1^p + \ldots + z_d^p\right]$ for all $z = (z_1, \ldots, z_d) \in [0,1]^d$, where $a_0 > 0$ is a constant depending on $\alpha, M$, and $p = d/\beta$ for $\beta \in (0, d/\alpha]$. The $\beta = 0$ case where $f_0 \equiv 0$ trivially holds. So we shall only consider the case of $\beta \in (0, d/\alpha]$.

We first show $f_0 \in \Sigma_\kappa^\alpha(M)$ with $\kappa = \infty$, provided that $a_0$ is sufficiently small. For any $j \leqslant k = \lfloor\alpha\rfloor$ and $\alpha_1 + \ldots + \alpha_d = j$, we have

$$\frac{\partial^j}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}} f_0(z) = \begin{cases} a_0 j! \cdot z_\ell^{p-j} & \text{if } \alpha_\ell = j, \ell \in [d]; \\ 0 & \text{otherwise.}\end{cases} \tag{85}$$

Because $z_1, \ldots, z_d \in [0,1]$ and $p = d/\beta \geqslant \alpha \geqslant j$, it's clear that $0 \leqslant \partial^j f_0(z)/\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d} \leqslant a_0\ j!$. In addition, for any $z, z' \in [0,1]^d$ and $\alpha_\ell = k$, $\ell \in [d]$, we have

$$\begin{aligned}\left|\frac{\partial^k}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}} f_0(z) - \frac{\partial^k}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}} f_0(z')\right| &\leqslant a_0 k! \cdot \left|[z_\ell]^{p-k} - [z'_\ell]^{p-k}\right| \quad (86)\\ &\leqslant a_0 k! \cdot \left|z_\ell - z'_\ell\right|^{\min\{p-k,1\}}, \quad (87)\end{aligned}$$

where the last inequality holds because $x^t$ is $\min\{t, 1\}$-Hölder continuous on $[0,1]$ for $t \geqslant 0$. The $|z_\ell - z'_\ell|^{\min\{p-k,1\}}$ term can be further upper bounded by $\|z - z'\|_\infty^{\alpha-k}$, because $p = d/\beta \geqslant \alpha$. By selecting $a_0 > 0$ to be sufficiently small (depending on $M$) we have $f_0 \in \Sigma_\infty^\alpha(M)$.

We next prove $f_0$ satisfies $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ with parameter $\beta$ depending on $a_0$ and $p$. For any $\epsilon > 0$, the level-set $L_{f_0}(\epsilon)$ can be expressed as $L_{f_0}(\epsilon) = \{z \in [0,1]^d : z_1^p + \ldots + z_d^p \leqslant \epsilon/a_0\}$. Subsequently,

$$\left[0, \left(\frac{\epsilon}{a_0 d}\right)^{1/p}\right]^d \subseteq L_{f_0}(\epsilon) \subseteq \left[0, \left(\frac{\epsilon}{a_0}\right)^{1/p}\right]^d. \tag{88}$$

Therefore,

$$[\epsilon/(a_0 d)]^{dp} \leqslant \mu_{f_0}(\epsilon) \leqslant [\epsilon/a_0]^{dp}. \tag{89}$$

Because $a_0, d$ are constants and $dp = \beta$, we established $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ for $\beta = dp$.

Finally, note that for any $\epsilon > 0$, $L_{f_0}(\epsilon)$ is sandwiched between two cubics whose volumes only differ by a constant. This proves (A2) and (A2') on the covering and packing numbers of $L_{f_0}(\epsilon)$. □

*Proof of Proposition 5.* By the Chernoff bound and the union bound, with probability $1 - O(n^{-1})$ uniformly over all $x \in G_n$, there are $\Omega(\sqrt{n_0}\log^2 n)$ uniform samples in $B^\infty_{h_0}(x; \mathcal{X})$. Because $h_0 \leqslant \zeta$ for sufficiently large $n_0$ ($\zeta$ is defined in condition (A1)), by Lemma 1 it holds that

$$\left|\check{f}_x(x') - f_x(x')\right| \lesssim h_0^\alpha + n_0^{-1/4} \lesssim n_0^{-\alpha/2d} + n_0^{-1/4}, \quad \forall x \in G_n,\ x' \in B^\infty_{h_0}(x; \mathcal{X}). \tag{90}$$

Also, using the standard Gaussian concentration inequality, with probability $1 - O(n^{-1})$ we have

$$\inf_{x' \in B^\infty_{h_0}(x;\mathcal{X})} f(x) - O(n_0^{-1/4}) \leqslant \overline{f}(x) \leqslant \sup_{x' \in B^\infty_{h_0}(x;\mathcal{X})} f(x) + O(n_0^{-1/4}) \quad \forall x \in G_n. \tag{91}$$

Let $x^*$ be the minimizer of $f$ on $\mathcal{X}$ and $x \in G_n$ such that $\|x - x^*\|_\infty \leqslant h_0$. By (90), we have with probability $1 - O(n^{-1})$ that $\inf_{x' \in B^\infty_{h_0}(x;\mathcal{X})} \check{f}_x(x') \leqslant f^* + O(n_0^{-\alpha/2d} + n_0^{-1/4}) \leqslant f^* + 1/2\log n$, where $f^* = f(x^*)$. Now consider arbitrary $z \in G_n$ such that $B^\infty_{h_0}(z; \mathcal{X}) \cap L_f(\kappa/2) = \varnothing$, meaning that for all $z' \in \mathcal{X}$, $\|z' - z\|_\infty \leqslant h_0$, $f(z') > \kappa/2$. By (90), $\overline{f}(z) \geqslant \kappa/2 - O(n_0^{-1/4}) \geqslant \kappa/2 - 1/2\log n$. Hence when $n_0$ is sufficiently large, $z \notin S'_0$, which is to be demonstrated. □

*Proof of Proposition 7.* The upper bound part of (53) trivially holds because the absolute values of every element in $\psi_{0,1}(z)\psi_{0,1}(z)^\top$ for $z \in \mathcal{X} = [0,1]^d$ is upper bounded by $O(1)$. To prove the lower bound part, we only need to show $\int_\mathcal{X} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z)$ is invertible. Assume the contrary. Then there exists $\upsilon \in \mathbb{R}^D\backslash\{0\}$ such that

$$\upsilon^\top\left[\int_\mathcal{X} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z)\right]\upsilon = \int_\mathcal{X} \left|\psi_{0,1}(z)^\top\upsilon\right|^2 \mathrm{d}U(z) = 0. \tag{92}$$

Therefore, $\langle\psi_{0,1}(z), \upsilon\rangle = 0$ almost everywhere on $z \in [0,1]^d$. Because $h > 0$, by re-scaling with constants this implies the existence of non-zero coefficient vector $\xi$ such that

$$P(z_1, \ldots, z_m) := \sum_{\alpha_1+\ldots+\alpha_m \leqslant k} \xi_{\alpha_1,\ldots,\alpha_m} z_1^{\alpha_1} \ldots z_m^{\alpha_m} = 0$$
$$\text{almost everywhere on } z \in [0,1]^d.$$

We next use induction to show that, for any degree-$k$ polynomial $P$ of $s$ variables $z_1, \ldots, z_s$ that has at least one non-zero coefficient, the set $\{z_1, \ldots, z_s \in [0,1]^d : P(z_1, \ldots, z_s) = 0\}$ must have zero measure. This would then result in the desired contradiction. For the base case of $s = 1$, the fundamental theorem of algebra asserts that $P(z_1) = 0$ can have at most $k$ roots, which is a finite set and of measure 0.

We next consider the case where $P(z_1, \ldots, z_s)$ takes on $s$ variables. Re-organizing the terms we have

$$P(z_1, \ldots, z_s) \equiv P_0(z_1, \ldots, z_{s-1}) + z_s P_1(z_1, \ldots, z_{s-1}) + \ldots + z_s^k P_k(z_1, \ldots, z_{s-1}), \quad (93)$$

where $P_1, \ldots, P_k$ are degree-$k$ polynomials of $z_1, \ldots, z_{s-1}$. Because $P$ has a non-zero coefficient, at least one $P_j$ must also have a non-zero coefficient. By the inductive hypothesis, the set $\{z_1, \ldots, z_{s-1} : P_j(z_1, \ldots, z_{s-1})\}$ has measure 0. On the other hand, if $P_j(z_1, \ldots, z_{s-1}) \neq 0$, then invoking the fundamental theorem of algebra again on $z_s$ we know that there are finitely many $z_s$ such that $P(z_1, \ldots, z_s) = 0$. Therefore, $\{z_1, \ldots, z_s : P(z_1, \ldots, z_s) = 0\}$ must also have measure zero. □

## References


[1] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2006.

[2] B. Reeja-Jayan, K. L. Harrison, K. Yang, C.-L. Wang, A. E. Yilmaz, and A. Manthiram, "Microwave-assisted low-temperature growth of thin films in solution," *Sci. Rep.*, vol. 2, Dec. 2012, Art. no. 1003.

[3] N. Nakamura, J. Seepaul, J. B. Kadane, and B. Reeja-Jayan, "Design for low-temperature microwave-assisted crystallization of ceramic thin films," *Appl. Stochastic Models Bus. Ind.*, vol. 33, no. 3, pp. 314–321, 2017.

[4] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics). New York, NY, USA: Springer, 2009.

[5] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications*. Boca Raton, FL, USA: CRC Press, 1996.

[6] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2879–2904, Oct. 2011.

[7] J. Scarlett, I. Bogunovic, and V. Cevher, "Lower bounds on regret for noisy Gaussian process bandit optimization," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2017, pp. 1723–1742.

[8] E. Hazan, A. Klivans, and Y. Yuan, "Hyperparameter optimization: A spectral approach," 2017, *arXiv:1706.00764*. [Online]. Available: https://arxiv.org/abs/1706.00764#

[9] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Hoboken, NJ, USA: Wiley, 1983.

[10] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2005, pp. 385–394.

[11] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2010, pp. 28–40.

[12] K. G. Jamieson, R. Nowak, and B. Recht, "Query complexity of derivative-free optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 2672–2680.

[13] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," *SIAM J. Optim.*, vol. 23, no. 1, pp. 213–240, 2013.

[14] S. Bubeck, Y. T. Lee, and R. Eldan, "Kernel-based methods for bandit convex optimization," in *Proc. 49th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, 2017, pp. 72–85.

[15] A. H. G. R. Kan and G. T. Timmer, "Stochastic global optimization methods part I: Clustering methods," *Math. Program.*, vol. 39, no. 1, pp. 27–56, 1987.

[16] A. H. G. R. Kan and G. T. Timmer, "Stochastic global optimization methods part II: Multi level methods," *Math. Program.*, vol. 39, no. 1, pp. 57–78, 1987.

[17] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "X-armed bandits," *J. Mach. Learn. Res.*, vol. 12, pp. 1655–1695, May 2011.

[18] C. Malherbe, E. Contal, and N. Vayatis, "A ranking approach to global optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016. AQ:2

[19] C. Malherbe and N. Vayatis, "Global optimization of Lipschitz functions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.

[20] R. D. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2005, pp. 697–704.

[21] S. Minsker, "Estimation of extreme values and associated level sets of a regression function via selective sampling," in *Proc. Conf. Learn. Theory (COLT)*, 2013, pp. 105–121.

[22] J.-B. Grill, M. Valko, and R. Munos, "Black-box optimization of noisy functions with unknown smoothness," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 667–675.

[23] S. Minsker, "Non-asymptotic bounds for prediction problems and density estimation," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, Georgia, 2012.

[24] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Stat.*, vol. 23, no. 3, pp. 462–466, 1952.

[25] E. Purzen, "On estimation of a probability density and mode," *Ann. Math. Statist.*, vol. 39, no. 3, pp. 1065–1076, 1962.

[26] H. Chen, "Lower rate of convergence for locating a maximum of a function," *Ann. Statist.*, vol. 16, no. 3, pp. 1330–1334, 1988.

[27] Z. B. Zabinsky and R. L. Smith, "Pure adaptive search in global optimization," *Math. Program.*, vol. 53, no. 1, pp. 323–338, 1992.

[28] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," *J. Comput. Syst. Sci.*, vol. 75, no. 1, pp. 78–89, 2009.

[29] S. Dasgupta, D. J. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 353–360.

[30] S. Hanneke, "A bound on the label complexity of agnostic active learning," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 353–360.

[31] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, pp. 1079–1105, Jun. 2006.

[32] W. Polonik, "Measuring mass concentrations and estimating density contour clusters-an excess mass approach," *Ann. Statist.*, vol. 23, no. 3, pp. 855–881, 1995.

[33] P. Rigollet and R. Vert, "Optimal rates for plug-in estimators of density level sets," *Bernoulli*, vol. 15, no. 4, pp. 1154–1178, 2009.

[34] A. Singh, C. Scott, and R. Nowak, "Adaptive Hausdorff estimation of density level sets," *Ann. Statist.*, vol. 37, no. 5B, pp. 2760–2782, 2009.

[35] K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. V. Luxburg, "Consistent procedures for cluster tree estimation and pruning," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7900–7912, Dec. 2014.

[36] S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman, "Cluster trees on manifolds," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2679–2687.

[37] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, 2006.

[38] E. Hazan, K. Levy, and S. Shalev-Shwartz, "Beyond convexity: Stochastic quasi-convex optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1594–1602.

[39] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2015, pp. 797–842.

[40] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *Proc. 49th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, 2017, pp. 1195–1199.

[41] Y. Carmon, O. Hinder, J. C. Duchi, and A. Sidford, "'Convex until proven guilty': Dimension-free acceleration of gradient descent on non-convex functions," 2017, *arXiv:1705.02766*. [Online]. Available: https://arxiv.org/abs/1705.02766

[42] Y. Zhang, P. Liang, and M. Charikar, "A hitting time analysis of stochastic gradient Langevin dynamics," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2017, pp. 1–43.

[43] Y. Zhu, S. Chatterjee, J. Duchi, and J. Lafferty, "Local minimax complexity of stochastic convex optimization," in *Proc. NIPS*, 2016, pp. 3431–3439.

[44] J. Duchi and F. Ruan, "Asymptotic optimality in stochastic optimization," 2016, *arXiv:1612.05612*. [Online]. Available: https://arxiv.org/abs/1612.05612

[45] A. Locatelli and A. Carpentier, "Adaptivity to smoothness in X-armed bandits," in *Proc. Conf. Learn. Theory (COLT)*, 2018, pp. 1463–1492.

[46] A. W. van der Vaart, *Asymptotic Statistics*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[47] C. Jin, L. T. Liu, R. Ge, and M. I. Jordan, "On the local minima of the empirical risk," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1–10.

[48] T. T. Cai and M. G. Low, "An adaptation theory for nonparametric confidence intervals," *Ann. Statist.*, vol. 32, no. 5, pp. 1805–1840, 2004.

AQ:3 [49] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*, vol. 82. Springer, 2012.

[50] R. M. Castro and R. D. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.

[51] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Proc. Int. Conf. Algorithmic Learn. Theory (ALT)*, 2009, pp. 23–37.

[52] O. V. Lepski, E. Mammen, and V. G. Spokoiny, "Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors," *Ann. Statist.*, vol. 25, no. 3, pp. 929–947, 1997.

[53] W. K. Newey, "Convergence rates and asymptotic normality for series estimators," *J. Econometrics*, vol. 79, no. 1, pp. 147–168, 1997.

[54] D. S. Ebert, *Texturing & Modeling: A Procedural Approach*. San Mateo, CA, USA: Morgan Kaufmann, 2003.

[55] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, vol. 20, no. 4, pp. 2217–2246, 2014.

[56] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

[57] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electron. Commun. Probab.*, vol. 17, no. 52, pp. 1–6, 2012.

[58] J. A. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, nos. 1–2, pp. 1–230, 2015.

**Yining Wang** received the B.Eng. degree in computer science and technology in 2014 from Tsinghua University, Beijing China, the M.S. degree in machine learning in 2017 from Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Ph.D. student in machine learning in the machine learning department at Carnegie Mellon University, Pittsburgh, PA, USA. His research interests are primarily in statistical machine learning, with emphasis on interactive methods, active learning, adaptive sampling.

**Sivaraman Balakrishnan** is an Assistant Professor in the Department of Statistics and Data Science at Carnegie Mellon University. Prior to this he received his Ph.D. from the School of Computer Science at Carnegie Mellon University and was a postdoctoral researcher in the Department of Statistics at UC Berkeley. His Ph.D. work was supported by several fellowships including the Richard King Mellon Fellowship and a grant from the Gates Foundation. He is broadly interested in problems that lie at the interface between computer science and statistics. Some particular areas that have provided motivation for his past and current research include the applications of statistical methods in ranking problems, computational biology, clustering, topological data analysis, nonparametric statistics, robust statistics and non-convex optimization.

**Aarti Singh** received the B.E. degree in electronics and communication engineering from the University of Delhi, New Delhi, India, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 2003 and 2008, respectively. She was a Postdoctoral Research Associate at the Program in Applied and Computational Mathematics, Princeton University, from 2008 to 2009, before joining the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, where she has been an Associate Professor since 2009. Her research interests include the intersection of machine learning, statistics and signal processing, and focus on designing statistically and computationally efficient algorithms that can leverage inherent structure of the data in the form of clusters, graphs, subspaces, and manifold using direct, compressive, and active queries. Her work is recognized by the NSF Career Award, the United States Air Force Young Investigator Award, A. Nico Habermann Faculty Chair Award, Harold A. Peterson Best Dissertation Award, and a best student paper award at Allerton.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

**PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.**

AQ:1 = Author: Please confirm or add details for any funding or financial support for the research of this article.

AQ:2 = Please provide the page range for Refs. [18] and [19].

AQ:3 = Please provide the publisher location for Ref. [49].

# Optimization of Smooth Functions With Noisy Observations: Local Minimax Rates

Yining Wang, Sivaraman Balakrishnan, and Aarti Singh

***Abstract*— We consider the problem of *global optimization* of an unknown non-convex smooth function with noisy zeroth-order feedback. We propose a *local minimax* framework to study the fundamental difficulty of optimizing smooth functions with adaptive function evaluations. We show that for functions with fast growth around their global minima, carefully designed optimization algorithms can identify a near global minimizer with many fewer queries than worst-case global minimax theory predicts. For the special case of strongly convex and smooth functions, our implied convergence rates match the ones developed for zeroth-order *convex* optimization problems. On the other hand, we show that in the worst case no algorithm can converge faster than the minimax rate of estimating an unknown function in the $\ell_\infty$-norm. Finally, we show that non-adaptive algorithms, though optimal in a global minimax sense, do not attain the optimal local minimax rate.**

***Index Terms*— Optimization of smooth functions, nonparametric statistics, local minimax analysis.**

## I. Introduction

Global function optimization with stochastic (zeroth-order) query oracles is an important problem in optimization, machine learning and statistics. To optimize an unknown bounded function $f : \mathcal{X} \mapsto \mathbb{R}$ defined on a known compact $d$-dimensional domain $\mathcal{X} \subseteq \mathbb{R}^d$, the data analyst makes $n$ *active* queries $x_1, \ldots, x_n \in \mathcal{X}$ and observes

$$y_t = f(x_t) + w_t, \qquad w_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1),^1 \;\; t = 1, \ldots, n. \quad (1)$$

The queries $x_1, \ldots, x_t$ are *active* in the sense that the selection of $x_t$ can depend on the previous queries and their responses $x_1, y_1, \ldots, x_{t-1}, y_{t-1}$. After $n$ queries, an estimate $\widehat{x}_n \in \mathcal{X}$ is produced that approximately minimizes the unknown function $f$. Such "active query" models are relevant in a broad range of (noisy) global optimization applications, for instance in hyper-parameter tuning of machine learning algorithms [1] and sequential design in material synthesis experiments where the goal is to maximize the strength of the synthesized material as a function of experimental settings [2], [3]. We refer the readers to Section II-A for a rigorous formulation of the active query model and contrast it with the classical passive query model.

