Proceedings of Machine Learning Research vol 75:1–98, 2018 31st Annual Conference on Learning Theory

Minimax Bounds on Stochastic Batched Convex Optimization

John Duchi JDUCHI@STANFORD.EDU
Feng Ruan FENGRUAN@STANFORD.EDU
Stanford University

Chulhee Yun CHULHEEY@MIT.EDU
Massachusetts Institute of Technology

Editors: Sebastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We study the stochastic batched convex optimization problem, in which we use many *parallel* observations to optimize a convex function given limited rounds of interaction. In each of M rounds, an algorithm may query for information at n points, and after issuing all n queries, it receives unbiased noisy function and/or (sub)gradient evaluations at the n points. After M such rounds, the algorithm must output an estimator. We provide lower and upper bounds on the performance of such batched convex optimization algorithms in zeroth and first-order settings for Lipschitz convex and smooth strongly convex functions. Our rates of convergence (nearly) achieve the fully sequential rate once $M = O(d \log \log n)$, where d is the problem dimension, but the rates may exponentially degrade as the dimension d increases, in distinction from fully sequential settings.

Keywords: Stochastic convex optimization, batched optimization, parallel computing

1. Introduction

Moore's law on the increasing speeds of computer processors, for reasons of basic physics, energy consumption, and area, is no longer true: computer clock speeds are no longer increasing (Fuller and Millett, 2011). As a consequence, processor manufacturers and algorithm designers have moved toward increased parallelism, with reduced communication among processors, as the way to continue to see increased computing performance (Fuller and Millett, 2011; Ballard et al., 2011). This has had wide-ranging influences, most saliently for us in the context of optimization, where a number of researchers, including Dekel et al. (2012), Duchi et al. (2012), and Niu et al. (2011), show how leveraging parallelism to compute many stochastic (sub)gradients of convex functions simultaneously during iterations of stochastic gradient-based procedures yields faster convergence.

In this paper, we attempt to delineate the tradeoffs between parallelism and sequential computation in stochastic optimization, providing upper and lower bounds on the convergence rates for algorithms as a function of the number of *rounds* of computation they may complete. To make this more precise, consider the problem of minimizing a convex function f subject to the constraint that $x \in \mathbf{D} \subset \mathbb{R}^d$, where $\mathbf{D}$ is a closed compact convex set. We consider algorithms based on noisy zeroth- or first-order oracles, which proceed iteratively by querying a point x, and then receive (conditional on x) either an unbiased estimate of $f(x)$ (the zeroth-order case) or an unbiased estimate of some $g \in \partial f(x)$ (the first-order case). Stochastic optimization procedures proceed in iterative *batches*, where in each batch, one chooses a set of points $x_1, \ldots, x_n$ at which to query the function f, receives the information about f, and then the algorithm may choose the next batch of points.

Given the growing expense of sequential computation as opposed to parallel computation, it is thus of interest to understand more precisely what the tradeoffs are between the number of batches—or rounds of interaction—and their size. Currently, the algorithms we develop are of intellectual rather than practical interest, but we hope that this investigation is a stepping stone toward a deeper understanding of sequential versus parallel optimization methods.

Perchet et al. (2016) inspired our interest in this problem with work on a more classical statistical setting: estimating the effect of a medical treatment. Perchet et al. study the *batched bandit* problem, where, motivated by multi-stage trials in medical settings, they ask the following: given noisy observations from distributions with means $\{\mu^{(1)}, \mu^{(2)}\}$, what is the regret of a procedure that may only update its strategy a small number of times? Perchet et al. (2016) show that in a two-armed bandit problem with n observations, $O(\log \log n)$ batches is sufficient to achieve optimal regret.