The error of the estimate $\widehat{x}_n$ is measured by the difference of $f(\widehat{x}_n)$ and the *global minimum* of $f$:

$$\mathfrak{L}(\widehat{x}_n; f) := f(\widehat{x}_n) - f^* \qquad \text{where } f^* := \inf_{x\in\mathcal{X}} f(x). \quad (2)$$

To simplify our presentation, throughout the paper we take the domain $\mathcal{X}$ to be the $d$-dimensional unit cube $[0, 1]^d$, while our results can be easily generalized to other compact domains satisfying minimal regularity conditions.

When $f$ belongs to a smoothness class, say the Hölder class with exponent $\alpha$, a straightforward global optimization method is to first sample $n$ points uniformly at random from $\mathcal{X}$ and then construct nonparametric estimates $\widehat{f}_n$ of $f$ using nonparametric regression methods such as kernel smoothing or local polynomial regression [4], [5]. Classical analysis shows that the sup-norm reconstruction error $\|\widehat{f}_n - f\|_\infty = \sup_{x\in\mathcal{X}} |\widehat{f}_n(x) - f(x)|$ can be upper bounded by $\widetilde{O}_{\mathbb{P}}(n^{-\alpha/(2\alpha+d)})^2$. This global reconstruction guarantee then implies an $\widetilde{O}_{\mathbb{P}}(n^{-\alpha/(2\alpha+d)})$ upper bound on $\mathfrak{L}(\widehat{x}_n; f)$ by considering an estimate $\widehat{x}_n \in \mathcal{X}$ for which $\widehat{f}_n(\widehat{x}_n) = \inf_{x\in\mathcal{X}} \widehat{f}_n(x)$ (such an $\widehat{x}_n$ exists because $\mathcal{X}$ is closed and bounded). Formally, we have the following proposition (proved in the Appendix) that converts a global reconstruction guarantee into an upper bound on the optimization error:

**Proposition 1.** *Suppose $\widehat{f}_n(\widehat{x}_n) = \inf_{x\in\mathcal{X}} \widehat{f}_n(x)$. Then $\mathfrak{L}(\widehat{x}_n; f) \leqslant 2\|\widehat{f}_n - f\|_\infty$.*

Typically, fundamental limits on the optimal optimization error are understood through the lens of *minimax analysis* where the object of study is the (global) minimax risk:

$$\inf_{\widehat{x}_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f \mathfrak{L}(\widehat{x}_n, f), \quad (3)$$

where $\mathcal{F}$ is a certain class of smooth functions such as the Hölder class. Although optimization appears to be easier than global reconstruction, we show in this paper that the $n^{-\alpha/(2\alpha+d)}$ rate is *not* improvable in the global minimax sense in over Hölder classes. Such a surprising phenomenon was also noted in previous works [6]–[8] for related problems. On the

AQ:1

Manuscript received August 10, 2018; revised April 21, 2019; accepted May 5, 2019. S. Balakrishnan was supported in part by the NSF under Grant DMS-17130003. Y. Wang and A. Singh were supported in part by the NSF under Grant CCF-1563918 and in part by the AFRL under Grant FA8750-17-2-0212. This paper was presented in part at the 2018 NeurIPS Conference.

Y. Wang and A. Singh are with the Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: yiningwa,aarti@cs.cmu.edu).

S. Balakrishnan is with the Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: siva@stat.cmu.edu).

Communicated by K. Chaudhuri, Associate Editor for Statistical Learning.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2921985

[1] The exact Gaussianity of the independent noise variables $\varepsilon_t$ is not crucial and our results can be easily generalized to sub-Gaussian noise.

[2] In the $\widetilde{O}(\cdot)$ or $\widetilde{O}_{\mathbb{P}}(\cdot)$ notation we suppress constant factors and terms that depend poly-logarithmically on $n$.

other hand, extensive empirical evidence suggests that non-uniform/active allocations of query points can significantly reduce optimization error in practical global optimization of smooth, non-convex functions [1]. This raises the interesting question of understanding, from a theoretical perspective, the conditions under which the global optimization of smooth functions is *easier* than their reconstruction, and the power of *active/feedback-driven* queries that play important roles in global optimization.

In this paper, we propose a theoretical framework that partially answers the above questions. In contrast to classical *global* minimax analysis of nonparametric estimation problems, we adopt a *local analysis* which characterizes the optimal convergence rate of optimization error when the underlying function $f$ is within a neighborhood of a "reference" function $f_0$. (See Section II-B for the rigorous local minimax formulation considered in this paper.) Our main results are to characterize the local convergence rates $R_n(f_0)$ for a wide range of reference functions $f_0 \in \mathcal{F}$. Concretely, our contributions can be summarized as follows:

1) We design an iterative (active) algorithm whose optimization error $\mathfrak{L}(\widehat{x}_n; f)$ converges at a rate of $R_n(f_0)$ depending on the reference function $f_0$. When the level-sets of $f_0$ satisfy certain regularity and polynomial growth conditions, the local rate $R_n(f_0)$ can be upper bounded by $R_n(f_0) = \widetilde{O}(n^{-\alpha/(2\alpha+d-\alpha\beta)})$, where $\beta \in [0, d/\alpha]$ is a parameter depending on $f_0$ that characterizes the volume growth of the *level-sets* of the reference function $f_0$. (See assumption (A2), Proposition 2 and Theorem 1 for details). The rate matches the global minimax convergence rate $n^{-\alpha/(2\alpha+d)}$ for worst-case $f_0$ where $\beta = 0$, but can be much faster when $\beta > 0$. We emphasize that our algorithm has no knowledge of the reference function $f_0$ and achieves this rate adaptively.
2) We prove *local* minimax lower bounds that match the $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ upper bound, up to logarithmic factors in $n$. More specifically, we show that *even if $f_0$ is known*, no (active) algorithm can estimate $f$ in close neighborhoods of $f_0$ at a rate faster than $n^{-\alpha/(2\alpha+d-\alpha\beta)}$. We further show that, if active queries are not available and queries $x_1, \ldots, x_n$ are i.i.d. uniformly sampled from $\mathcal{X}$, then the $n^{-\alpha/(2\alpha+d)}$ global minimax rate also applies locally regardless of how large $\beta$ is. Thus, there is an explicit gap between local minimax rates in the active and uniform query models when $\beta$ is large.
3) In the special case when $f$ is *convex*, the global optimization problem is usually referred to as *zeroth-order convex optimization* and this problem has been widely studied [9]–[14]. Our results imply that, when $f_0$ is *strongly* convex and smooth, the local minimax rate $R_n(f_0)$ is on the order of $\widetilde{O}(n^{-1/2})$, which matches the convergence rates in [11]. Additionally, our negative results (Theorem 2) indicate that the $n^{-1/2}$ rate cannot be achieved if $f_0$ is merely convex, which seems to contradict $n^{-1/2}$ results in [13], [14] that do not require strong convexity of $f$. However, it should be noted that mere convexity of $f_0$ does *not* imply convexity of $f$ in a neighborhood of $f_0$ (e.g., $\|f - f_0\|_\infty \leqslant \varepsilon$). Our results show significant differences in the intrinsic difficulty of zeroth-order optimization of convex and near-convex functions.

## *A. Related Work*

*Global optimization*, known variously as *black-box optimization*, *Bayesian optimization* and the *continuum-armed bandit*, has a long history in the optimization research community [15], [16] and has also received a significant amount of recent interest in statistics and machine learning [1], [6], [8], [17]–[19]. Many previous works [17], [20] have derived rates for non-convex smooth payoffs in "continuum-armed" bandit problems.

The papers [21], [22] are closely related to our work. They studied the related problem of estimating the set of all optima of a smooth function in the Hausdorff distance. For Hölder smooth functions with polynomial growth, the paper [21] derives an $n^{-1/(2\alpha+d-\alpha\beta)}$ minimax rate for $\alpha < 1$ (subsequently improved to include $\alpha \geqslant 1$ in [23]). This result is similar to our Propositions 2 and 3. The papers [21], [22] also discussed adaptivity to unknown smoothness parameters. We however remark on several differences between our work and the papers [21], [22]. First, in [21], [22] only functions with polynomial growth are considered, while in our Theorems 1 and 2 functionals $\varepsilon_n^{\mathsf{U}}(f_0)$ and $\varepsilon_n^{\mathsf{L}}(f_0)$ are proposed for general reference functions $f_0$ satisfying mild regularity conditions, which include functions with polynomial growth as special cases. In addition, [21] considers the harder problem of estimating maxima sets in Hausdorff distance, as opposed to the problem of producing a single approximately optimal solution $\widehat{x}_T$. As a result, the minimax lower bounds in [21] do not apply to this latter setting. An algorithm, without distinguishing between two functions with different optima sets, can nevertheless produce a good approximate optimizer as long as the two functions under consideration have *overlapping* optima sets. New constructions and information-theoretic techniques are therefore required to prove lower bounds under the weaker (one-point) approximate optimization framework. Finally, we prove minimax lower bounds when only *uniform* query points are available and demonstrate a significant gap between algorithms having access to uniformly sampled or adaptively chosen data points.

The papers [18], [19] imposed additional assumptions on the level-sets of the underlying function to obtain an improved convergence rate. The level-set assumptions considered in the mentioned references are rather restrictive and essentially require the underlying function to be uni-modal, while our assumptions are much more flexible and apply to multi-modal functions as well. In addition, [18], [19] considered a *noiseless* setting in which exact function evaluations $f(x_t)$ can be obtained, while our paper studies the noise corrupted model in (1) for which vastly different convergence rates are derived. Finally, no matching lower bounds were proved in the papers [18], [19].

The (stochastic) global optimization problem is similar to *mode estimation* of either densities or regression functions, which has a rich literature [24]–[26]. An important difference between statistical mode estimation and global optimization is the way sample/query points $x_1, \ldots, x_n \in \mathcal{X}$ are distributed: in mode estimation it is customary to assume the samples are independently and identically distributed, while in global optimization sequential designs of samples/queries are typical. Furthermore, to estimate/locate the mode of an unknown density or regression function, such a mode has to be well-defined; on the other hand, producing an estimate $\widehat{x}_n$ with small $\mathfrak{L}(\widehat{x}_n, f)$ is easier and results in weaker conditions imposed on the underlying function.

Methodology-wise, our proposed algorithm is conceptually similar to the abstract *Pure Adaptive Search (PAS)* framework proposed and analyzed in [27]. The iterative procedure also resembles disagreement-based active learning methods [28]–[30] and the "successive rejection" algorithm in bandit problems [31]. The intermediate steps of candidate point elimination can also be viewed as level-set estimation problems [32]–[34] or cluster-tree estimation problems [35], [36] with active queries.

Another line of research has focused on *first-order* optimization of quasi-convex or non-convex functions [37]–[42], in which exact or unbiased evaluations of function *gradients* are available at query points $x \in \mathcal{X}$. The paper [42] considered a Cheeger's constant restriction on level-sets which is similar to our level-set regularity assumptions (A2 and A2'). The papers [43], [44] studied local minimax rates for the first-order optimization of convex functions. First-order optimization differs significantly from our setting because unbiased gradient estimation is generally impossible in the model of (1). Furthermore, most works on (first-order) non-convex optimization focus on obtaining stationary points or local minima, while we consider the problem of finding a (near) global minima.

### B. Comparison with the HOO Algorithm

The HOO algorithm [17], as well as similar algorithms such as Algorithm 2 in [45] and the POO algorithm in [22], are theoretically well-studied methods for global optimization. Below we summarize the differences of our results and the ones from these works.

(a) Weaker Smoothness Conditions I: In Algorithm 1, we use local polynomial estimation as a sub-routine to obtain local estimates of the objective function $f$. Compared to the sample average approach in HOO (e.g., Algorithm 2 in [45]), local polynomial estimates have the advantage of being unbiased for the estimation of low-degree polynomials. This translates to the improved (A1) Hölder-continuity condition that *only* restricts the $\lfloor\alpha\rfloor$-th order derivatives of objective functions. More specifically, the actual function values of $f(x)$ and $f(x')$ for $x, x'$ close to each other can be very different, as long as such differences can be perfectly modeled by low-degree polynomials. This is in contrast to the smoothness conditions imposed in [17], [45] which essentially require $f(x)$ to be close to $f(x^*)$ for $x$ close to $x^*$ the optima of $f$.

(b) Weaker Smoothness Conditions II: Our results in Section IV-C hold on functions that are only assumed to be smooth in regions close to its global minimum, in contrast to Definition 1 in [45] and many other existing works that place smoothness assumptions on the entire domain of the objective function $f$.

(c) Spatially Restricted Queries: Our proposed algorithm is "grid" based, and can be run on any sufficiently dense finite grid $G_n$ in $\mathcal{X}$ and does not need to have the capacity to query arbitrary points in $\mathcal{X}$. As a result, our algorithm can be run in experimental settings where queries are restricted to belong to a large pool of a-priori chosen points.

(d) Results for any Smooth Function: Our algorithm and lower bounds yield essentially tight results for the complexity of optimization of arbitrary smooth functions. While these rates are most interpretable under the level-set growth conditions (also studied in [45]) our results also yield nearly matching guarantees for other (arbitrary, smooth) functions $f_0$.

## II. Background and Notation

We first review standard asymptotic notation that will be used throughout this paper. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $\limsup_{n\to\infty} |a_n|/|b_n| < \infty$, or equivalently $b_n = \Omega(a_n)$ or $b_n \gtrsim a_n$. Denote $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We also write $a_n = o(b_n)$ or equivalently $b_n = \omega(a_n)$ if $\lim_{n\to\infty} |a_n|/|b_n| = 0$. For two sequences of random variables $\{A_n\}_{n=1}^{\infty}$ and $\{B_n\}_{n=1}^{\infty}$, denote $A_n = O_{\mathbb{P}}(B_n)$ if for every $\epsilon > 0$, there exists $C > 0$ such that $\limsup_{n\to\infty} \Pr[|A_n| > C|B_n|] \leqslant \epsilon$. For $r > 0$, $1 \leqslant p \leqslant \infty$ and $x \in \mathbb{R}^d$, we denote by $B_r^p(x) := \{z \in \mathbb{R}^d : \|z - x\|_p \leqslant r\}$ the $d$-dimensional $\ell_p$-ball of radius $r$ centered at $x$, where the vector $\ell_p$ norm is defined as $\|x\|_p := (\sum_{j=1}^d |x_j|^p)^{1/p}$ for $1 \leqslant p < \infty$ and $\|x\|_\infty := \max_{1\leqslant j\leqslant d} |x_j|$. For any subset $S \subseteq \mathbb{R}^d$ we denote by $B_r^p(x; S)$ the set $B_r^p(x) \cap S$.

### A. Passive and Active Query Models

Let $U$ be a known random quantity defined on a probability space $\mathcal{U}$. The following definitions characterize all passive and active optimization algorithms:

**Definition 1** (The passive query model)**.** *Let $x_1, \ldots, x_n$ be i.i.d. points uniformly sampled on $\mathcal{X}$ and $y_1, \ldots, y_n$ be observations from the model* (1)*. A* passive optimization algorithm *$\mathcal{A}$ with $n$ queries is parameterized by a mapping $\phi_n$ : $(x_1, y_1, \ldots, x_n, y_n, U) \mapsto \widehat{x}_n$ that maps the i.i.d. observations $\{(x_i, y_i)\}_{i=1}^n$ to an estimated optimum $\widehat{x}_n \in \mathcal{X}$, potentially randomized by $U$.*

**Definition 2** (The active query model)**.** *An* active optimization algorithm *can be parameterized by mappings $(\chi_1, \ldots, \chi_n, \phi_n)$, where for $t = 1, \ldots, n$,*

$$\chi_t : (x_1, y_1, \ldots, x_{t-1}, y_{t-1}, U) \mapsto x_t$$

*produces a query point* $x_t \in \mathcal{X}$ *based on previous observations* $\{(x_i, t_i)\}_{i=1}^{t-1}$, *and*

$$\phi_n : (x_1, y_1, \ldots, x_n, y_n, U) \mapsto \widehat{x}_n$$

*produces the final estimate. All mappings* $(\chi_1, \ldots, \chi_n, \phi_n)$ *can be randomized by* $U$.

### B. Local Minimax Rates

We use a classical *local minimax analysis* [46] to understand the fundamental information-theoretic limits of noisy global optimization of smooth functions. On the upper bound side, we seek (active) estimators $\widehat{x}_n$ such that

$$\sup_{f_0 \in \Theta} \sup_{f \in \Theta', \|f - f_0\|_\infty \leqslant \varepsilon_n(f_0)} \Pr_f \left[\mathfrak{L}(\widehat{x}_n; f) \geqslant C_1 \cdot R_n(f_0)\right] \leqslant 1/4, \tag{4}$$

where $C_1 > 0$ is a positive constant. Here $f_0 \in \Theta$ is referred to as the *reference function*, and $f \in \Theta'$ is the true underlying function to be optimized, which is assumed to be "near" $f_0$ (in the $\ell_\infty$ norm). The minimax convergence rate of $\mathfrak{L}(\widehat{x}_n; f)$ is then characterized *locally* by $R_n(f_0)$ which depends on the reference function $f_0$. The constant of $1/4$ is chosen arbitrarily and any small constant leads to similar conclusions. To establish negative results (i.e., local minimax lower bounds), in contrast to the upper bound formulation, we assume the potential active optimization estimator $\widehat{x}_n$ has *perfect knowledge* about the reference function $f_0 \in \Theta$. We then prove local minimax lower bounds of the form

$$\inf_{\widehat{x}_n} \sup_{f \in \Theta', \|f - f_0\|_\infty \leqslant \varepsilon_n(f_0)} \Pr_f \left[\mathfrak{L}(\widehat{x}_n; f) \geqslant C_2 \cdot R_n(f_0)\right] \geqslant 1/3, \tag{5}$$

where $C_2 > 0$ is another positive constant and $\varepsilon_n(f_0)$, $R_n(f_0)$ are desired local convergence rates for functions near the reference $f_0$.

Although in some sense classical, the local minimax definition we propose warrants further discussion:

1) **Roles of $\Theta$ and $\Theta'$:** The reference function $f_0$ and the true functions $f$ are assumed to belong to different but closely related function classes $\Theta$ and $\Theta'$. In particular, in our paper $\Theta \subseteq \Theta'$, meaning that less restrictive assumptions are imposed on the true underlying function $f$ compared to those imposed on the reference function $f_0$ on which $R_n$ and $\varepsilon_n$ are based.
2) **Upper Bounds:** It is worth emphasizing that the estimator $\widehat{x}_n$ has no knowledge of the reference function $f_0$. From the perspective of upper bounds, we can consider the simpler task of producing $f_0$-dependent bounds (eliminating the second supremum) to instead study the (already interesting) quantity:
$$\sup_{f_0 \in \Theta} \Pr_{f_0} \left[\mathfrak{L}(\widehat{x}_n; f_0) \geqslant C_1 R_n(f_0)\right] \leqslant 1/4.$$
As indicated above we maintain the double-supremum in the definition because fewer assumptions are imposed directly on the true underlying function $f$, and further because it allows to more directly compare our upper and lower bounds.
3) **Lower Bounds and the choice of the "localization radius" $\varepsilon_n(f_0)$:** Our lower bounds allow the estimator knowledge of the reference function (this makes establishing the lower bound more challenging). The lower bound in (5) implies that no estimator $\widehat{x}_n$ can effectively optimize a function $f$ close to $f_0$ beyond the convergence rate of $R_n(f_0)$, even if perfect knowledge of the reference function $f_0$ is available a priori. The $\varepsilon_n(f_0)$ parameter that decides the "range" in which local minimax rates apply is taken to be on the same order as the actual local rate $R_n(f_0)$ in this paper. This is (up to constants) the smallest radius for which we can hope to obtain non-trivial lower-bounds: if we consider a much smaller radius than $R_n(f_0)$ then the trivial estimator which outputs the minimizer of the reference function would achieve a faster rate than $R_n(f_0)$. On the other hand selecting the smallest possible radius makes establishing the lower bound most challenging but provides a refined picture of the complexity of zeroth-order optimization.

We remark that our primary motivation for the local-minimax analysis stems from the fact that for natural function classes the global-minimax rate for the optimization complexity is excessively pessimistic, while the local minimax analysis provides a more refined picture. In machine learning applications, there are several cases where the population risk is well-behaved (smooth, potentially non-convex) but we are only able to access/query the empirical risk which we want to minimize. Using standard concentration bounds the empirical risk and population risk are close, and the resulting problem is then to minimize the approximate-smooth empirical risk (see for instance [42], [47] for a more detailed discussion).

## III. Main Results

With this background in place we now turn our attention to our main results. We begin by collecting our assumptions about the true underlying function and the reference function in Section III-A. We state and discuss the consequences of our upper and lower bounds in Sections III-B and III-C respectively. We defer most technical proofs to Section V and turn our attention to our optimization algorithm in Section IV.

### A. Assumptions

We first state and motivate assumptions that will be used. The first assumption states that $f$ is locally Hölder smooth on its level-sets.

(A1) There exist constants $\kappa, \alpha, M, \zeta > 0$ such that $f$ restricted to $\mathcal{X}_{f,\kappa,\zeta} := \{x \in \mathcal{X} : \inf_{z \in \mathcal{X}, \|z-x\|_\infty \leqslant \zeta} f(z) \leqslant f^* + \kappa\}$ belongs to the Hölder class $\Sigma^\alpha(M)$, meaning that $f$ is $k$-times differentiable on $\mathcal{X}_{f,\kappa,\zeta}$ and furthermore for any $x, x' \in \mathcal{X}_{f,\kappa,\zeta}$[3],
$$\sum_{\alpha_1 + \ldots + \alpha_d = k} \frac{|f^{(\boldsymbol{\alpha},k)}(x) - f^{(\boldsymbol{\alpha},k)}(x')|}{\|x - x'\|_\infty^{\alpha-k}} \leqslant M. \tag{6}$$

[3]We use the $\ell_\infty$-norm for convenience and it can be replaced by any equivalent vector norm.

Here $k = \lfloor \alpha \rfloor$ is the largest integer lower bounding $\alpha$ and $f^{(\boldsymbol{\alpha},j)}(x) := \partial^j f(x)/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}$.

We use $\Sigma_\kappa^\alpha(M)$ to denote the class of all functions satisfying (A1). We remark that (A1) is weaker than the usual Hölder assumption in two ways. First, (6) only imposes stability conditions on the $\lfloor \alpha \rfloor$-th order derivatives of the function $f$, in contrast to conditions involving all orders of derivatives in previous works [17], [45]. Second, (A1) only imposes the Hölder smoothness assumption on certain regions of $\mathcal{X}$, because regions with function values larger than $f^* + \kappa$ can be easily detected and removed by a pre-processing step, highlighting an important difference between optimization and $\ell_\infty$-norm estimation. We give further details of the pre-processing step in Section IV-C.

Our next assumption concerns the "regularity" of the *level-sets* of the "reference" function $f_0$. Define $L_{f_0}(\epsilon) := \{x \in \mathcal{X} : f_0(x) \leqslant f_0^* + \epsilon\}$ as the $\epsilon$-level-set of $f_0$, and $\mu_{f_0}(\epsilon) := \lambda(L_{f_0}(\epsilon))$ as the Lebesgue measure of $L_{f_0}(\epsilon)$, which we refer to as the *distribution function*. Define, $N(L_{f_0}(\epsilon), \delta)$ as the smallest number of $\ell_2$-balls of radius $\delta$ that cover $L_{f_0}(\epsilon)$. Then we make the following assumption:

(A2) There exist constants $c_0 > 0$ and $C_0 > 0$ such that $N(L_{f_0}(\epsilon), \delta) \leqslant C_0[1 + \mu_{f_0}(\epsilon)\delta^{-d}]$ for all $\epsilon, \delta \in (0, c_0]$.

We use $\Theta_{\mathbf{C}}$ to denote all functions that satisfy (A2) with respect to parameters $\mathbf{C} = (c_0, C_0)$.

At a high-level, the regularity condition (A2) assumes that the level-sets are sufficiently "regular" such that covering them with small-radius balls does not require significantly larger total volume. For example, consider the perfectly regular case when $L_{f_0}(\epsilon)$ is the $d$-dimensional $\ell_2$ ball of radius $r$: $L_{f_0}(\epsilon) = \{x \in \mathcal{X} : \|x - x^*\|_2 \leqslant r\}$. Clearly, $\mu_{f_0}(\epsilon) \asymp r^d$. In addition, the $\delta$-covering number in $\ell_2$ of $L_{f_0}(\epsilon)$ is on the order of $1 + (r/\delta)^d \asymp 1 + \mu_{f_0}(\epsilon)\delta^{-d}$, which satisfies the scaling in (A2).

When (A2) holds, uniform confidence intervals for $f$ on its level-sets are easier to construct because little statistical efficiency is lost by slightly enlarging the level-sets so that complete (sufficiently small) $d$-dimensional cubes are contained in the enlarged level-sets. On the other hand, when regularity of level-sets fails to hold such nonparametric estimation can be very difficult or even impossible. As an extreme example, suppose the level-set $L_{f_0}(\epsilon)$ consists of $\mathfrak{n}$ standalone and well-spaced points in $\mathcal{X}$: the Lebesgue measure of $L_{f_0}(\epsilon)$ would be zero, but at least $\Omega(\mathfrak{n})$ queries are necessary to construct uniform confidence intervals on $L_{f_0}(\epsilon)$. It is clear that such $L_{f_0}(\epsilon)$ violates (A2), because $N(L_{f_0}(\epsilon), \delta) \geqslant \mathfrak{n}$ as $\delta \to 0^+$ but $\mu_{f_0}(\epsilon) = 0$.

### B. Upper Bound

The following theorem is our main result that provides an upper bound on the local minimax rate of noisy global optimization with active queries.

**Theorem 1.** *For any $\alpha, M, \kappa, c_0, C_0 > 0$ and $f_0 \in \Sigma_\kappa^\alpha(M) \cap \Theta_{\mathbf{C}}$, where $\mathbf{C} = (c_0, C_0)$, define*

$$\varepsilon_n^{\mathsf{U}}(f_0) := \sup\left\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n/\log^\omega n\right\}, \quad (7)$$

*where $\omega > 5 + d/\alpha$ is a large constant. Suppose also that $\varepsilon_n^{\mathsf{U}}(f_0) \to 0$ as $n \to \infty$. Then for sufficiently large $n$, there exists an estimator $\widehat{x}_n$ with access to $n$ active queries $x_1, \ldots, x_n \in \mathcal{X}$, a constant $C_R > 0$ depending only on $\alpha, M, \kappa, c, c_0, C_0$ and a constant $\gamma > 0$ depending only on $\alpha$ and $d$ such that*

$$\sup_{f_0 \in \Sigma_\kappa^\alpha(M) \cap \Theta_{\mathbf{C}}} \ \sup_{\substack{f \in \Sigma_\kappa^\alpha(M), \\ \|f - f_0\|_\infty \leqslant \varepsilon_n^{\mathsf{U}}(f_0)}} \Pr_f\Big[\mathfrak{L}(\widehat{x}_n, f) > C_R \log^\gamma n \cdot (\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2})\Big] \leqslant 1/4. \quad (8)$$

**Remark 1.** *Unlike the (local) smoothness class $\Sigma_\kappa^\alpha(M)$, the additional function class $\Theta_{\mathbf{C}}$ that encapsulates (A2) is imposed only on the "reference" function $f_0$ but not the true function $f$ to be estimated. This makes the assumptions considerably weaker because the true function $f$ may violate (A2) while our results remain valid.*

**Remark 2.** *The estimator $\widehat{x}_n$ does* not *require knowledge of parameters $\kappa, c_0, C_0$ or $\varepsilon_n^{\mathsf{U}}(f_0)$, and automatically adapts to them, as shown in the next section. While the knowledge of smoothness parameters $\alpha$ and $M$ is in general unavoidable in non-parametric regression (see [48]), in the zeroth-order optimization problem it is possible to adapt to $\alpha$ and $M$ by running $O(\log^2 n)$ parallel sessions of $\widehat{x}_n$ on $O(\log n)$ grids of $\alpha$ and $M$ values, and then using $\Omega(n/\log^2 n)$ single-point queries to decide on the location with the smallest function value. This adaptive strategy was suggested in [22] to remove an additional condition in [21], and also applies to our setting.*

**Remark 3.** *When the distribution function $\mu_{f_0}(\epsilon)$ does not change abruptly with $\epsilon$ the expression of $\varepsilon_n^{\mathsf{U}}(f_0)$ can be significantly simplified. In particular, if for all $\epsilon \in (0, c_0]$ it holds that*

$$\mu_{f_0}(\epsilon/\log n) \geqslant \mu_{f_0}(\epsilon)/[\log n]^{O(1)}, \quad (9)$$

*then $\varepsilon_n^{\mathsf{U}}(f_0)$ can be upper bounded as*

$$\varepsilon_n^{\mathsf{U}}(f_0) \leqslant [\log n]^{O(1)} \cdot \sup\left\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n\right\}. \quad (10)$$

*If $\mu_{f_0}(\epsilon)$ scales polynomially with $\epsilon$, i.e. $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ for some constant $\beta \geqslant 0$, then* (9) *and* (10) *are both satisfied.*

The quantity $\varepsilon_n^{\mathsf{U}}(f_0) = \sup\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n/\log^\omega n\}$ is crucial in determining the convergence rate of optimization error of $\widehat{x}_n$ *locally* around the reference function $f_0$. While the definition of $\varepsilon_n^{\mathsf{U}}(f_0)$ is mostly implicit and involves solving an inequality involving the distribution function $\mu_{f_0}(\cdot)$, we remark that it admits a simple form when $\mu_{f_0}$ has a polynomial growth rate similar to a local Tsybakov noise condition [4], [49], as shown in the following proposition:

**Proposition 2.** *Suppose $\mu_{f_0}(\epsilon) \lesssim \epsilon^\beta$ for some constant $\beta \in [0, 2 + d/\alpha)$. Then $\varepsilon_n^{\mathsf{U}}(f_0) = \widetilde{O}(n^{-\alpha/(2\alpha+d-\alpha\beta)})$. In addition, if $\beta \in [0, d/\alpha]$ then $\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2} \lesssim \varepsilon_n^{\mathsf{U}}(f_0) = \widetilde{O}(n^{-\alpha/(2\alpha+d-\alpha\beta)})$.*

We remark that, following Proposition 1 of [45], $\alpha$, $\beta$ and $d$ must satisfy the relationship that $\beta \leqslant d/\alpha$. Proposition 2 can be easily verified by solving the system $\varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n/\log^{\omega} n$ with the condition $\mu_{f_0}(\epsilon) \lesssim \epsilon^{\beta}$. We therefore omit its proof. The following two examples give some simple reference functions $f_0$ that satisfy the $\mu_{f_0}(\epsilon) \lesssim \epsilon^{\beta}$ condition in Proposition 2 with particular values of $\beta$.

**Example 1.** *The constant function $f_0 \equiv 0$ satisfies (A1) through (A3) with $\beta = 0$.*

**Example 2.** *$f_0 \in \Sigma_{\kappa}^{2}(M)$ that is* strongly convex[4] *satisfies (A1) through (A3) with $\beta = d/2$.*

Example 1 is simple to verify, as the volume of level-sets of the constant function $f_0 \equiv 0$ exhibit a phase transition at $\epsilon = 0$ and $\epsilon > 0$. Consequently, $\beta = 0$ is the only parameter for which $\mu_{f_0}(\epsilon) \lesssim \epsilon^{\beta}$. Example 2 is more involved, and holds because the strong convexity of $f_0$ *lower bounds* the growth rate of $f_0$ when moving away from its minimum. We give a rigorous proof for Example 2 in the appendix. We also remark that $f_0$ does *not* need to be exactly strongly convex for $\beta = d/2$ to hold, and the example is valid for, e.g., piecewise strongly convex functions with a constant number of pieces too.

To best interpret the results in Theorem 1 and Proposition 2, it is instructive to compare the "local" rate $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ with the baseline rate $n^{-\alpha/(2\alpha+d)}$, which can be attained by reconstructing $f$ in sup-norm and applying Proposition 1. Since $\beta \geqslant 0$, the local convergence rate established in Theorem 1 is never slower, and the improvement compared to the baseline rate $n^{-\alpha/(2\alpha+d)}$ is dictated by $\beta$, which governs the growth rate of volume of level-sets of the reference function $f_0$. In particular, for functions that grows fast when moving away from its minimum, the parameter $\beta$ is large and therefore the local convergence rate around $f_0$ could be much faster than $n^{-\alpha/(2\alpha+d)}$.

Theorem 1 also implies concrete convergence rates for special functions considered in Examples 1 and 2. For the constant reference function $f_0 \equiv 0$, Example 1 and Theorem 1 yield that $R_n(f_0) \asymp n^{-\alpha/(2\alpha+d)}$, which matches the baseline rate $n^{-\alpha/(2\alpha+d)}$ and suggests that $f_0 \equiv 0$ is the worst-case reference function. This is intuitive, because $f_0 \equiv 0$ has a drastic level-set change at $\epsilon \to 0^+$ and therefore small perturbations of $f_0$ result in changes to the optimal location. On the other hand, if $f_0$ is strongly smooth and convex as in Example 2, Theorem 1 leads to the bound of $R_n(f_0) \asymp n^{-1/2}$, which is significantly better than the $n^{-2/(4+d)}$ baseline rate[5] and also matches existing works on zeroth-order optimization of convex functions [11]. The faster rate holds intuitively because strongly convex functions grow quickly when moving away from the minimum. An active query algorithm can focus most of its queries on the small level-sets of the underlying function, resulting in more accurate local function reconstruction and faster optimization error rate.

Our proof of Theorem 1 is constructive, by upper bounding the local minimax optimization error of an explicit algorithm. Roughly, our algorithm partitions the $n$ active queries evenly into $\log n$ epochs, and level-sets of $f$ are estimated at the end of each epoch by comparing (uniform) confidence intervals on a dense grid on $\mathcal{X}$. It is then proved that the volume of the estimated level-sets contracts *geometrically*, until the target convergence rate $R_n(f_0)$ is attained. The algorithm is described in more detail in Section IV and the complete proof of Theorem 1 is in Section V-B.

### C. Lower Bounds

We prove local minimax lower bounds that match the upper bounds in Theorem 1 up to logarithmic terms. As we remarked in Section II-B, in the local minimax lower bound formulation we assume the data analyst has full knowledge of the reference function $f_0$, which makes the lower bounds stronger as more information is available a priori.

To facilitate such local minimax lower bounds, the following additional condition is imposed on the reference function $f_0$ of which the data analyst has perfect information.

(A2') There exist constants $c_0', C_0' > 0$ such that $M(L_{f_0}(\epsilon), \delta) \geqslant C_0' \mu_{f_0}(\epsilon)\delta^{-d}$ for all $\epsilon, \delta \in (0, c_0']$, where $M(L_{f_0}(\epsilon), \delta)$ is the maximum number of disjoint $\ell_2$ balls of radius $\delta$ that can be packed into $L_{f_0}(\epsilon)$.

We denote $\Theta'_{\mathbf{C}'}$ as the class of functions that satisfy (A2') with respect to parameters $\mathbf{C}' = (c_0', C_0') > 0$. Intuitively, (A2') can be regarded as a converse of (A2).

We are now ready to state our main negative result, which shows, from an information-theoretic perspective, that the upper bound in Theorem 1 is not improvable.

**Theorem 2.** *Suppose $\alpha, c_0, C_0, c_0', C_0' > 0$ and $\kappa = \infty$. Denote $\mathbf{C} = (c_0, C_0)$ and $\mathbf{C}' = (c_0', C_0')$. For any $f_0 \in \Theta_{\mathbf{C}} \cap \Theta'_{\mathbf{C}'}$, define*

$$\varepsilon_n^{\mathsf{L}}(f_0) := \sup\left\{\varepsilon > 0 : \varepsilon^{-(2+d/\alpha)}\mu_{f_0}(\varepsilon) \geqslant n\right\}. \tag{11}$$

*Then there exists a constant $M > 0$ depending on $\alpha, d, \mathbf{C}$ and $\mathbf{C}'$ such that, for any $f_0 \in \Sigma_{\kappa}^{\alpha}(M/2) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$,*

$$\inf_{\widehat{x}_n} \sup_{\substack{f \in \Sigma_{\kappa}^{\alpha}(M), \\ \|f-f_0\|_\infty \leqslant 2\varepsilon_n^{\mathsf{L}}(f_0)}} \Pr_f\left[\mathfrak{L}(\widehat{x}_n; f) \geqslant \varepsilon_n^{\mathsf{L}}(f_0)\right] \geqslant \frac{1}{3}. \tag{12}$$

**Remark 4.** *We note in passing that for any $f_0$ and $n$ it always holds that $\varepsilon_n^{\mathsf{L}}(f_0) \leqslant \varepsilon_n^{\mathsf{U}}(f_0)$.*

**Remark 5.** *If the distribution function $\mu_{f_0}(\epsilon)$ satisfies* (9) *(i.e. it does not change too abruptly) in Remark 3, then $\varepsilon_n^{\mathsf{L}}(f_0) \geqslant \varepsilon_n^{\mathsf{U}}(f_0)/[\log n]^{O(1)}$. Consequently, the upper and lower bounds for these functions match up to logarithmic factors.*

The following proposition derives an explicit expression for $\varepsilon_n^{\mathsf{L}}(f_0)$ for reference functions whose distribution functions have a polynomial growth, which matches the upper bound in Proposition 2 up to $\log n$ factors. The proof of this Proposition is straightforward and is omitted.