We consider a different problem, as we do not study regret: we study only stochastic optimization, where the optimizer need only output some estimate $\widehat{x}$ such that the optimality gap

$$f(\widehat{x}) - \inf_{x^\star \in \mathbf{D}} f(x^\star) \tag{1}$$

is small; we do not care which points are queried during iterations of the algorithm, and we do not measure the sequential error or regret $\sum_i f(x_i) - \inf_{x^\star \in \mathbf{D}} f(x^\star)$. This problem differs substantially from the linear bandits case, and deriving near optimal algorithms (as well as proving lower bounds) is harder. Indeed, if all we care about is stochastic accuracy, the linear problem that underpins the typical multi-armed bandit problem is quite solvable—at least in terms of achieving accuracy that is optimal in the sample size n, ignoring dimensional issues. To make this clear, consider the 2-dimensional setting, where $f(x) = \langle \mu, x \rangle$ for some vector $\mu \in \mathbb{R}^2$ that we assume satisfies $\|\mu\|_\infty \le 1$, and we wish to minimize f over the simplex $\mathbf{D} = \{x \in \mathbb{R}^d_+ \mid \sum_j x_j = 1\}$ given observations of the form $f(x) + \varepsilon$ for sub-Gaussian, mean-zero noise ε. Then we sample $x_1, \ldots, x_n$ alternating between the 2 basis vectors $\mathbf{e}_1, \mathbf{e}_2$, observing $y_i = f(x_i) + \varepsilon_i$. Letting $\widehat{\mu}_j = \frac{2}{n} \sum_{i : x_i = \mathbf{e}_j} y_i$ and defining the estimator $\widehat{j} = \operatorname{argmin}_j \{\widehat{\mu}_j\}$, then noting that $te^{-\frac{\alpha}{2}t2} \le \sqrt{e/\alpha}$, we obtain

$$\mathbb{E}[f(e_{\widehat{j}}) - \min_j f(e_j)] = |\mu_1 - \mu_2| \cdot \mathbb{P}(\widehat{j} \neq 1) \le |\mu_1 - \mu_2| \cdot \exp\left(-\frac{n|\mu_1 - \mu_2|^2}{4}\right) \le \sqrt{\frac{2e}{n}}.$$

We can thus solve the linear stochastic *optimization* problem with no rounds of interaction.

In contrast with the linear case, we show that for general *convex* optimization, the number of rounds of interaction to solve convex problems even to accuracy $n^{-\frac{1}{2}}$ must scale at least as $d \log \log n$ when n is the total number of observations. We shall be more precise in the coming sections, but roughly, our results are as follows. We work in an oracle model of optimization (Nemirovski and Yudin, 1983) where in each of M rounds, the algorithm may query n points $x_1, \ldots, x_n \in \mathbf{D}$. After issuing all of the queries, the algorithm receives noisy function evaluations $f(x_i) + \varepsilon_i$ (a zeroth-order oracle) or noisy (sub)gradient evaluations g_i satisfying $\mathbb{E}[g_i] \in \partial f(x_i)$ (the first-order oracle). After M such rounds, the algorithm must output an estimator $\widehat{x}$. For a given information oracle (noisy function or subgradient evaluations), we evaluate the performance of the algorithm in a worst-case sense as $\sup_{f \in \mathcal{F}} \mathbb{E}[f(\widehat{x})] - \inf_{x^\star \in \mathbf{D}} f(x^\star)$, where $\mathbb{E}$ denotes the expectation taken over randomness in the algorithm and in the noisy evaluation oracle, and $\mathcal{F}$ is a collection of convex functions defined on $\mathbf{D}$. We provide a number of lower and upper bounds on these quantities, but roughly, we show that for the case of zeroth-order oracles, for any (possibly

randomized) algorithm using M rounds with n function computations in each round,

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}[f(\widehat{x})] - \inf_{x^\star \in \mathbf{D}} f(x^\star) \right\} \geq c \cdot n^{-\frac{1}{2}\left(1 - \left(\frac{d}{d+2\kappa}\right)^M\right)} \tag{2}$$

where $\mathcal{F}$ is either the class of 1-Lipschitz convex functions (in which case $\kappa = 1$) or 1-strongly-convex and $O(1)$-strongly smooth functions (in which case $\kappa = 2$), and c is a constant depending on problem parameters. In the first order case, we show a similar result, except the lower bound in the strongly convex case becomes $n^{-(1-(\frac{d}{d+2(\kappa-1)})^M)}$. Let us perform an asymptotic comparison, in which d is held fixed as $n \to \infty$. The gold-standard in such cases is for fully sequential algorithms that receive n queries, in which case the minimax rates for the two settings scale (ignoring polynomials in dimension d) as $n^{-\frac{1}{2}}$ and n^{-1}, respectively (Agarwal et al., 2012; Shamir, 2013); achieving these comparatively good rates requires a number of rounds scaling as $\frac{d}{\kappa} \log \log n$.