**Proposition 3.** *Suppose $\mu_{f_0}(\epsilon) \gtrsim \epsilon^{\beta}$ for some $\beta \in [0, 2 + d/\alpha)$. Then $\varepsilon_n^{\mathsf{L}}(f_0) = \Omega(n^{-\alpha/(2\alpha+d-\alpha\beta)})$.*

[4]A twice differentiable function $f_0$ is strongly convex if there exists $\sigma > 0$ such that $\nabla^2 f_0(x) \succeq \sigma I, \forall x \in \mathcal{X}$.

[5]Note that $f_0$ being strongly smooth corresponds to $\alpha = 2$ in the local smoothness assumption.

The following proposition additionally shows the existence of $f_0 \in \Sigma_\infty^\alpha(M) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$ that satisfies $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ for any values of $\alpha > 0$ and $\beta \in [0, d/\alpha]$. Its proof is given in the Appendix.

**Proposition 4.** *Fix arbitrary $\alpha, M > 0$ and $\beta \in [0, d/\alpha]$. There exists $f_0 \in \Sigma_\kappa^\alpha(M) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$ for $\kappa = \infty$ and constants $\mathbf{C} = (c_0, C_0)$, $\mathbf{C}' = (c_0', C_0')$ that depend only on $\alpha, \beta, M$ and $d$ such that $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$.*

Theorem 2 and Proposition 3 show that the $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ upper bound on local minimax convergence rate established in Theorem 1 is not improvable up to logarithmic factors of $n$. Such information-theoretic lower bounds on the convergence rates hold *even if the data analyst has perfect information of* $f_0$, the reference function on which the $n^{-\alpha/(2\alpha+d-\alpha\beta)}$ local rate is based. Our results also imply an $n^{-\alpha/(2\alpha+d)}$ minimax lower bound over all $\alpha$-Hölder smooth functions, showing that without additional assumptions, noisy optimization of smooth functions is as difficult as reconstructing the unknown function in sup-norm.

Our proof of Theorem 2 also differs from those of existing minimax lower bounds for active nonparametric models [50]. The classical approach is to invoke Fano's inequality and to upper bound the KL divergence between different underlying functions $f$ and $g$ using $\|f - g\|_\infty$, corresponding to the point $x \in \mathcal{X}$ that leads to the largest KL divergence. Such an approach, however, does not produce tight lower bounds for our problem. To overcome such difficulties, we borrow the lower bound analysis for bandit pure exploration problems in [51]. In particular, our analysis considers the query distribution of any active query algorithm $\mathcal{A} = (\varphi_1, \ldots, \varphi_n, \phi_n)$ under the reference function $f_0$ and bounds the perturbation in query distributions between $f_0$ and $f$ using Le Cam's lemma. Afterwards, an adversarial function choice $f$ can be made based on the query distributions of the considered algorithm $\mathcal{A}$. We defer the complete proof of Theorem 2 to Section V-C.

Theorem 2 applies to any global optimization method that makes *active* queries, corresponding to the query model in Definition 2. The following theorem, on the other hand, shows that for passive algorithms (Definition 1) the $n^{-\alpha/(2\alpha+d)}$ optimization rate is not improvable even with additional level-set assumptions imposed on $f_0$. This demonstrates an explicit gap between passive and adaptive query models in global optimization problems.

**Theorem 3.** *Suppose $\alpha, c_0, C_0, c_0', C_0' > 0$ and $\kappa = \infty$. Denote $\mathbf{C} = (c_0, C_0)$ and $\mathbf{C}' = (c_0', C_0')$. Then there exist constants $M > 0$ depending on $\alpha, d, \mathbf{C}, \mathbf{C}'$ and $N$ depending on $M$ such that, for any $f_0 \in \Sigma_\kappa^\alpha(M/2) \cap \Theta_{\mathbf{C}} \cap \Theta_{\mathbf{C}'}$ satisfying $\varepsilon_n^{\mathsf{L}}(f_0) \leqslant \widetilde{\varepsilon}_n^{\mathsf{L}} =: [\log n/n]^{\alpha/(2\alpha+d)}$,*

$$\inf_{\check{x}_n} \sup_{\substack{f \in \Sigma_\kappa^\alpha(M), \\ \|f - f_0\|_\infty \leqslant 2\widetilde{\varepsilon}_n^{\mathsf{L}}}} \Pr_f\left[\mathfrak{L}(\widehat{x}_n; f) \geqslant \widetilde{\varepsilon}_n^{\mathsf{L}}\right] \geqslant \frac{1}{3} \quad \textit{for all } n \geqslant N. \tag{13}$$

Intuitively, the apparent gap demonstrated by Theorems 2 and 3 between the active and passive query models stems from the observation that, a passive algorithm $\mathcal{A}$ only has access to uniformly sampled query points $x_1, \ldots, x_n$ and therefore cannot focus on a small level-set of $f$ in order to improve query efficiency. In addition, for functions that grow faster when moving away from their minima (implying a larger value of $\beta$), the gap between passive and active query models becomes bigger as active queries can more effectively exploit the restricted level-sets of such functions.

## IV. Our Algorithm

In this section we describe a concrete algorithm that attains the upper bound in Theorem 1. We start with a cleaner algorithm that operates under the slightly stronger condition that $\kappa = \infty$ in (A1), meaning that $f$ is $\alpha$-Hölder smooth on the entire domain $\mathcal{X}$. The generalization to $\kappa > 0$ being a constant is given in Section IV-C with an additional pre-processing step.

Let $G_n \in \mathcal{X}$ be a *finite* grid of points in $\mathcal{X}$. We assume the finite grid $G_n$ satisfies the following two mild conditions:

(B1) Points in $G_n$ are sampled i.i.d. from an unknown distribution $P_X$ on $\mathcal{X}$; furthermore, the density $p_X$ associated with $P_X$ satisfies $\underline{p}_0 \leqslant p_X(x) \leqslant \overline{p}_0$ for all $x \in \mathcal{X}$, where $0 < \underline{p}_0 \leqslant \overline{p}_0 < \infty$ are universal constants;

(B2) $|G_n| \gtrsim n^3$ and $\log|G_n| = O(\log n)$.

**Remark 6.** *Although typically the choices of the grid points $G_n$ belong to the data analyst, in some applications the choices of design points are not completely unconstrained. For example, in material synthesis experiments described previously some environmental parameter settings (e.g., temperature and pressure) might not be allowed due to budget or physical constraints. Thus, we choose to consider less restrictive conditions imposed on the design grid $G_n$, allowing it to be more flexible in real-world applications.*

**Remark 7.** *Condition (B2) ensures that the grid $G_n$ is sufficiently dense, such that even with the smallest bandwidth our algorithm possibly uses ($h_t(x) = 1/n^2$, see (18)), each $x \in G_n$ has abundant neighboring points in $G_n$, so that the local polynomial estimates in* (15) *are well-defined.*

For any subset $S \subseteq G_n$ and a "weight" function $\varrho : G_n \to \mathbb{R}^+$, define the *extension* $S^\circ(\varrho)$ of $S$ with respect to $\varrho$ as

$$S^\circ(\varrho) := \bigcup_{x \in S} B^\infty_{\varrho(x)}(x; G_n) \quad \text{where}$$
$$B^\infty_{\varrho(x)}(x; G_n) = \{z \in G_n : \|z - x\|_\infty \leqslant \varrho(x)\}. \tag{14}$$

The algorithm can then be formulated as two levels of iterations, with the outer loop shrinking the "active set" $S_\tau$ and the inner loop collecting data in order to reduce the lengths of the confidence intervals on the points in the active set. A pseudocode description of our proposed algorithm is given in Figure 1.

### A. Local Polynomial Regression

We use local polynomial regression [5] to obtain the estimate $\widehat{f}$. In particular, for any $x \in G_n$ and a bandwidth

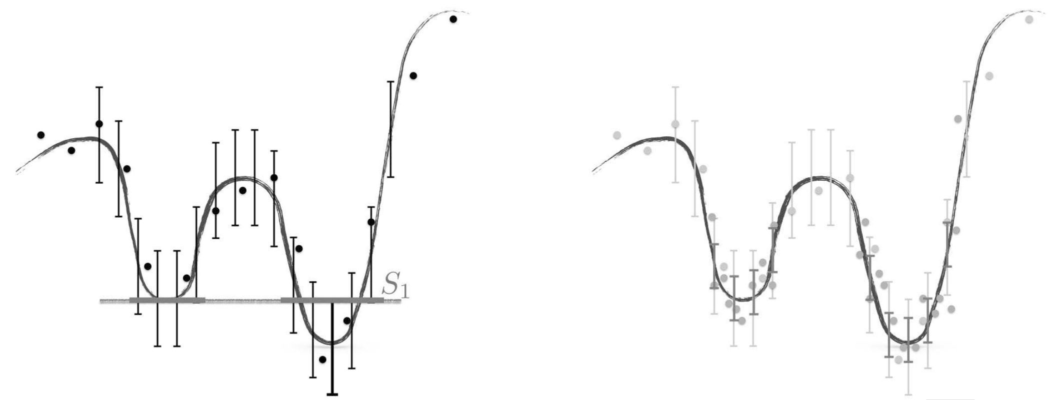


Fig. 1. An informal illustration of Algorithm 1. Solid blue curves depict the underlying function $f$ to be optimized, black and red solid dots denote the query points and their responses $\{(x_t, y_t)\}$, and black and red vertical line segments correspond to uniform confidence intervals on function evaluations constructed using the current batch of data observed. The left figure illustrates the first epoch of our algorithm, where query points are uniformly sampled from the entire domain $\mathcal{X}$. Afterwards, sub-optimal locations based on the constructed confidence intervals are removed, and a shrunken "candidate set" $S_1$ is obtained. The algorithm then proceeds to the second epoch, illustrated in the right figure, where query points (in red) are sampled only from the restricted candidate set and shorter confidence intervals (also in red) are constructed and updated. The procedure is repeated until $O(\log n)$ epochs are completed.

**Parameters**: $\alpha$, $M$, $\delta$, $n$
**Output**: $\widehat{x}_n$, the final prediction
Initialization: $S_0 = G_n$, $\varrho_0(x) \equiv \infty$, $T = \lfloor \log_2 n \rfloor$, $n_0 = \lfloor n/T \rfloor$;
**for** $\tau = 1, 2, \ldots, T$ **do**
  Compute "extended" sample set $S^{\circ}_{\tau-1}(\varrho_{\tau-1})$ defined in (14);
  **for** $t = (\tau-1)n_0 + 1$ *to* $\tau n_0$ **do**
    Sample $x_t$ uniformly at random from $S^{\circ}_{\tau-1}(\varrho_{\tau-1})$ and observe $y_t = f(x_t) + w_t$;
  **end**
  For every $x \in S_{\tau-1}$, compute bandwidth $h_\tau(x)$ using (18) and build the confidence interval $[\ell_\tau(x), u_\tau(x)]$ in (19);
  $S_\tau := \{x \in S_{\tau-1} : \ell_\tau(x) \leqslant \min_{x' \in S_{\tau-1}} u_\tau(x')\}$,
  $\varrho_\tau(x) := \min\{\varrho_{\tau-1}(x), h_\tau(x)\}$,
**end**
Final processing: for every $x \in S_T$ let $\widehat{f}_{h_T,x}(\cdot)$ be the local polynomial estimates constructed in (15) at $x$.
Output $\widehat{x}_n = \arg\min_{x \in S_T} \min_{x' \in B^{\infty}_{h_T(x)}(x;\mathcal{X})} \widehat{f}_{h_T,x}(x')$.

**Algorithm 1** The Main Algorithm

parameter $h > 0$, consider the least squares polynomial estimate

$$\widehat{f}_h \in \operatorname*{argmin}_{g \in \mathcal{P}_k} \sum_{t'=1}^{t} \mathbb{I}[x_{t'} \in B^{\infty}_h(x)] \cdot (y_{t'} - g(x_{t'}))^2, \qquad (15)$$

where $B^{\infty}_h(x) := \{x' \in \mathcal{X} : \|x' - x\|_\infty \leqslant h\}$ and $\mathcal{P}_k$ denotes all polynomials of degree $k$ on $\mathcal{X}$.

To analyze the performance of $\widehat{f}_h$ evaluated at a certain point $x \in \mathcal{X}$, define the mapping

$$\psi_{x,h} : z \mapsto (1, \psi^1_{x,h}(z), \ldots, \psi^k_{x,h}(z))$$

where $\psi^j_{x,h} : z \mapsto [\prod_{\ell=1}^{j} h^{-1}(z_{i_\ell} - x_{i_\ell})]^d_{i_1,\ldots,i_j=1}$ is the degree-$j$ polynomial mapping from $\mathbb{R}^d$ to $\mathbb{R}^{d^j}$. Also define $\Psi_{t,h} := (\psi_{x,h}(x_{t'}))_{1 \leqslant t' \leqslant t, x_{t'} \in B_h(x)}$ as the $m \times D$ aggregated design matrix, where $m = \sum_{t'=1}^{t} \mathbb{I}[x_{t'} \in B^{\infty}_h(x)]$ and $D = 1 + d + \ldots + d^k$, $k = \lfloor \alpha \rfloor$. The estimate $\widehat{f}_h$ defined in (15) then admits the following closed-form expression:

$$\widehat{f}_h(z) \equiv \psi_{x,h}(z)^\top (\Psi^\top_{t,h}\Psi_{t,h})^\dagger \Psi^\top_{t,h} Y_{t,h}, \qquad (16)$$

where $Y_{t,h} = (y_{t'})_{1 \leqslant t' \leqslant t, x_{t'} \in B^{\infty}_h(x)}$ and $A^\dagger$ is the Moore-Penrose pseudo-inverse of $A$.

The following lemma gives a finite-sample analysis of the error of $\widehat{f}_h(x)$:

**Lemma 1.** *Suppose that $f$ satisfies (6) on $B^{\infty}_h(x;\mathcal{X})$, $\max_{z \in B^{\infty}_h(x;\mathcal{X})} \|\psi_{x,h}(z)\|_2 \leqslant b$ and $\frac{1}{m}\Psi^\top_{t,h}\Psi_{t,h} \succeq \sigma I_{D\times D}$ for some $\sigma > 0$. Then for any $\delta \in (0, 1/2)$, with probability $1-\delta$*

$$\left|\widehat{f}_h(x') - f(x')\right| \leqslant \underbrace{\frac{b^2}{\sigma} M d^k h^\alpha}_{\mathfrak{b}_{h,\delta}(x)} + \underbrace{b\sqrt{\frac{5D\ln(1/\delta)}{\sigma m}}}_{\mathfrak{s}_{h,\delta}(x)} =: \eta_{h,\delta}(x), \quad \forall x' \in B^{\infty}_h(x;\mathcal{X}). \qquad (17)$$

**Remark 8.** *$\mathfrak{b}_{h,\delta}(x)$, $\mathfrak{s}_{h,\delta}(x)$ and $\eta_{h,\delta}(x)$ depend on $x$ because $\sigma$ depends on $\Psi_{t,h}$, which further depends on the sample points in the neighborhood $B^{\infty}_h(x;\mathcal{X})$ of $x$.*

In the rest of the paper we define $\mathfrak{b}_{h,\delta}(x) := (b^2/\sigma)Md^k h^\alpha$ and $\mathfrak{s}_{h,\delta}(x) := b\sqrt{5D\ln(1/\delta)/\sigma m}$ as the bias and standard deviation terms in the error of $\widehat{f}_h(x)$, respectively. We also denote $\eta_{h,\delta}(x) := \mathfrak{b}_{h,\delta}(x) + \mathfrak{s}_{h,\delta}(x)$ as the overall error in $\widehat{f}_h(x)$.

Notice that when bandwidth $h$ increases, the bias term $\mathfrak{b}_{h,\delta}(x)$ increases because of the $h^\alpha$ term; on the other hand, with $h$ increasing the local neighborhood $B^{\infty}_h(x;\mathcal{X})$ grows and would potentially contain more samples, implying a larger $m$ and smaller standard deviation term $\mathfrak{s}_{h,\delta}(x)$. A careful selection of the bandwidth $h$ balances $\mathfrak{b}_{h,\delta}(x)$ and $\mathfrak{s}_{h,\delta}(x)$ and yields appropriate confidence intervals on $f(x)$, and we turn our attention to this in the next section.

### B. Bandwidth Selection and Confidence Intervals

Given the expressions of bias $\mathfrak{b}_{h,\delta}(x)$ and standard deviation $\mathfrak{s}_{h,\delta}(x)$ in (17), the bandwidth $h_\tau(x) > 0$ at epoch $\tau$ and point $x$ is selected as

$$h_\tau(x) := \frac{j_\tau(x)}{n^2} \quad \text{where} \quad j_\tau(x) := \arg\max\big\{ j \in \mathbb{N}^+, \; j \leqslant n^2 : \mathfrak{b}_{j/n^2,\delta}(x) \leqslant \mathfrak{s}_{j/n^2,\delta}(x)\big\}. \tag{18}$$

More specifically, $h_\tau(x)$ is the largest positive value in an evenly spaced grid $\{j/n^2\}$ such that the bias of $\widehat{f}_{h_\tau}(x)$ is smaller than its standard deviation. This bandwidth selection is in principle similar to the Lepski's method [52], with the exception that an upper bound on the bias for any bandwidth parameter is known and does not need to be estimated from data.

With the selection of bandwidth $h_\tau(x)$ at epoch $\tau$ and query point $x$, a confidence interval on $f(x')$ for all $x' \in B^{\infty}_{h_\tau(x)}(x; \mathcal{X})$ is constructed as

$$\ell_\tau(x) := \max_{1\leqslant t'\leqslant \tau} \sup_{x'\in B^{\infty}_{h_\tau(x)}(x;\mathcal{X})} \Big\{ \widehat{f}_{h_{t'}(x)}(x') - \eta_{h_{t'}(x),\delta}(x) \Big\};$$
$$u_\tau(x) := \min_{1\leqslant t'\leqslant \tau} \inf_{x'\in B^{\infty}_{h_\tau(x)}(x;\mathcal{X})} \Big\{ \widehat{f}_{h_{t'}(x)}(x') + \eta_{h_{t'}(x),\delta}(x) \Big\}. \tag{19}$$

Note that for any $x \in \mathcal{X}$, the lower confidence edge $\ell_\tau(x)$ is a non-decreasing function in $\tau$ and the upper confidence edge $u_\tau(x)$ is a non-increasing function in $\tau$.