We are not the first to study these questions of interactivity and sequential versus batch access in the case of convex optimization. Perchet et al. (2016) study the problem in its most natural statistical setting, as the design of experiments in a bandit problem, and provide a comprehensive literature review. Much of the statistical and medical literature on batched sample access focuses on testing hypotheses: can we determine which treatment (of a set of treatments) is best, or at least reject a null hypothesis with a desired *a priori* power (Dantzig, 1940; Stein, 1945); Hardwick and Stout (2002) provide an elegant treatment of multi-stage experimental design. More recent work in the statistical learning theory literature focuses on so-called "switching bandits," in which an algorithm plays a certain strategy and pays a penalty for switching between strategies (Cesa-Bianchi et al., 2013a,b). In most of these cases, the problems are different from general (convex) stochastic optimization, in that one has a linear function or hypothesis to test (the standard multi-armed bandit scenario), and one must control the regret rather than the optimality gap (1). The results of Smith et al. (2017) are related; they study the interaction necessary for locally differentially private estimation. Their results suggest that roughly $\log \frac{1}{\epsilon}$ rounds of function queries are necessary for ϵ-accurate optimization of a d-dimensional convex function, but it is hard to compare this with the current work; most saliently, our lower and upper bounds indicate that the number of rounds (batches) must at least scale linearly in the dimension, though it may be sub-logarithmic in the sample size n. In addition, our lower bound arguments are information-theoretic.

Notation Throughout the paper, we consider the domain $\mathbf{D} = [0,1]^d$. We use $\mathbf{B}_u^p(\delta)$ to denote an ℓ_p ball centered at $u \in \mathbb{R}^d$ with radius δ. As is standard, a δ-*packing* of S with respect to the metric ρ is a set $S' \subset S$ such that for $v, v' \in S'$ with $v \neq v'$, $\rho(v, v') \geq \delta$. A maximal δ-packing is any δ-packing S' of S with maximum cardinality. For integers $a \leq b$, let $a : b := \{a, a+1, \ldots, b\}$. For any $m \in \mathbb{N}$ we use $[m]$ to denote the set $\{1, 2, \ldots, m\}$.

2. Problem Formulation

As desribed in the introduction, we consider convex optimization problems of minimizing a convex objective f over the domain $\mathbf{D} = [0,1]^d$. We consider algorithms that proceed in a fixed number of *batches* or *rounds*, where in each round,the algorithm chooses n points $x_1, x_2, \ldots, x_n$ from the domain $\mathbf{D}$ to query *in parallel*, receiving noisy information about the function f.

Here we formalize the definition of the sequential optimization procedure. Let $M \in \mathbb{N}$ denote the total number of rounds. For each $t \in [M]$, let $X_{1:n}^{(t)} = \{X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)}\}$ be the points the

algorithm queries in round t. We consider the usual noisy oracle model of optimization (Nemirovski and Yudin, 1983; Agarwal et al., 2012), where we represent information available to the algorithm via an *information oracle* $\phi \in \mathcal{O}$ from a family of oracles $\mathcal{O}$. We consider one of two oracle families. The first is the family $\mathcal{O}_0$ of **zeroth-order oracles**, which, when queried at a point $x \in \mathbf{D}$ for the function f, return $Y = f(x) + \varepsilon$, where $\mathbb{E}[\varepsilon \mid x] = 0$. For the **first-order oracle** family $\mathcal{O}_1$, the information consists of the pair (Y, Z), where $\mathbb{E}[Z \mid x] \in \partial f(x)$ is a stochastic subgradient.