### C. Pre-processing

We describe a pre-processing step that relaxes the smoothness condition from $\kappa = \infty$ to $\kappa = \Omega(1)$, meaning that only local smoothness of $f$ around its minimum values is required. Let $n_0 = \lfloor n/\log n \rfloor$, $x_1, \ldots, x_{n_0}$ be points i.i.d. uniformly sampled from $\mathcal{X}$ and $y_1, \ldots, y_{n_0}$ be their corresponding responses. For every grid point $x \in G_n$, perform the following:

1) Compute $\breve{f}_x(\cdot)$ as the local polynomial fits of all $y_i$ corresponding to $\|x_i - x\|_\infty \leqslant n_0^{-1/2d} \log^3 n =: h_0$;
2) Compute $\overline{f}(x)$ as the sample average of all $y_i$ corresponding to $\|x_i - x\|_\infty \leqslant h_0$;
3) Remove all $x \in G_n$ from $S_0$ if $\overline{f}(x) \geqslant \min_{z\in G_n} \inf_{z'\in B^{\infty}_{h_0}(z;\mathcal{X})} \breve{f}_z(z') + 1/\log n$.

**Remark 9.** *The $1/\log n$ term in the removal condition $\breve{f}(x) \geqslant \min_{z\in G_n} \breve{f}(z) + 1/\log n$ is not important, and can be replaced with any sequence $\{\omega_n\}$ such that $\lim_{n\to\infty} \omega_n = 0$ and $\lim_{n\to\infty} \omega_n n^t = \infty$ for any $t > 0$. The readers are referred to the proof of Proposition 5 in the appendix for the motivation of this term as well as the selection of the pre-processing bandwidth $h_0$.*

To analyze the pre-processing step, we state the following proposition:

**Proposition 5.** *Assume $f \in \Sigma^{\alpha}_{\kappa}(M)$ and let $S_0'$ be the screened grid after step 2 of the pre-processing procedure. Then for sufficiently large $n$, with probability $1 - O(n^{-1})$ we have*

$$B^{\infty}_{h_0}(x; \mathcal{X}) \cap L_f(\kappa/2) \neq \varnothing, \qquad \forall x \in S_0', \tag{20}$$

*where $L_f(\kappa/2) = \{x \in \mathcal{X} : f(x) \leqslant f^* + \kappa/2\}$.*

To interpret Proposition 5, note that for sufficiently large $n$, $f \in \Sigma^{\alpha}_{\kappa}(M)$ implies $f$ being $\alpha$-Hölder smooth (i.e., $f$ satisfies (6)) on $\bigcup_{x\in L_f(\kappa/2)} B^{\infty}_{h_0}(x; \mathcal{X})$, because $\kappa > 0$ is a constant and $h_0 \to 0$ as $n \to \infty$. Subsequently, the proposition shows that with high probability, the pre-processing step will remove all grid points in $G_n$ in non-smooth regions of $f$, while maintaining the global optimal solution. This justifies the pre-processing step for $f \in \Sigma^{\alpha}_{\kappa}(M)$, because $f$ is smooth on the grid and its close neighborhood after pre-processing.

The proof of Proposition 5 uses the fact that the local mean estimation is large provided that all data points in the local mean estimator are large, regardless of their underlying smoothness. The complete proof of Proposition 5 is deferred to the Appendix.

## V. Proofs of Main Theorems

### A. Proof of Lemma 1

Our proof closely follows the analysis of asymptotic convergence rates for series estimators in [53]. We further work out all constants in the error bounds to arrive at a completely finite-sample result, which is then used to construct finite-sample confidence intervals.

We start with as polynomial interpolation results for all Hölder smooth functions in $B^{\infty}_{h_t}(x; \mathcal{X})$.

**Lemma 2.** *Suppose $f$ satisfies (6) on $B^{\infty}_{h}(x; \mathcal{X})$. Then there exists $\widetilde{f}_x \in \mathcal{P}_k$ such that*

$$\sup_{z\in B^{\infty}_{h}(x;\mathcal{X})} \big|f(z) - \widetilde{f}_x(z)\big| \leqslant M d^k h^\alpha. \tag{21}$$

*Proof.* Consider

$$\widetilde{f}_x(z) := f(x) + \sum_{j=1}^{k} \sum_{\alpha_1+\ldots+\alpha_d=j} \frac{\partial^j f(x)}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} \prod_{\ell=1}^{d} (z_\ell - x_\ell)^{\alpha_\ell}. \tag{22}$$

By Taylor expansion with Lagrangian remainders, there exists $\xi \in (0, 1)$ such that

$$\big|\widetilde{f}_x(z) - f(z)\big| \leqslant \sum_{\alpha_1+\ldots+\alpha_d=k} \big|f^{(\boldsymbol{\alpha})}(x+\xi(z-x)) - f^{(\boldsymbol{\alpha})}(x)\big| \cdot \prod_{\ell=1}^{d} |z_\ell - x_\ell|^{\alpha_\ell}.$$

Because $f$ satisfies (6) on $B^{\infty}_{h}(x; \mathcal{X})$, we have that $|f^{(\boldsymbol{\alpha})}(x + \xi(z-x)) - f^{(\boldsymbol{\alpha})}(x)| \leqslant M \cdot \|z-x\|_\infty^{\alpha-k}$. Also note that $|z_\ell - x_\ell| \leqslant \|z - x\|_\infty \leqslant h$ for all $z \in B^{\infty}_{h}(x; \mathcal{X})$. The lemma is thus proved. □

Using (16), the local polynomial estimate $\widehat{f}_h$ can be written as $\widehat{f}_h(z) \equiv \psi_{x,h}(z)^\top \widehat{\theta}_h$, where

$$\widehat{\theta}_h = (\Psi_{t,h}^\top \Psi_{t,h})^{-1} \Psi_{t,h}^\top Y_{t,h}. \tag{23}$$

In addition, because $\widetilde{f}_x \in \mathcal{P}_k$, there exists $\widetilde{\theta} \in \mathbb{R}^D$ such that $\widetilde{f}_x(z) \equiv \psi_{x,h}(z)^\top \widetilde{\theta}$. Denote also that $F_{t,h} := (f(x_{t'}))_{1\leqslant t'\leqslant t, x_{t'}\in B^{\infty}_{h}(x)}$, $\Delta_{t,h} := (f(x_{t'}) - \widetilde{f}_x(x_{t'}))$

$_{1\leqslant t'\leqslant t, x_{t'}\in B_h^\infty(x)}$ and $W_{t,h} := (w_{t'})_{1\leqslant t'\leqslant t, x_{t'}\in B_h^\infty(x)}$. (23) can then be re-formulated as

$$\widehat{\theta}_h = (\Psi_{t,h}^\top \Psi_{t,h})^{-1}\Psi_{t,h}^\top\left[\Psi_{t,h}\widetilde{\theta} + \Delta_{t,h} + W_{t,h}\right] \tag{24}$$

$$= \widetilde{\theta} + \left[\frac{1}{m}\Psi_{t,h}^\top \Psi_{t,h}\right]^{-1}\left[\frac{1}{m}\Psi_{t,h}^\top(\Delta_{t,h} + W_{t,h})\right]. \tag{25}$$

Because $\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h} \succeq \sigma I_{D\times D}$ and $\sup_{z\in B_h^\infty(x)}\|\psi_{x,h}(z)\|_2 \leqslant b$, we have that

$$\|\widehat{\theta}_h - \widetilde{\theta}\|_2 \leqslant \frac{b}{\sigma}\|\Delta_{t,h}\|_\infty + \left\|\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\frac{1}{m}\Psi_{t,h}^\top W_t\right\|_2. \tag{26}$$

Invoking Lemma 2 we have $\|\Delta_{t,h}\|_\infty \leqslant Md^k h^\alpha$. In addition, because $W_t \sim \mathcal{N}_m(0, I_{m\times n})$, we have that

$$\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\frac{1}{m}\Psi_{t,h}^\top W_t \sim \mathcal{N}_D\left(0, \frac{1}{m}\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\right). \tag{27}$$

Applying concentration inequalities for quadratic forms of Gaussian random vectors (Lemma 10), with probability $1-\delta$ it holds that

$$\left\|\left[\frac{1}{m}\Psi_{t,h}^\top\Psi_{t,h}\right]^{-1}\frac{1}{m}\Psi_{t,h}^\top W_t\right\|_2 \leqslant \sqrt{\frac{5D\log(1/\delta)}{\sigma m}}. \tag{28}$$

We then have that with probability $1-\delta$ that

$$\|\widehat{\theta}_h - \widetilde{\theta}\|_2 \leqslant \frac{b}{\sigma_h}Md^k h_t^\alpha + \sqrt{\frac{5D\log(1/\delta)}{\sigma m}}. \tag{29}$$

Finally, noting that for all $x' \in B_h^\infty(x;\mathcal{X})$, $\|\psi_{x,h}(x')\|_2 \leqslant b$ by definition, we have that

$$|\widehat{f}_h(x') - f(x')| = |\widehat{f}_h('x) - \widetilde{f}_x(x')|$$
$$= |\psi_{x,h}(x')^\top(\widehat{\theta}_h - \widetilde{\theta})| \leqslant b\|\widehat{\theta}_h - \widetilde{\theta}\|_2,$$

which completes the proof of Lemma 1.

### *B. Proof of Theorem 1*

In this section we prove Theorem 1. We prove the theorem by considering every reference function $f_0 \in \Sigma_\kappa^\alpha(M)\cap\Theta_{\mathbf{C}}$ separately. For simplicity, we assume $\kappa = \infty$ throughout the proof. The $0 < \kappa < \infty$ can be handled by replacing $\mathcal{X}$ with $S_0$ which is the grid after the pre-processing step described in Section IV-C. We also suppress dependency on $d, \alpha, M, \mathbf{C}, \underline{p}_0, \overline{p}_0$ in $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, $\gtrsim$, $\lesssim$ and $\asymp$ notations. We further suppress logarithmic terms of $n$ in $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ notations.

The following lemma is our main lemma, which shows that the active set $S_\tau$ in our proposed algorithm shrinks geometrically before it reaches a certain level. To simplify notations, denote $\widetilde{c}_0 := 10c_0$ and (A2) then hold for all $\epsilon, \delta \in [0, \widetilde{c}_0]$ for all $f_0 \in \Theta_{\mathbf{C}}$.

**Lemma 3.** *For $\tau = 1,\ldots,T$ define $\varepsilon_\tau := \max\{\widetilde{c}_0 \cdot 2^{-\tau}, C_3[\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n\}$, where $C_3 > 0$ is a constant depending only on $d, \alpha, M, \underline{p}_0, \overline{p}_0$ and $\mathbf{C}$. Denote also $\rho_\tau^* := \max_{x\in S_\tau}\varrho_\tau(x)$. Then for sufficiently large $n$, with probability $1 - O(n^{-1})$ the following holds uniformly for all outer iterations $\tau = 1,\ldots,T$:*

$$B_{\rho_\tau^*}^\infty(x;\mathcal{X}) \cap L_f(\varepsilon_\tau) \neq \emptyset. \tag{30}$$

Lemma 3 shows that the level $\varepsilon_\tau$ in $L_f(\varepsilon_\tau)$ that contains $S_{\tau-1}$ shrinks *geometrically*, until the condition $\varepsilon_\tau \geqslant C_3[\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n$ is violated. If the condition is never violated, then at the end of the last epoch $\tau^*$ we have $\varepsilon_{\tau^*} = O(n^{-1})$ because $\tau^* = \log n$. On the other hand, because $S_\tau \subseteq S_{\tau-1}$ always holds, we have $\varepsilon_{\tau^*} \lesssim [\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n$. Combining both cases we have that $\varepsilon_{\tau^*} \lesssim [\varepsilon_n^{\mathsf{U}}(f_0) + n^{-1/2}]\log^2 n + n^{-1}$. Theorem 1 is thus proved.

In the rest of this section we prove Lemma 3. We need several technical lemmas and propositions. Except for Proposition 6 that is straightforward, the proofs of the other technical lemmas are deferred to the end of this section.

Denote $x_n^* := \operatorname{argmin}_{x\in G_n} f(x)$ as the point on the grid $G_n$ with the smallest objective value The following proposition shows that with high probability, the confidence intervals constructed in the algorithm are truthful and the successive rejection procedure will never exclude the true optimizer of $f$ on $G_n$.

**Proposition 6.** *Suppose $\delta = 1/n^4|G_n|$. Then with probability $1 - O(n^{-1})$ the following hold:*

1) *$f(x') \in [\ell_t(x), u_t(x)]$ for all $1 \leqslant t \leqslant n$ and $x \in G_n$, $x' \in B_{h_t(x)}^\infty(x;\mathcal{X})$;*
2) *$x_n^* \in S_\tau$ for all $0 \leqslant \tau \leqslant n$.*

*Proof.* The first property is true by applying the union bound over all $t = 1,\ldots,n$ and $x \in G_n$. The second property then follows, because $\ell_t(x_n^*) \leqslant f(x_n^*)$ and $\min_{x\in S_{\tau-1}} u_t(x) \geqslant f(x_n^*)$ for all $\tau$. □

The following lemma shows that every small box centered around a certain sample point $x \in G_n$ contains a sufficient number of sample points whose least eigenvalue can be bounded with high probability under the polynomial mapping $\psi_{x,h}$ defined in Section III-B.

**Lemma 4.** *For any $x \in G_n$, $1 \leqslant m \leqslant n$ and $h > 0$, let $K_{h,m}^1(x),\ldots,K_{h,m}^n(x)$ be $n$ independent point sets, where each point set consists of $m$ points sampled i.i.d. uniformly at random from $B_h^\infty(x;G_n) = G_n \cap B_h^\infty(x;\mathcal{X})$. With probability $1 - O(n^{-1})$ the following holds true uniformly for all $x \in G_n$, $h \in \{j/n^2 : j \in \mathbb{N}, j \leqslant n^2\}$ and $K_{h,m}^\ell(x)$, $\ell \in [n]$ as $n \to \infty$:*

1) *$\sup_{h>0}\sup_{z\in B_h^\infty(x)}\|\psi_{x,h}(z)\|_2 \asymp \Theta(1)$;*
2) *$|B_h^\infty(x;G_n)| \asymp h^d|G_n|$;*
3) *$\sigma_{\min}(K_{h,m}^\ell(x)) \asymp \Theta(1)$ for all $m \geqslant \Omega(\log^2 n)$ and $m \leqslant |G_n|$, where $\sigma_{\min}(K_{h,m}^\ell(x))$ is the least eigenvalue of $\frac{1}{m}\sum_{z\in K_{h,m}^\ell(x)}\psi_{x,h}(z)\psi_{x,h}(z)^\top$.*

**Remark 10.** *It is possible to improve the concentration result in (48) using the strategies adopted in [35] based on sharper Bernstein type concentration inequalities. Such improvements are, however, not important in establishing the main results of this paper.*

The next lemma shows that, the bandwidth $h_t$ selected at the end of each outer iteration $\tau$ is near-optimal, being sandwiched between two quantities determined by the size of the active sample grid $\widetilde{S}_{\tau-1} := S^{\circ}_{\tau-1}(\varrho_{\tau-1})$.

**Lemma 5.** *There exist constants $C_1, C_2 > 0$ depending only on $d, \alpha, M, \underline{p}_0, \overline{p}_0$ and $\mathbf{C}$ such that with probability $1 - O(n^{-1})$, the following holds for every outer iteration $\tau \in \{1, \ldots, T\}$ and all $x \in S_{\tau-1}$:*

$$\begin{aligned} C_1[\widetilde{\nu}_{\tau-1} n_0]^{-1/(2\alpha+d)} - \tau/n &\leqslant \varrho_\tau(x) \\ &\leqslant h_t(x) \leqslant C_2[\widetilde{\nu}_{\tau-1} n_0]^{-1/(2\alpha+d)} \log n + \tau/n, \end{aligned} \tag{31}$$

*where $\widetilde{\nu}_{\tau-1} := |G_n|/|\widetilde{S}_{\tau-1}|$.*

We are now ready to state the proof of Lemma 3, which is based on an inductive argument over the epochts $\tau = 1, \ldots, T$.

*Proof.* We use induction to prove this lemma. For the base case $\tau = 1$, because $\|f - f_0\|_\infty \leqslant \varepsilon^{\mathsf{U}}_n(f_0)$ and $\varepsilon^{\mathsf{U}}_n(f_0) \to 0$ as $n \to \infty$, it suffices to prove that $B^{\infty}_{\rho_1^*}(x; \mathcal{X}) \cap L_{f_0}(\widetilde{c}_0/4) \neq \varnothing$ for all $x \in S_1$ and sufficiently large $n$. Because $\widetilde{S}_0 = S_0 = G_n$, invoking Lemmas 5 and 1 we have that $|\eta_{h_t(x),\delta}(x)| = \widetilde{O}(n^{-\alpha/(2\alpha+d)})$ for all $x \in G_n$ with high probability at the end of the first outer iteration $\tau = 1$. Therefore, for sufficiently large $n$ we conclude that $\sup_{x \in G_n} |\eta_{h_t(x),\delta}(x)| \leqslant c_0/16$ and hence $B^{\infty}_{\rho_1^*}(x; \mathcal{X}) \cap L_{f_0}(\widetilde{c}_0/4) \neq \varnothing$ for all $x \in S_1$.

We now prove the lemma for $\tau \geqslant 2$, assuming it holds for $\tau - 1$. We also assume that $n$ (and hence $n_0$) is sufficiently large, such that the maximum CI length $\max_{x \in G} |\eta_{h_t(x),\delta}(x)|$ after the first outer iteration $\tau = 1$ is smaller than $c_0/2$.

Because $\|f - f_0\|_\infty \leqslant \varepsilon^{\mathsf{U}}_n(f_0)$ and $\varepsilon_{\tau-1} \geqslant C_3 \varepsilon^{\mathsf{U}}_n(f_0) \log^2 n$, for appropriately chosen constant $C_3$ that is not too small, we have that $\|f - f_0\|_\infty \leqslant \varepsilon_{\tau-1}$. By the inductive hypothesis we have

$$\forall x \in S_{\tau-1}, \quad B^{\infty}_{\rho^*_{\tau-1}}(x; \mathcal{X}) \cap L_f(\varepsilon_{\tau-1}) \neq \varnothing;$$

Equivalently,

$$\begin{aligned} S_{\tau-1} \subseteq L^{\circ}_f(\varepsilon_{\tau-1}, \rho^*_{\tau-1}) &\subseteq L^{\circ}_{f_0}(\varepsilon_{\tau-1} + \|f - f_0\|_\infty, \rho^*_{\tau-1}) \\ &\subseteq L^{\circ}_{f_0}(2\varepsilon_{\tau-1}, \rho^*_{\tau-1}). \end{aligned} \tag{32}$$

Subsequently,

$$\widetilde{S}_{\tau-1} = S^{\circ}_{\tau-1} \subseteq L^{\circ}_{f_0}(2\varepsilon_{\tau-1}, 2\rho^*_{\tau-1}). \tag{33}$$

Let $\bigcup_{x \in H_n} B^2_{2\rho^*_{\tau-1}}(x)$ be the smallest covering set of $L_{f_0}(2\varepsilon_{\tau-1})$, meaning that $L_{f_0}(2\varepsilon_{\tau-1}) \subseteq \bigcup_{x \in H_n} B^2_{2\rho^*_{\tau-1}}(x)$, where $B^2_{2\rho^*_{\tau-1}}(x) = \{z \in \mathcal{X} : \|z - x\|_2 \leqslant 2\rho^*_{\tau-1}\}$ is the $\ell_2$ ball of radius $2\rho^*_{\tau-1}$ centered at $x$. By (A2), we know that $|H_n| \lesssim 1 + [\rho^*_{\tau-1}]^{-d} \mu_{f_0}(2\varepsilon_{\tau-1})$. In addition, the enlarged level-set satisfies $L^{\circ}_{f_0}(2\varepsilon_{\tau-1}, 2\rho^*_{\tau-1}) \subseteq \bigcup_{x \in H_n} B^{\infty}_{4\rho^*_{\tau-1}}(x)$. Subsequently,

$$\mu^{\circ}_{f_0}(2\varepsilon_{\tau-1}, \rho^*_{\tau-1}) \lesssim |H_n| \cdot [\rho^*_{\tau-1}]^d \lesssim \mu_{f_0}(2\varepsilon_{\tau-1}) + [\rho^*_{\tau-1}]^d. \tag{34}$$

By Lemma 5, the monotonicity of $|\widetilde{S}_{\tau-1}|$ and the fact that $\underline{p}_0 \leqslant p_X(z) \leqslant \overline{p}_0$ for all $z \in \mathcal{X}$, we have

$$\begin{aligned} \rho^*_{\tau-1} &\lesssim [\mu^{\circ}_f(\varepsilon_{\tau-1}, \rho^*_{\tau-1})]^{1/(2\alpha+d)} n_0^{-1/(2\alpha+d)} \log n & (35) \\ &\leqslant [\mu^{\circ}_{f_0}(2\varepsilon_{\tau-1}, \rho^*_{\tau-1})]^{1/(2\alpha+d)} n_0^{-1/(2\alpha+d)} \log n & (36) \\ &\lesssim \left(\mu_{f_0}(2\varepsilon_{\tau-1}) + [\rho^*_{\tau-1}]^d\right)^{1/(2\alpha+d)} n_0^{-1/(2\alpha+d)} \log n. & (37) \end{aligned}$$

Re-arranging terms on both sides of (37) we have

$$\rho^*_{\tau-1} \lesssim \max\left\{[\mu_{f_0}(2\varepsilon_{\tau-1})]^{\frac{1}{2\alpha+d}} n_0^{-\frac{1}{2\alpha+d}} \log n,\ n_0^{-\frac{1}{2\alpha}} \log n\right\}. \tag{38}$$

On the other hand, according to the selection procedure of the bandwidth $h_t(x)$, we have that $\eta_{h_t(x),\delta}(x) \lesssim \mathfrak{b}_{h_t(x),\delta}(x)$. Invoking Lemma 5 we have for all $x \in S_{\tau-1}$ that

$$\begin{aligned} \eta_{h_t(x),\delta}(x) \lesssim \mathfrak{b}_{h_t(x),\delta}(x) &\lesssim [h_t(x)]^\alpha & (39) \\ &\lesssim [\widetilde{\nu}_{\tau-1} n_0]^{-\alpha/(2\alpha+d)} \log n & (40) \\ &\lesssim [\widetilde{\nu}_{\tau-2} n_0]^{-\alpha/(2\alpha+d)} \log n & (41) \\ &\lesssim [\rho^*_{\tau-1}]^\alpha \log n. & (42) \end{aligned}$$

Here (40) holds by invoking the upper bound on $h_t(x)$ in Lemma 5, (41) holds because $\widetilde{\nu}_{\tau-1} \geqslant \widetilde{\nu}_{\tau-2}$, and (42) holds by again invoking the lower bound on $\varrho_{\tau-1}(x)$ in Lemma 5. Combining Eqs. (38,42) we have

$$\begin{aligned} &\max_{x \in S_{\tau-1}} \eta_{h_t(x),\delta}(x) & (43) \\ &\lesssim \max\left\{[\mu_{f_0}(2\varepsilon_{\tau-1})]^{\frac{\alpha}{2\alpha+d}} n_0^{-\frac{\alpha}{2\alpha+d}} \log^2 n,\ n_0^{-\frac{1}{2}} \log n\right\}. & (44) \end{aligned}$$

Recall that $n_0 = n/\log n$ and $\varepsilon^{\mathsf{U}}_n(f_0) \leqslant \varepsilon_{\tau-1}$, provided that $C_3$ is not too small. By definition, every $\varepsilon \geqslant \varepsilon^{\mathsf{U}}_n(f_0)$ satisfies $\varepsilon^{-(2+d/\alpha)} \mu_{f_0}(\varepsilon) \leqslant n/\log^\omega n$ for some large constant $\omega > 5 + d/\alpha$. Subsequently,

$$\begin{aligned} &[\mu_{f_0}(2\varepsilon_{\tau-1})]^{\frac{\alpha}{2\alpha+d}} n_0^{-\frac{\alpha}{2\alpha+d}} \log^2 \\ &\lesssim 2\varepsilon_{\tau-1} n^{\frac{\alpha}{2\alpha+d}} \log^{-\frac{\omega\alpha}{2\alpha+d}} n \cdot n_0^{-\frac{\alpha}{2\alpha+d}} \log^2 n & (45) \\ &\lesssim \varepsilon_{\tau-1}/[\log n]^{\frac{(\omega-5-d/\alpha)\alpha}{2\alpha+d}}. & (46) \end{aligned}$$

Because $\omega > 5 + d/\alpha$, the right-hand side of (46) is asymptotically dominated [6] by $\varepsilon_{\tau-1}$. In addition, $n_0^{-1/2} \log n$ is also asymptotically dominated by $\varepsilon_{\tau-1}$ because $\varepsilon_{\tau-1} \geqslant C_3 n^{-1/2} \log^\omega n$. Therefore, for sufficiently large $n$ we have

$$\max_{x \in S_{\tau-1}} \eta_{h_t(x),\delta}(x) \leqslant \varepsilon_{\tau-1}/4. \tag{47}$$

Lemma 3 is thus proved. □

[6] We say $\{a_n\}$ is asymptotically dominated by $\{b_n\}$ if $\lim_{n\to\infty} |a_n|/|b_n| = 0$.

*1) Proof of Lemma 4:*

*Proof.* We first show that the first property holds almost surely. Recall the definition of $\psi_{x,h}$, we have that $1 \leqslant \|\psi_{x,h}(z)\|_2 \leqslant D \cdot [\max_{1\leqslant j\leqslant d} h^{-1}|z_j - x_j|]^k$. Because $\|z-x\|_\infty \leqslant h$ for all $z \in B_h^\infty(x)$, $\sup_{z\in B_h^\infty(x)} \|\psi_{x,h}(z)\|_2 \lesssim O(1)$ for all $h>0$. Thus, $\sup_{h>0}\sup_{z\in B_h^\infty(x)} \|\psi_{x,h}(z)\|_2 \asymp \Theta(1)$ for all $x \in G_n$.

For the second property, by Hoeffding's inequality (Lemma 9) and the union bound, with probability $1-O(n^{-1})$ we have that

$$\max_{x,h}\left|\frac{|B_h^\infty(x;G_n)|}{|G_n|} - P_X(z\in B_h^\infty(x))\right| \lesssim \sqrt{\frac{\log n}{|G_n|}}. \quad (48)$$

In addition, note that $P_X(z \in B_h^\infty(x;\mathcal{X})) \geqslant \underline{p}_0\lambda(B_h^\infty(x;\mathcal{X})) \gtrsim h^d$ and $P_X(z \in B_h^\infty(x;\mathcal{X})) \leqslant \overline{p}_0\lambda(B_h^\infty(x;\mathcal{X})) \lesssim h^d$, where $\lambda(\cdot)$ denotes the Lebesgue measure on $\mathcal{X}$. Subsequently, $|B_h^\infty(x;G_n)|$ is lower bounded by $\Omega(h^d|G_n| - \sqrt{|G_n|\log n})$ and upper bounded by $O(h^d|G_n| + \sqrt{|G_n|\log n})$. The second property is then proved by noting that $h_d \gtrsim n^{-d}$ and $|G_n| \gtrsim n^{3d/\min(\alpha,1)}$.

We next prove the third property. Because $\underline{p}_0 \leqslant p_X(z) \in \overline{p}_0$ for all $z\in\mathcal{X}$, we have that

$$\underline{p}_0\int_{B_h^\infty(x;\mathcal{X})}\psi_{x,h}(z)\psi_{x,h}(z)^\top \mathrm{d}U_{x,h}(z)$$

$$\preceq \mathbb{E}\left[\frac{1}{m}\sum_{z\in K_{h,m}^\ell}\psi_{x,h}(z)\psi_{x,h}(z)^\top\right] \quad (49)$$

$$\preceq \overline{p}_0\int_{B_h^\infty(x;\mathcal{X})}\psi_{x,h}(z)\psi_{x,h}(z)^\top \mathrm{d}U_{x,h}(z), \quad (50)$$

where $U_{x,h}$ is the uniform distribution on $B_h^\infty(x;\mathcal{X})$. Note also that

$$\int_{\mathcal{X}}\psi_{0,1}(z)\psi_{0,1}(z)^\top\mathrm{d}U(z)$$

$$\preceq \int_{B_h^\infty(x;\mathcal{X})}\psi_{x,h}(z)\psi_{x,h}(z)^\top \mathrm{d}U_{x,h}(z) \quad (51)$$

$$\preceq 2^d\int_{\mathcal{X}}\psi_{0,1}(z)\psi_{0,1}(z)^\top\mathrm{d}U(z) \quad (52)$$

where $U$ is the uniform distribution on $\mathcal{X}=[0,1]^d$. The following proposition upper and lower bounds the eigenvalues of $\int_{\mathcal{X}}\psi_{0,1}(z)\psi_{0,1}(z)^\top\mathrm{d}U(z)$, which is proved in the appendix.

**Proposition 7.** *There exist constants $0<\psi_0\leqslant\Psi_0<\infty$ depending only on $d, D$ such that*

$$\psi_0\ I_{D\times D} \preceq \int_{\mathcal{X}}\psi_{0,1}(z)\psi_{0,1}(z)^\top\mathrm{d}U(z) \preceq \Psi_0 I_{D\times D}. \quad (53)$$

Using Proposition 7 and Eqs. (51,52), we conclude that

$$\Omega(1)\cdot I_{D\times D} \preceq \mathbb{E}\left[\frac{1}{m}\sum_{z\in K_{h,m}^\ell}\psi_{x,h}(z)\psi_{x,h}(z)^\top\right] \preceq O(1)\cdot I_{D\times D}. \quad (54)$$

Applying matrix Chernoff bound (Lemma 11) and the union bound, we have that with probability $1-O(n^{-1})$,

$$\max_{x,h,m,\ell}\left\|\frac{1}{m}\sum_{z\in K_{h,m}^\ell(x)}\psi_{x,h}(z)\psi_{x,h}(z)^\top - \mathbb{E}\left[\psi_{x,h}(z)\psi_{x,h}(z)^\top|z\in B_h^\infty(x)\right]\right\|_{\mathrm{op}} \lesssim \sqrt{\frac{\log n}{m}}.$$

Combining Eqs. (54,55) and applying Weyl's inequality (Lemma 12) we have

$$\Omega(1) - O(\sqrt{\log n/m}) \lesssim \sigma_{\min}(K_{h,m}^\ell(x))$$

$$\lesssim O(1) - O(\sqrt{\log n/m}). \quad (55)$$

The third property is therefore proved. □

*2) Proof of Lemma 5: Proof.* We use induction to prove this lemma. For the base case of $\tau=1$, we have $\widetilde{S}_0 = S_0 = G_n$ and therefore $\widetilde{\nu}_{\tau-1}=1$. Furthermore, applying Lemma 4 we have that for all $h=j/n^2$,

$$\mathfrak{b}_{h,\delta}(x) \asymp h^\alpha,\ \ \mathfrak{s}_{h,\delta}(x) \asymp \sqrt{\frac{\log n}{h^d n_0}}. \quad (56)$$

Thus, for $h$ selected according to (18) as the largest bandwidth of the form $j/n^2$, $j\in\mathbb{N}$ such that $\mathfrak{b}_{h,\delta}(x)\leqslant\mathfrak{s}_{h,\delta}(x)$, both $\mathfrak{b}_{h,\delta}(x),\mathfrak{s}_{h,\delta}(x)$ are on the order of $n_0^{-1/(2\alpha+d)}$ up to logarithmic terms of $n$, and therefore one can pick appropriate constants $C_1, C_2>0$ such that $C_1 n_0^{-1/(2\alpha+d)} \leqslant \varrho_1(x) \leqslant C_2 n_0^{-1/(2\alpha+d)}\log n$ holds for all $x\in G_n$.

We next prove the lemma for $\tau>1$, assuming it holds for $\tau-1$. We first establish the lower bound part. Define $\rho_{\tau-1}^* := \min_{z\in S_{\tau-1}}\varrho_{\tau-1}(z)$. By inductive hypothesis, $\rho_{\tau-1}^* \geqslant C_1[\widetilde{\nu}_{\tau-2}n_0]^{-1/(2\alpha+d)} - (\tau-1)/n$. Note also that $\widetilde{\nu}_{\tau-1}\geqslant\widetilde{\nu}_{\tau-2}$ because $\widetilde{S}_{\tau-1}\subseteq\widetilde{S}_{\tau-2}$, which holds because $S_{\tau-1}\subseteq S_{\tau-2}$ and $\varrho_{\tau-1}(z)\leqslant\varrho_{\tau-2}(z)$ for all $z$. Let $h_t^*$ be the smallest number of the form $j_t^*/n^2$, $j_t^*\in[n^2]$ such that $h_t^*\geqslant C_1[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} - \tau/n$. We then have $h_t^*\leqslant\rho_{\tau-1}^*$ and therefore query points in epoch $\tau$ are uniformly distributed in $B_{h_t^*}^\infty(x;G_n)$. Subsequently, applying Lemma 4 we have with probability $1-O(n^{-1})$ that

$$\mathfrak{b}_{h_t^*,\delta}(x)\leqslant C'[h_t^*]^\alpha,\ \ \mathfrak{s}_{h_t^*,\delta}(x)\geqslant C''\sqrt{\frac{\log n}{[h_t^*]^d\widetilde{\nu}_{\tau-1}n}}, \quad (57)$$

where $C', C''>0$ are constants that depend on $d,\alpha,M,\underline{p}_0,\overline{p}_0$ and $\mathbf{C}$, but *not* $C_1$, $C_2$, $\tau$ or $h_t^*$. By choosing $C_1$ appropriately (depending on $C'$ and $C''$) we can make $\mathfrak{b}_{h_t^*,\delta}(x)\leqslant\mathfrak{s}_{h_t^*,\delta}(x)$ holds for all $x\in S_{\tau-1}$, thus establishing $\varrho_\tau(x)\geqslant\min\{\varrho_{\tau-1}(x),h_t^*\}\geqslant C_1[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)}-\tau/n$.

We next prove the upper bound part. For any $h_t=j_t/n^2$ where $j_t\in[n^2]$, invoking Lemma 4 we have that

$$\mathfrak{b}_{h,\delta}(x)\geqslant\widetilde{C}'h^\alpha,\ \mathfrak{s}_{h,\delta}(x)\leqslant\widetilde{C}''\sqrt{\frac{\log n}{\min\{h,\rho_{\tau-1}^*\}^d\cdot\widetilde{\nu}_{\tau-1}n_0}}, \quad (58)$$

where $\widetilde{C}'$ and $\widetilde{C}''$ are again constants depending on $d, \alpha, M, \underline{p}_0, \overline{p}_0$ and $\mathbf{C}$, but *not* $C_1, C_2$. Note also that $\rho^*_{\tau-1} \geqslant C_1[\widetilde{\nu}_{\tau-2}n_0]^{-1/(2\alpha+d)} - (\tau-1)/n \geqslant C_1[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} - \tau/n$, because $\widetilde{\nu}_{\tau-1} \geqslant \widetilde{\nu}_{\tau-2}$. By selecting constant $C_2 > 0$ carefully (depending on $\widetilde{C}', \widetilde{C}''$ and $C_1$), we can ensure $\mathfrak{b}_{h,\delta}(x) > \mathfrak{s}_{h,\delta}(x)$ for all $h \geqslant C_2[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} + \tau/n$. Therefore, $\varrho_\tau(x) \leqslant h_t(x) \leqslant C_2[\widetilde{\nu}_{\tau-1}n_0]^{-1/(2\alpha+d)} + \tau/n$. □

### *C. Proof of Theorem 2*

In this section we prove the main negative result in Theorem 2. To simplify presentation, we suppress dependency on $\alpha, d, c_0$ and $C_0$ in $\lesssim, \gtrsim, \asymp, O(\cdot)$ and $\Omega(\cdot)$ notations. However, we do *not* suppress dependency on $\underline{C}_R$ or $M$ in any of the above notations.

Let $\varphi_0 : [-2,2]^d \to \mathbb{R}^*$ be a non-negative function defined on $\mathcal{X}$ such that $\varphi_0 \in \Sigma_\kappa^{\lceil\alpha\rceil}(1)$ with $\kappa = \infty$, $\sup_{x\in\mathcal{X}} \varphi_0(x) = \Omega(1)$ and $\varphi_0(z) = 0$ for all $\|z\|_2 \geqslant 1$. Here $\lceil\alpha\rceil$ denotes the smallest integer that upper bounds $\alpha$. Such functions exist and are the cornerstones of the construction of information-theoretic lower bounds in nonparametric estimation problems [50]. One typical example is the "smoothstep" function (see for example [54])

$$S_N(x) := \frac{1}{Z}x^{N+1}\sum_{n=0}^{N}\binom{N+n}{n}\binom{2N+1}{N-n}(-x)^n, \quad N = 0,1,2,\ldots,$$

where $Z > 0$ is a scaling parameter. The smoothstep function $S_N$ is defined on $[0,1]$ and satisfies the Hölder condition in (6) of order $\alpha = N$ on $[0,1]$. It can be easily extended to $\widetilde{S}_{N,d} : [-2,2]^d \to \mathbb{R}$ by considering $\widetilde{S}_{N,d}(x) := 1/Z - S_N(a\|x\|_1)$ where $\|x\|_1 = |x_1| + \ldots + |x_d|$ and $a = 1/(2d)$. It is easy to verify that, with $Z$ chosen appropriately, $\widetilde{S}_{N,d} \in \Sigma_\infty^N(1)$, $\sup_{x\in\mathcal{X}} \widetilde{S}_{N,d}(x) = 1/Z = \Omega(1)$ and $\widetilde{S}_{N,d}(z) = 0$ for all $\|z\|_2 \geqslant 1$, where $M > 0$ is a constant.