Let $Y_{1:n}^{(t)} = \{Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_n^{(t)}\}$ be the (random) received noisy function values at the query points $X_{1:n}^{(t)}$ when the oracle is zeroth-order oracle, and $Z_{1:n}^{(t)}$ the noisy gradient values at $X_{1:n}^{(t)}$. Then a *batched* optimization algorithm A consists of a series of conditional distributions Q, each defined on the space $\mathbf{D}^n$ (within a round, the points $X_{1:n}^{(t)}$ may have arbitrary dependence), of the querying points $X_{1:n}^{(t)}$ for $t \in [T]$. At round t, the conditional $Q^{(t)}$ is defined given all past information $\{X_{1:n}^{(i)}, Y_{1:n}^{(i)}, Z_{1:n}^{(i)}\}_{i=1}^{t-1}$; the algorithm consists of these conditionals and the final conditional distribution Q of the estimate $\widehat{X}$ for the minimizer of the function f given $\{X_{1:n}^{(t)}, Y_{1:n}^{(t)}, Z_{1:n}^{(t)}\}_{t=1}^{M}$. A *batched* optimization algorithm A is representable as the collection of these conditional distributions,

$$\mathsf{A} := \left\{Q(\widehat{X} \mid X_{1:n}^{(1:M)}, Y_{1:n}^{(1:M)})\right\} \cup \bigcup_{t=1}^{M} \left\{Q^{(t)}(X_{1:n}^{(t)} \mid X_{1:n}^{(1:t-1)}, Y_{1:n}^{(1:t-1)})\right\} \tag{3}$$

when the oracle is zeroth-order, or

$$\mathsf{A} := \left\{Q(\widehat{X} \mid X_{1:n}^{(1:M)}, Y_{1:n}^{(1:M)}, Z_{1:n}^{(1:M)})\right\} \cup \bigcup_{t=1}^{M} \left\{Q^{(t)}(X_{1:n}^{(t)} \mid X_{1:n}^{(1:t-1)}, Y_{1:n}^{(1:t-1)}, Z_{1:n}^{(1:t-1)})\right\} \tag{4}$$

when the oracle is first-order.

With this algorithmic setting, we define the risk of an M-round algorithm A, for a given oracle ϕ and function $f : \mathbf{D} \to \mathbb{R}$, by the expected gap

$$\mathcal{R}(\phi, \mathsf{A}, f) = \mathbb{E}_f\left[f(\widehat{X}) - f^\star\right],$$

where the expectation is taken over any randomness in A and the oracle ϕ. We evaluate the performance of an algorithm A in a uniform sense (Wald, 1939) by considering its maximum risk for a collection of functions $\mathcal{F}$. Now, letting $\mathcal{A}_M$ be the collection of all M-batch algorithms (as in Eqs. (3) and (4)), the M-batch minimax risk of the function class $\mathcal{F}$ for an oracle family $\mathcal{O}$ is

$$\mathfrak{M}_M(\mathcal{F}, \mathcal{O}) := \sup_{\phi \in \mathcal{O}} \inf_{\mathsf{A} \in \mathcal{A}_M} \sup_{f \in \mathcal{F}} \mathcal{R}(\phi, \mathsf{A}, f). \tag{5}$$

Throughout this paper, we consider the following important classes of convex functions:

i. The class of λ strongly convex and H strongly smooth functions $\mathcal{F}_{H,\lambda}$,

$$\mathcal{F}_{H,\lambda} = \left\{f : \lambda \left\|x - x'\right\|_2^2 \le \left\langle \nabla f(x) - \nabla f(x'), x - x' \right\rangle \le H \left\|x - x'\right\|_2^2 \text{ for all } x, x' \in \mathbf{D}\right\}.$$

ii. The class of L Lipschitz convex functions $\mathcal{F}_L$,

$$\mathcal{F}_L = \left\{f : \left|f(x) - f(x')\right| \le L \left\|x - x'\right\|_2 \text{ for all } x, x' \in \mathbf{D}\right\}.$$

Finally, the minimax risk (5) and our algorithms are highly dependent on the oracle class $\mathcal{O}$. Throughout this paper, we consider *subgaussian* oracles, defined as follows. Recall the definition.