For any $x \in \mathcal{X}$ and $h > 0$, define $\varphi_{x,h} : \mathcal{X} \to \mathbb{R}^*$ as

$$\varphi_{x,h}(z) := \mathbb{I}[z \in B_h^\infty(x)] \cdot \frac{Mh^\alpha}{2}\varphi_0\left(\frac{z-x}{h}\right). \tag{59}$$

It is easy to verify that $\varphi_{x,h} \in \Sigma_\infty^\alpha(M/2)$, and furthermore $\sup_{z\in\mathcal{X}} \varphi_{x,h}(z) \asymp Mh^\alpha$ and $\varphi_{x,h}(z) = 0$ for all $z \notin B_h^\infty(x)$.

Let $L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0))$ be the level-set of $f_0$ at $\varepsilon_n^{\mathsf{L}}(f_0)$. Let $H_n \subseteq L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0))$ be the largest *packing* set such that $B_h^\infty(x)$ are disjoint for all $x \in H_n$, and $\bigcup_{x\in H_n} B_h^\infty(x) \subseteq L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0))$. By (A2') and the definition of $\varepsilon_n^{\mathsf{L}}(f_0)$, we have that

$$\begin{aligned} |H_n| &\geqslant M(L_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0)), 2\sqrt{d}h) \\ &\gtrsim \mu_{f_0}(\varepsilon_n^{\mathsf{L}}(f_0)) \cdot h^{-d} \geqslant [\varepsilon_n^{\mathsf{L}}(f_0)]^{2+d/\alpha} \cdot nh^{-d}. \end{aligned} \tag{60}$$

For any $x \in H_n$, construct $f_x : \mathcal{X} \to \mathbb{R}$ as

$$f_x(z) := f_0(z) - \varphi_{x,h}(z). \tag{61}$$

Let $\mathcal{F}_n := \{f_x : x \in H_n\}$ be the class of functions indexed by $x \in H_n$. Let also $h \asymp (\varepsilon_n^{\mathsf{L}}(f_0)/M)^{1/\alpha}$ such that $\|\varphi_{x,h}\|_\infty = 2\varepsilon_n^{\mathsf{L}}(f_0)$. We then have that $\|f_x - f_0\|_\infty \leqslant 2\varepsilon_n^{\mathsf{L}}(f_0)$ and $f_x \in \Sigma_\infty^\alpha(M)$, because $f_0, \varphi_{x,h} \in \Sigma_\infty^\alpha(M/2)$.

The next lemma shows that, with $n$ adaptive queries to the noisy zeroth-order oracle $y_t = f(x_t) + w_t$, it is information theoretically not possible to identify a certain $f_x$ in $\mathcal{F}_n$ with high probability.

**Lemma 6.** *Suppose $|\mathcal{F}_n| \geqslant 2$. Let $\mathcal{A}_n = (\chi_1, \ldots, \chi_n, \phi_n)$ be an active optimization algorithm operating with a sample budget n, which consists of* samplers $\chi_\ell : \{(x_i, y_i)\}_{i=1}^{\ell-1} \mapsto x_\ell$ *and an* estimator $\phi_n : \{(x_i, y_i)\}_{i=1}^n \mapsto \widehat{f}_x \in \mathcal{F}_n$, *both can be deterministic or randomized functions. Then*

$$\inf_{\mathcal{A}_n} \sup_{f_x\in\mathcal{F}_n} \Pr_{f_x}\left[\widehat{f}_x \neq f_x\right] \geqslant \frac{1}{2} - \sqrt{\frac{n \cdot \sup_{f_x\in\mathcal{F}_n}\|f_x - f_0\|_\infty^2}{2|\mathcal{F}_n|}}. \tag{62}$$

**Lemma 7.** *There exists constant $M > 0$ depending on $\alpha, d, c_0, C_0$ such that the right-hand side of (62) is lower bounded by* $1/3$.

Lemmas 6 and 7 are proved at the end of this section. Combining both lemmas and noting that for any distinct $f_x, f_{x'} \in \mathcal{F}_n$ and $z \in \mathcal{X}$, $\max\{\mathfrak{L}(z; f_x), \mathfrak{L}(z; f_{x'})\} \geqslant \varepsilon_n^{\mathsf{L}}(f_0)$, we proved the minimax lower bound formulated in Theorem 2.

*1) Proof of Lemma 6:* Our proof is inspired by the negative result of multi-arm bandit pure exploration problems established in [51].

*Proof.* For any $x \in H_n$, define

$$n_x := \mathbb{E}_{f_0}\left[\sum_{i=1}^n \mathbb{I}[x \in B_h^\infty(x)]\right]. \tag{63}$$

Because $B_h^\infty(x)$ are disjoint for $x \in H_n$, we have $\sum_{x\in H_n} n_x \leqslant n$. Also define, for every $x \in H_n$,

$$\wp_x := \Pr_{f_0}\left[\widehat{f}_x = f_x\right]. \tag{64}$$

Because $\sum_{x\in H_n} \wp_x = 1$, by pigeonhole principle there is at most one $x \in H_n$ such that $\wp_x > 1/2$. Let $x_1, x_2 \in H_n$ be the points that have the smallest and second smallest $n_x$. Then there exists $x \in \{x_1, x_2\}$ such that $\wp_x \leqslant 1/2$ and $n_x \leqslant 2n/|\mathcal{F}_n|$. By Le Cam's and Pinsker's inequality (see, for example, [4]) we have that

$$\begin{aligned} \Pr_{f_x}\left[\widehat{f}_x = f_x\right] &\leqslant \Pr_{f_0}\left[\widehat{f}_x = f_x\right] + d_{\mathrm{TV}}(P_{f_0}^{\mathcal{A}_n}\|P_{f_x}^{\mathcal{A}_n}) && (65)\\ &\leqslant \Pr_{f_0}\left[\widehat{f}_x = f_x\right] + \sqrt{\frac{1}{2}\mathrm{KL}(P_{f_0}^{\mathcal{A}_n}\|P_{f_x}^{\mathcal{A}_n})} && (66)\\ &= \wp_x + \sqrt{\frac{1}{2}\mathrm{KL}(P_{f_0}^{\mathcal{A}_n}\|P_{f_x}^{\mathcal{A}_n})} && (67)\\ &\leqslant \frac{1}{2} + \sqrt{\frac{1}{2}\mathrm{KL}(P_{f_0}^{\mathcal{A}_n}\|P_{f_x}^{\mathcal{A}_n})}. && (68) \end{aligned}$$

It remains to upper bound KL divergence of the active queries made by $\mathcal{A}_n$. Using the standard lower bound analysis for active learning algorithms [50], [55] and the fact that

$f_x \equiv f_0$ on $\mathcal{X}\backslash B_h^\infty(x)$, we have

$$\mathrm{KL}(P_{f_0}^{\mathcal{A}_n}\|P_{f_x}^{\mathcal{A}_n}) = \mathbb{E}_{f_0,\mathcal{A}_n}\left[\log\frac{P_{f_0,\mathcal{A}_n}(x_{1:n},y_{1:n})}{P_{f_x,\mathcal{A}_n}(x_{1:n},y_{1:n})}\right] \tag{69}$$

$$= \mathbb{E}_{f_0,\mathcal{A}_n}\left[\log\frac{\prod_{i=1}^n P_{f_0}(y_i|x_i)P_{\mathcal{A}_n}(x_i|x_{1:(i-1)},y_{1:(i-1)})}{\prod_{i=1}^n P_{f_x}(y_i|x_i)P_{\mathcal{A}_n}(x_i|x_{1:(i-1)},y_{1:(i-1)})}\right] \tag{70}$$

$$= \mathbb{E}_{f_0,\mathcal{A}_n}\left[\log\frac{\prod_{i=1}^n P_{f_0}(y_i|x_i)}{\prod_{i=1}^n P_{f_x}(y_i|x_i)}\right] \tag{71}$$

$$= \mathbb{E}_{f_0,\mathcal{A}_n}\left[\sum_{x_i\in B_h^\infty(x)}\log\frac{P_{f_0}(y_i|x_i)}{P_{f_x}(y_i|x_i)}\right] \tag{72}$$

$$\leqslant n_x\cdot\sup_{z\in B_h^\infty(x;\mathcal{X})}\mathrm{KL}(P_{f_0}(\cdot|z)\|P_{f_x}(\cdot|z)) \tag{73}$$

$$\leqslant n_x\cdot\|f_0-f_x\|_\infty^2. \tag{74}$$

Therefore,

$$\Pr_{f_x}\left[\widehat{f}_x = f_x\right] \leqslant \frac{1}{2}+\sqrt{\frac{1}{4}n_x\varepsilon_n^2} \leqslant \frac{1}{2}+\sqrt{\frac{n\|f_x-f_0\|_\infty^2}{2|\mathcal{F}_n|}}. \tag{75}$$

□

*2) Proof of Lemma 7:*

*Proof.* By construction, $n\sup_{f_x\in\mathcal{F}_x}\|f_x-f_0\|_\infty^2 \lesssim M^2\ nh^{2\alpha}$ and $|\mathcal{F}_n| = |H_n| \gtrsim [\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)]^{2+d/\alpha}nh^{-d}$. Note also that $h \asymp (\varepsilon/M)^{1/\alpha} \asymp (\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)/M)^{1/\alpha}$ because $\|f_x-f_0\|_\infty = \varepsilon = \underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)$. Subsequently,

$$\frac{n\sup_{f_x\in\mathcal{F}_x}\|f_x-f_0\|_\infty^2}{2|\mathcal{F}_n|} \lesssim \frac{n[\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)]^2}{n[\underline{C}_\varepsilon\varepsilon_n^{\mathsf{L}}(f_0)]^2\cdot M^{d/\alpha}} = M^{-d/\alpha}. \tag{76}$$

By choosing the constant $M>0$ to be sufficiently large, the right-hand side of the above inequality is upper bounded by $1/36$. The lemma is thus proved. □

### D. Proof of Theorem 3

The proof of Theorem 3 is similar to the proof of Theorem 2, but is much more standard by invoking the *Fano's inequality* [4]. In particular, adapting the Fano's inequality on any finite function class $\mathcal{F}_n$ constructed we have the following lemma:

**Lemma 8** (Fano's inequality)**.** *Suppose $|\mathcal{F}_n| \geqslant 2$, and $\{(x_i,y_i)\}_{i=1}^n$ are i.i.d. random variables. Then*

$$\inf_{\widehat{f}_x}\sup_{f_x\in\mathcal{F}_n}\Pr_{f_x}\left[\widehat{f}_x \neq f_x\right] \geqslant 1-\frac{\log 2+n\cdot\sup_{f_x,f_{x'}\in\mathcal{F}_n}\mathrm{KL}(P_{f_x}\|P_{f_{x'}})}{\log|\mathcal{F}_n|}, \tag{77}$$

*where $P_{f_x}$ denotes the distribution of $(x,y)$ under the law of $f_x$.*

Let $\mathcal{F}_n$ be the function class constructed in the previous proof of Theorem 2, corresponding to the largest packing set $H_n$ of $L_{f_0}(\tilde{\varepsilon}_n^{\mathsf{L}})$ such that $B_h^\infty(x)$ for all $x\in H_n$ are disjoint, where $h \asymp (\tilde{\varepsilon}_n^{\mathsf{L}}/M)^{1/\alpha}$ such that $\|\varphi_{x,h}\|_\infty = 2\tilde{\varepsilon}_n^{\mathsf{L}}$ for all $x\in H_n$. Because $f_0$ satisfies (A2'), we have that $|\mathcal{F}_n| = |H_n| \gtrsim \mu_{f_0}(\tilde{\varepsilon}_n^{\mathsf{L}})h^{-d}$. Under the condition that $\varepsilon_n^{\mathsf{U}}(f_0) \leqslant \tilde{\varepsilon}_n^{\mathsf{L}}$, it holds that $\mu_{f_0}(\tilde{\varepsilon}_n^{\mathsf{L}}) \geqslant [\tilde{\varepsilon}_n^{\mathsf{L}}]^{2+d/\alpha}n$. Therefore,

$$|\mathcal{F}_n| \gtrsim [\tilde{\varepsilon}_n^{\mathsf{L}}]^{2+d/\alpha}\cdot nh^{-d} \gtrsim [\tilde{\varepsilon}_n^{\mathsf{L}}]^2\cdot nM^{d/\alpha}. \tag{78}$$

Because $\log(n/\tilde{\varepsilon}_n^{\mathsf{L}}) \gtrsim \log n$ and $M>0$ is a constant, we have that $\log|\mathcal{F}_n| \geqslant c\log n$ for all $n\geqslant N$, where $c>0$ is a constant depending only on $\alpha, d$ and $N\in\mathbb{N}$ is a constand depending on $M$.

Let $U$ be the uniform distribution on $\mathcal{X}$. Because $x\sim U$ and $f_x \equiv f_{x'}$ on $\mathcal{X}\backslash B_h^\infty(x)$, we have that

$$\mathrm{KL}(P_{f_x}\|P_{f_{x'}}) = \frac{1}{2}\int_{\mathcal{X}}|f_x(z)-f_{x'}(z)|^2\mathrm{d}U(z) \tag{79}$$

$$\leqslant \frac{1}{2}\Pr_U\left[z\in B_h^\infty(x)\right]\cdot\|f_x-f_{x'}\|_\infty^2 \tag{80}$$

$$\leqslant \frac{1}{2}\lambda(B_h^\infty(x))\cdot[\varepsilon_n^{\mathsf{L}}]^2 \tag{81}$$

$$\lesssim h^d[\tilde{\varepsilon}_n^{\mathsf{L}}]^2 \lesssim [\tilde{\varepsilon}_n^{\mathsf{L}}]^{2+d/\alpha}/M^{d/\alpha}. \tag{82}$$

By choosing $M$ to be sufficiently large, the right-hand side of (77) can be lower bounded by an absolute constant. The theorem is then proved following the same argument as in the proof of Theorem 2.

## Appendix A
## Some Concentration Inequalities

In this section, to ease readability of our paper, we provide some concentration inequalities and other standard results that we use extensively.

**Lemma 9** ( [56])**.** *Suppose $X_1,\ldots,X_n$ are i.i.d. random variables such that $a\leqslant X_i\leqslant b$ almost surely. Then for any $t>0$,*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i-\mathbb{E}X\right|>t\right] \leqslant 2\exp\left\{-\frac{nt^2}{2(b-a)^2}\right\}.$$

**Lemma 10** ( [57])**.** *Suppose $x\sim\mathcal{N}_d(0,I_{d\times d})$ and let $A$ be a $d\times d$ positive semi-definite matrix. Then for all $t>0$,*

$$\Pr\left[x^\top Ax > \mathrm{tr}(A)+2\sqrt{\mathrm{tr}(A^2)t}+2\|A\|_{\mathrm{op}}t\right] \leqslant e^{-t}.$$

**Lemma 11** ( [58], simplified)**.** *Suppose $A_1,\ldots,A_n$ are i.i.d. positive semidefinite random matrices of dimension $d$ and $\|A_i\|_{\mathrm{op}}\leqslant R$ almost surely. Then for any $t>0$,*

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^n A_i-\mathbb{E}A\right\|_{\mathrm{op}}>t\right] \leqslant 2\exp\left\{-\frac{nt^2}{8R^2}\right\}.$$

**Lemma 12** (Weyl's inequality)**.** *Let $A$ and $A+E$ be $d\times d$ matrices with $\sigma_1,\ldots,\sigma_d$ and $\sigma_1',\ldots,\sigma_d'$ be their singular values, sorted in descending order. Then* $\max_{1\leqslant i\leqslant d}|\sigma_i-\sigma_i'|\leqslant\|E\|_{\mathrm{op}}$.

## Appendix B
## Additional Proofs

*Proof of Proposition 1.* Consider arbitrary $x^* \in \mathcal{X}$ such that $f(x^*) = \inf_{x\in\mathcal{X}} f(x)$. Then we have that $\mathfrak{L}(\widehat{x}_n; f) = f(\widehat{x}_n) - f(x^*) \leqslant [\widehat{f}_n(\widehat{x}_n) + \|\widehat{f}_n - f\|_\infty] - [\widehat{f}_n(x^*) - \|\widehat{f}_n - f\|_\infty] \leqslant 2\|\widehat{f}_n - f\|_\infty$, where the last inequality holds because $\widehat{f}_n(\widehat{x}_n) \leqslant \widehat{f}_n(x^*)$ by optimality of $\widehat{x}_n$. □

*Proof of Example 2.* Because $f_0 \in \Sigma_\kappa^2(M)$ is strongly convex, there exists $\sigma > 0$ such that $\nabla^2 f_0(x) \succeq \sigma I$ for all $x \in \mathcal{X}_{f_0,\kappa}$, where $\mathcal{X}_{f_0,\kappa} := L_{f_0}(\kappa)$ is the $\kappa$-level-set of $f_0$. Let $x^* = \arg\min_{x\in\mathcal{X}} f_0(x)$, which is unique because $f_0$ is strongly convex. The smoothness and strong convexity of $f_0$ implies that

$$f_0^* + \frac{\sigma}{2}\|x - x^*\|_\infty^2 \leqslant f_0(x) \leqslant f_0^* + \frac{M}{2}\|x - x^*\|_\infty^2 \quad \forall x \in \mathcal{X}_{f_0,\kappa}. \tag{83}$$

Subsequently, there exist constants $c_0, C_1, C_2 > 0$ depending only on $\sigma, M, \kappa$ and $d$ such that for all $\epsilon \in (0, c_0]$,

$$B^\infty_{C_1\sqrt{\epsilon}}(x^*; \mathcal{X}) \subseteq L_{f_0}(\epsilon) \subseteq B^\infty_{C_2\sqrt{\epsilon}}(x^*; \mathcal{X}). \tag{84}$$

The property $\mu_{f_0}(\epsilon) \lesssim \epsilon^\beta$ holds because $\mu(L_{f_0}(\epsilon)) \leqslant \mu(B^\infty_{C_1\sqrt{\epsilon}}(x^*; \mathcal{X})) \lesssim \epsilon^{d/2}$. To prove (A2), note that $N(L_{f_0}(\epsilon), \delta) \leqslant N(B^\infty_{C_2\sqrt{\epsilon}}(x^*; \mathcal{X}), \delta) \lesssim 1 + (\sqrt{\epsilon}/\delta)^d$. Because $\epsilon^{d/2} \lesssim \mu(L_{f_0}(\epsilon)) = \mu_{f_0}(\epsilon)$, we conclude that $N(L_{f_0}(\epsilon), \delta) \lesssim 1 + \delta^{-d}\mu_{f_0}(\epsilon)$ and (A2) is thus proved. □

*Proof of Proposition 4.* Consider $f_0 \equiv 0$ if $\beta = 0$ and $f_0(z) := a_0\left[z_1^p + \ldots + z_d^p\right]$ for all $z = (z_1, \ldots, z_d) \in [0,1]^d$, where $a_0 > 0$ is a constant depending on $\alpha, M$, and $p = d/\beta$ for $\beta \in (0, d/\alpha]$. The $\beta = 0$ case where $f_0 \equiv 0$ trivially holds. So we shall only consider the case of $\beta \in (0, d/\alpha]$.