Definition 1 *A random vector $W \in \mathbb{R}^d$ is σ^2-*subgaussian *if for all $v \in \mathbb{R}^d$ and $t \in \mathbb{R}$ we have*

$$\mathbb{E}\left[\exp(t\langle v, W - \mathbb{E}[W]\rangle)\right] \le \exp\left(\frac{t^2\sigma^2 \left\|v\right\|_2^2}{2}\right).$$

Throughout this paper, we use $\mathcal{O}_0$ and $\mathcal{O}_1$ to denote an (otherwise arbitrary) noise oracle family with the following properties:

i. The zeroth-order oracle class $\mathcal{O}_0$. For any $\phi \in \mathcal{O}_0$, given the query $x \in \mathbf{D}$, the oracle outputs $y = f(x) + \epsilon$, where $\epsilon \in \mathbb{R}$ is (conditionally on x) mean-zero and σ^2-subgaussian.

ii. The first-order oracle class $\mathcal{O}_1$. For any $\phi \in \mathcal{O}_1$, given the query $x \in \mathbf{D}$, the oracle outputs $y = f(x) + \epsilon_1$ and the noisy gradient value $z = g + \epsilon_2$, where $g(x) \in \partial f(x)$ and $\epsilon_2 \in \mathbb{R}^d$ is (conditionally on x) mean-zero and σ^2-subgaussian.

We assume that the noise additions are independent conditional on the query points $X_{1:n}^{(1:M)}$.

3. The big ideas

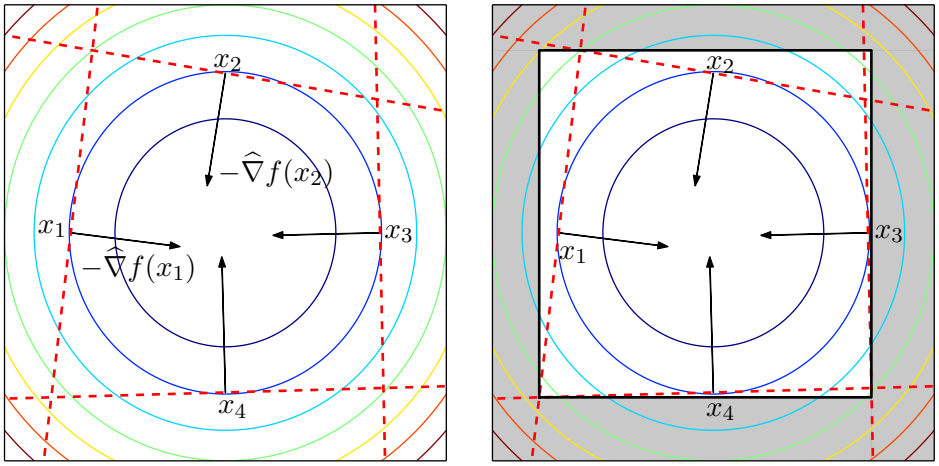
The main contributions of this paper are twofold: (i) the construction of lower bounds for batched convex optimization and (ii) the construction of matching—in terms of the sample size n, though quite far from tight in dimension d—upper bounds through algorithmic developments. At this point, the algorithms we develop should be seen more as intellectual contributions rather than practically useful tools, but we believe it interesting to further understand this area and that these serve as a useful first step in that direction.

3.1. Achieving reasonable convergence rates

We begin by giving a heuristic description of the convergence guarantees we might hope to achieve by describing the idea of the algorithm for smooth strongly convex functions ($\mathcal{F}_{H,\lambda}$) and first-order oracles ($\mathcal{O}_1$). Our algorithms are sequential, in that in each batch or round they collect (stochastic) gradient information, then update and make a decision. Each algorithm maintains a box of radius r_t over iterations t, call this $\mathcal{B}_t = c_t + [-r_t/2, r_t/2]^d$ for some center $c_t \in \mathbb{R}^d$. Then within an iteration, the algorithm chooses a number of points $x_1, x_2, \ldots, x_m$, distributed around B_t, and computes estimated gradients $\widehat{\nabla} f(x_1), \widehat{\nabla} f(x_2), \ldots, \widehat{\nabla} f(x_m)$. If these gradients were accurate, then for each $x \in \{x_1, \ldots, x_m\}$, the standard cutting plane bound guarantees that $x^\star \in \{y \mid \langle \nabla f(x), y - x\rangle \le 0\}$. Of course, we cannot so precisely cut space, but we can guarantee that (with high probability)