We first show $f_0 \in \Sigma_\kappa^\alpha(M)$ with $\kappa = \infty$, provided that $a_0$ is sufficiently small. For any $j \leqslant k = \lfloor\alpha\rfloor$ and $\alpha_1 + \ldots + \alpha_d = j$, we have

$$\frac{\partial^j}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} f_0(z) = \begin{cases} a_0 j! \cdot z_\ell^{p-j} & \text{if } \alpha_\ell = j, \ell \in [d]; \\ 0 & \text{otherwise.} \end{cases} \tag{85}$$

Because $z_1, \ldots, z_d \in [0,1]$ and $p = d/\beta \geqslant \alpha \geqslant j$, it's clear that $0 \leqslant \partial^j f_0(z)/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d} \leqslant a_0\ j!$. In addition, for any $z, z' \in [0,1]^d$ and $\alpha_\ell = k$, $\ell \in [d]$, we have

$$\begin{aligned} \left|\frac{\partial^k}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} f_0(z) - \frac{\partial^k}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} f_0(z')\right| & \\ \leqslant a_0 k! \cdot \left|[z_\ell]^{p-k} - [z'_\ell]^{p-k}\right| & \quad (86) \\ \leqslant a_0 k! \cdot \left|z_\ell - z'_\ell\right|^{\min\{p-k,1\}}, & \quad (87) \end{aligned}$$

where the last inequality holds because $x^t$ is $\min\{t,1\}$-Hölder continuous on $[0,1]$ for $t \geqslant 0$. The $|z_\ell - z'_\ell|^{\min\{p-k,1\}}$ term can be further upper bounded by $\|z - z'\|_\infty^{\alpha-k}$, because $p = d/\beta \geqslant \alpha$. By selecting $a_0 > 0$ to be sufficiently small (depending on $M$) we have $f_0 \in \Sigma_\infty^\alpha(M)$.

We next prove $f_0$ satisfies $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ with parameter $\beta$ depending on $a_0$ and $p$. For any $\epsilon > 0$, the level-set $L_{f_0}(\epsilon)$ can be expressed as $L_{f_0}(\epsilon) = \{z \in [0,1]^d : z_1^p + \ldots + z_d^p \leqslant \epsilon/a_0\}$. Subsequently,

$$\left[0, \left(\frac{\epsilon}{a_0 d}\right)^{1/p}\right]^d \subseteq L_{f_0}(\epsilon) \subseteq \left[0, \left(\frac{\epsilon}{a_0}\right)^{1/p}\right]^d. \tag{88}$$

Therefore,

$$[\epsilon/(a_0 d)]^{dp} \leqslant \mu_{f_0}(\epsilon) \leqslant [\epsilon/a_0]^{dp}. \tag{89}$$

Because $a_0, d$ are constants and $dp = \beta$, we established $\mu_{f_0}(\epsilon) \asymp \epsilon^\beta$ for $\beta = dp$.

Finally, note that for any $\epsilon > 0$, $L_{f_0}(\epsilon)$ is sandwiched between two cubics whose volumes only differ by a constant. This proves (A2) and (A2') on the covering and packing numbers of $L_{f_0}(\epsilon)$. □

*Proof of Proposition 5.* By the Chernoff bound and the union bound, with probability $1 - O(n^{-1})$ uniformly over all $x \in G_n$, there are $\Omega(\sqrt{n_0}\log^2 n)$ uniform samples in $B^\infty_{h_0}(x; \mathcal{X})$. Because $h_0 \leqslant \zeta$ for sufficiently large $n_0$ ($\zeta$ is defined in condition (A1)), by Lemma 1 it holds that

$$\left|\check{f}_x(x') - f_x(x')\right| \lesssim h_0^\alpha + n_0^{-1/4} \lesssim n_0^{-\alpha/2d} + n_0^{-1/4}, \qquad \forall x \in G_n,\ x' \in B^\infty_{h_0}(x; \mathcal{X}). \tag{90}$$

Also, using the standard Gaussian concentration inequality, with probability $1 - O(n^{-1})$ we have

$$\begin{aligned} &\inf_{x' \in B^\infty_{h_0}(x;\mathcal{X})} f(x) - O(n_0^{-1/4}) \\ &\leqslant \overline{f}(x) \leqslant \sup_{x' \in B^\infty_{h_0}(x;\mathcal{X})} f(x) + O(n_0^{-1/4}) \qquad \forall x \in G_n. \quad (91) \end{aligned}$$

Let $x^*$ be the minimizer of $f$ on $\mathcal{X}$ and $x \in G_n$ such that $\|x - x^*\|_\infty \leqslant h_0$. By (90), we have with probability $1 - O(n^{-1})$ that $\inf_{x' \in B^\infty_{h_0}(x;\mathcal{X})} \check{f}_x(x') \leqslant f^* + O(n_0^{-\alpha/2d} + n_0^{-1/4}) \leqslant f^* + 1/2\log n$, where $f^* = f(x^*)$. Now consider arbitrary $z \in G_n$ such that $B^\infty_{h_0}(z; \mathcal{X}) \cap L_f(\kappa/2) = \varnothing$, meaning that for all $z' \in \mathcal{X}$, $\|z' - z\|_\infty \leqslant h_0$, $f(z') > \kappa/2$. By (90), $\overline{f}(z) \geqslant \kappa/2 - O(n_0^{-1/4}) \geqslant \kappa/2 - 1/2\log n$. Hence when $n_0$ is sufficiently large, $z \notin S'_0$, which is to be demonstrated. □

*Proof of Proposition 7.* The upper bound part of (53) trivially holds because the absolute values of every element in $\psi_{0,1}(z)\psi_{0,1}(z)^\top$ for $z \in \mathcal{X} = [0,1]^d$ is upper bounded by $O(1)$. To prove the lower bound part, we only need to show $\int_{\mathcal{X}} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z)$ is invertible. Assume the contrary. Then there exists $v \in \mathbb{R}^D \backslash \{0\}$ such that

$$\begin{aligned} &v^\top \left[\int_{\mathcal{X}} \psi_{0,1}(z)\psi_{0,1}(z)^\top \mathrm{d}U(z)\right] v \\ &= \int_{\mathcal{X}} \left|\psi_{0,1}(z)^\top v\right|^2 \mathrm{d}U(z) = 0. \quad (92) \end{aligned}$$

Therefore, $\langle \psi_{0,1}(z), v\rangle = 0$ almost everywhere on $z \in [0,1]^d$. Because $h > 0$, by re-scaling with constants this implies the existence of non-zero coefficient vector $\xi$ such that

$$P(z_1, \dots, z_m) := \sum_{\alpha_1 + \ldots + \alpha_m \leqslant k} \xi_{\alpha_1, \ldots, \alpha_m} z_1^{\alpha_1} \ldots z_m^{\alpha_m} = 0$$

$$\text{almost everywhere on } z \in [0,1]^d.$$

We next use induction to show that, for any degree-$k$ polynomial $P$ of $s$ variables $z_1, \ldots, z_s$ that has at least one non-zero coefficient, the set $\{z_1, \ldots, z_s \in [0,1]^d : P(z_1, \ldots, z_s) = 0\}$ must have zero measure. This would then result in the desired contradiction. For the base case of $s = 1$, the fundamental theorem of algebra asserts that $P(z_1) = 0$ can have at most $k$ roots, which is a finite set and of measure 0.

We next consider the case where $P(z_1, \ldots, z_s)$ takes on $s$ variables. Re-organizing the terms we have

$$P(z_1, \ldots, z_s) \equiv P_0(z_1, \ldots, z_{s-1}) + z_s P_1(z_1, \ldots, z_{s-1}) + \ldots + z_s^k P_k(z_1, \ldots, z_{s-1}), \quad (93)$$

where $P_1, \ldots, P_k$ are degree-$k$ polynomials of $z_1, \ldots, z_{s-1}$. Because $P$ has a non-zero coefficient, at least one $P_j$ must also have a non-zero coefficient. By the inductive hypothesis, the set $\{z_1, \ldots, z_{s-1} : P_j(z_1, \ldots, z_{s-1})\}$ has measure 0. On the other hand, if $P_j(z_1, \ldots, z_{s-1}) \neq 0$, then invoking the fundamental theorem of algebra again on $z_s$ we know that there are finitely many $z_s$ such that $P(z_1, \ldots, z_s) = 0$. Therefore, $\{z_1, \ldots, z_s : P(z_1, \ldots, z_s) = 0\}$ must also have measure zero. □

## References


[1] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2006.

[2] B. Reeja-Jayan, K. L. Harrison, K. Yang, C.-L. Wang, A. E. Yilmaz, and A. Manthiram, "Microwave-assisted low-temperature growth of thin films in solution," *Sci. Rep.*, vol. 2, Dec. 2012, Art. no. 1003.

[3] N. Nakamura, J. Seepaul, J. B. Kadane, and B. Reeja-Jayan, "Design for low-temperature microwave-assisted crystallization of ceramic thin films," *Appl. Stochastic Models Bus. Ind.*, vol. 33, no. 3, pp. 314–321, 2017.

[4] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics). New York, NY, USA: Springer, 2009.

[5] J. Fan and I. Gijbels, *Local Polynomial Modelling and its Applications*. Boca Raton, FL, USA: CRC Press, 1996.

[6] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *J. Mach. Learn. Res.*, vol. 12, pp. 2879–2904, Oct. 2011.

[7] J. Scarlett, I. Bogunovic, and V. Cevher, "Lower bounds on regret for noisy Gaussian process bandit optimization," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2017, pp. 1723–1742.

[8] E. Hazan, A. Klivans, and Y. Yuan, "Hyperparameter optimization: A spectral approach," 2017, *arXiv:1706.00764*. [Online]. Available: https://arxiv.org/abs/1706.00764#

[9] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*. Hoboken, NJ, USA: Wiley, 1983.

[10] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proc. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2005, pp. 385–394.

[11] A. Agarwal, O. Dekel, and L. Xiao, "Optimal algorithms for online convex optimization with multi-point bandit feedback," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2010, pp. 28–40.

[12] K. G. Jamieson, R. Nowak, and B. Recht, "Query complexity of derivative-free optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 2672–2680.

[13] A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," *SIAM J. Optim.*, vol. 23, no. 1, pp. 213–240, 2013.

[14] S. Bubeck, Y. T. Lee, and R. Eldan, "Kernel-based methods for bandit convex optimization," in *Proc. 49th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, 2017, pp. 72–85.

[15] A. H. G. R. Kan and G. T. Timmer, "Stochastic global optimization methods part I: Clustering methods," *Math. Program.*, vol. 39, no. 1, pp. 27–56, 1987.

[16] A. H. G. R. Kan and G. T. Timmer, "Stochastic global optimization methods part II: Multi level methods," *Math. Program.*, vol. 39, no. 1, pp. 57–78, 1987.

[17] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "X-armed bandits," *J. Mach. Learn. Res.*, vol. 12, pp. 1655–1695, May 2011.

[18] C. Malherbe, E. Contal, and N. Vayatis, "A ranking approach to global optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016. AQ:2

[19] C. Malherbe and N. Vayatis, "Global optimization of Lipschitz functions," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.

[20] R. D. Kleinberg, "Nearly tight bounds for the continuum-armed bandit problem," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2005, pp. 697–704.

[21] S. Minsker, "Estimation of extreme values and associated level sets of a regression function via selective sampling," in *Proc. Conf. Learn. Theory (COLT)*, 2013, pp. 105–121.

[22] J.-B. Grill, M. Valko, and R. Munos, "Black-box optimization of noisy functions with unknown smoothness," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 667–675.

[23] S. Minsker, "Non-asymptotic bounds for prediction problems and density estimation," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, Georgia, 2012.

[24] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *Ann. Math. Stat.*, vol. 23, no. 3, pp. 462–466, 1952.

[25] E. Purzen, "On estimation of a probability density and mode," *Ann. Math. Statist.*, vol. 39, no. 3, pp. 1065–1076, 1962.

[26] H. Chen, "Lower rate of convergence for locating a maximum of a function," *Ann. Statist.*, vol. 16, no. 3, pp. 1330–1334, 1988.

[27] Z. B. Zabinsky and R. L. Smith, "Pure adaptive search in global optimization," *Math. Program.*, vol. 53, no. 1, pp. 323–338, 1992.

[28] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," *J. Comput. Syst. Sci.*, vol. 75, no. 1, pp. 78–89, 2009.

[29] S. Dasgupta, D. J. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 353–360.

[30] S. Hanneke, "A bound on the label complexity of agnostic active learning," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, 2007, pp. 353–360.

[31] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, pp. 1079–1105, Jun. 2006.

[32] W. Polonik, "Measuring mass concentrations and estimating density contour clusters-an excess mass approach," *Ann. Statist.*, vol. 23, no. 3, pp. 855–881, 1995.

[33] P. Rigollet and R. Vert, "Optimal rates for plug-in estimators of density level sets," *Bernoulli*, vol. 15, no. 4, pp. 1154–1178, 2009.

[34] A. Singh, C. Scott, and R. Nowak, "Adaptive Hausdorff estimation of density level sets," *Ann. Statist.*, vol. 37, no. 5B, pp. 2760–2782, 2009.

[35] K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. V. Luxburg, "Consistent procedures for cluster tree estimation and pruning," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7900–7912, Dec. 2014.

[36] S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman, "Cluster trees on manifolds," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2679–2687.

[37] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, 2006.

[38] E. Hazan, K. Levy, and S. Shalev-Shwartz, "Beyond convexity: Stochastic quasi-convex optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 1594–1602.

[39] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—Online stochastic gradient for tensor decomposition," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2015, pp. 797–842.

[40] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, "Finding approximate local minima faster than gradient descent," in *Proc. 49th Annu. ACM SIGACT Symp. Theory Comput. (STOC)*, 2017, pp. 1195–1199.

[41] Y. Carmon, O. Hinder, J. C. Duchi, and A. Sidford, "'Convex until proven guilty': Dimension-free acceleration of gradient descent on non-convex functions," 2017, *arXiv:1705.02766*. [Online]. Available: https://arxiv.org/abs/1705.02766

[42] Y. Zhang, P. Liang, and M. Charikar, "A hitting time analysis of stochastic gradient Langevin dynamics," in *Proc. Annu. Conf. Learn. Theory (COLT)*, 2017, pp. 1–43.

[43] Y. Zhu, S. Chatterjee, J. Duchi, and J. Lafferty, "Local minimax complexity of stochastic convex optimization," in *Proc. NIPS*, 2016, pp. 3431–3439.

[44] J. Duchi and F. Ruan, "Asymptotic optimality in stochastic optimization," 2016, *arXiv:1612.05612*. [Online]. Available: https://arxiv.org/abs/1612.05612

[45] A. Locatelli and A. Carpentier, "Adaptivity to smoothness in X-armed bandits," in *Proc. Conf. Learn. Theory (COLT)*, 2018, pp. 1463–1492.

[46] A. W. van der Vaart, *Asymptotic Statistics*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[47] C. Jin, L. T. Liu, R. Ge, and M. I. Jordan, "On the local minima of the empirical risk," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1–10.

[48] T. T. Cai and M. G. Low, "An adaptation theory for nonparametric confidence intervals," *Ann. Statist.*, vol. 32, no. 5, pp. 1805–1840, 2004.

[49] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*, vol. 82. Springer, 2012.

[50] R. M. Castro and R. D. Nowak, "Minimax bounds for active learning," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2339–2353, May 2008.

[51] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Proc. Int. Conf. Algorithmic Learn. Theory (ALT)*, 2009, pp. 23–37.

[52] O. V. Lepski, E. Mammen, and V. G. Spokoiny, "Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors," *Ann. Statist.*, vol. 25, no. 3, pp. 929–947, 1997.

[53] W. K. Newey, "Convergence rates and asymptotic normality for series estimators," *J. Econometrics*, vol. 79, no. 1, pp. 147–168, 1997.

[54] D. S. Ebert, *Texturing & Modeling: A Procedural Approach*. San Mateo, CA, USA: Morgan Kaufmann, 2003.

[55] R. M. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, vol. 20, no. 4, pp. 2217–2246, 2014.

[56] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

[57] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electron. Commun. Probab.*, vol. 17, no. 52, pp. 1–6, 2012.

[58] J. A. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, nos. 1–2, pp. 1–230, 2015.

**Yining Wang** received the B.Eng. degree in computer science and technology in 2014 from Tsinghua University, Beijing China, the M.S. degree in machine learning in 2017 from Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Ph.D. student in machine learning in the machine learning department at Carnegie Mellon University, Pittsburgh, PA, USA. His research interests are primarily in statistical machine learning, with emphasis on interactive methods, active learning, adaptive sampling.

**Sivaraman Balakrishnan** is an Assistant Professor in the Department of Statistics and Data Science at Carnegie Mellon University. Prior to this he received his Ph.D. from the School of Computer Science at Carnegie Mellon University and was a postdoctoral researcher in the Department of Statistics at UC Berkeley. His Ph.D. work was supported by several fellowships including the Richard King Mellon Fellowship and a grant from the Gates Foundation. He is broadly interested in problems that lie at the interface between computer science and statistics. Some particular areas that have provided motivation for his past and current research include the applications of statistical methods in ranking problems, computational biology, clustering, topological data analysis, nonparametric statistics, robust statistics and non-convex optimization.

**Aarti Singh** received the B.E. degree in electronics and communication engineering from the University of Delhi, New Delhi, India, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Wisconsin–Madison, Madison, WI, USA, in 2003 and 2008, respectively. She was a Postdoctoral Research Associate at the Program in Applied and Computational Mathematics, Princeton University, from 2008 to 2009, before joining the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, where she has been an Associate Professor since 2009. Her research interests include the intersection of machine learning, statistics and signal processing, and focus on designing statistically and computationally efficient algorithms that can leverage inherent structure of the data in the form of clusters, graphs, subspaces, and manifold using direct, compressive, and active queries. Her work is recognized by the NSF Career Award, the United States Air Force Young Investigator Award, A. Nico Habermann Faculty Chair Award, Harold A. Peterson Best Dissertation Award, and a best student paper award at Allerton.