$$x^\star \in \{y \mid \langle \widehat{\nabla} f(x), y - x\rangle \le \mathsf{d}(x, y)\}$$

for a distance-like function d. Once we know we can cut off these regions of the space, we may construct a new center c_{t+1} and box $\mathcal{B}_{t+1} = c_t + [-r_{t+1}/2, r_{t+1}/2]^d$ containing $x^\star$ with high probability. For a graphical illustration, see Figure 1, which shows the current box $\mathcal{B}_0$ on the left with approximate gradients and contours of f, and the updated box $\mathcal{B}_1$ on the right, with points that the algorithm has cut shaded gray.



(a) (b)

Figure 1: Illustration of upper bound argument and algorithm on contours of the function $f(x) = \frac{1}{2}\|x\|_2^2$ for $x \in \mathbb{R}^2$. (a) Points $x_1, \ldots, x_4 \in \mathbb{R}^2$ at locations in box, with stochastic gradients $\widehat{\nabla} f(x_i)$ represented for each x. Hyperplanes they support, $\{y \in \mathbb{R}^2 \mid \langle \widehat{\nabla} f(x), y - x \rangle = 0\}$, denoted as dotted red lines. (b) A box containing the points the algorithm is *confident* enough to cut off, containing most of the area allowed by the hyperplanes $\widehat{\nabla} f(x)$ define.

The key insight is that the updated radius r_t is nearly contracting, in that

$$r_{t+1} \le \nu r_t^\beta \tag{6}$$

for some $\beta > 0$ and $\nu < 1$. The recursion (6) for the radius r_t then, by a calculation, implies that

$$r_t \le \nu^{\frac{\beta^t - 1}{\beta - 1}} r_0^{\beta^t} \tag{7}$$

for $\beta \neq 1$, and is $r_t \le \nu^t r_0$ when $\beta = 1$. The size of β of course governs the rate of convergence of the sequence (6), where $\beta > 1$ guarantees superlinear convergence of r_t to zero, whereas if $\beta < 1$, we require ν to be near zero to guarantee that r_t is small enough. Indeed, in our setting, $\beta < 1$, while because of the sampling we perform in each round, the contraction multiplier ν scales as an inverse polynomial in n, so that we may take $\nu = n^{-c}$ for some constant $c > 0$. The question then becomes, compared to the fully sequential case, how large must t be to achieve the "standard" accuracy. In the convex optimization scenarios we consider, the optimal accuracy given n gradient samples is of order n^{-1}, so simplifying expression (7) by assuming $r_0 = 1$, we must find ν sufficiently small and t sufficiently large that $\nu^{\frac{\beta^t - 1}{\beta - 1}} \lesssim n^{-1}$, the typical rate of convergence for convex minimization.

Let us solve for the number of iterations necessary in this case, where we note that at $t = \infty$ we require that $\nu^{-\frac{1}{\beta - 1}} \le n^{-1}$, and for $\nu = n^{-c}$ this implies we must choose our sampling schemes such that $\frac{c}{1 - \beta} \ge 1$. Making this concrete, let us assume we wish to have t large enough that $n^{-c\frac{\beta^t - 1}{\beta - 1}} \le An^{-1}$ for some constant $A > 1$. This occurs if and only if

$$\log A - \log n \ge -c\frac{\beta^t - 1}{\beta - 1} \log n \text{ iff } \left(1 - \frac{c}{1 - \beta}\right) + \frac{c\beta^t}{1 - \beta} \le \frac{\log A}{\log n},$$

which, using that $\frac{c}{1-\beta} \geq 1$, is in turn implied by

$$\beta^t \leq \frac{1-\beta}{c} \frac{\log A}{\log n} \text{ or } t \geq \frac{\log \log n - \log \log A}{-\log \beta} + \frac{\log(1-\beta)}{\log \beta} - \frac{\log c}{\log \beta}.$$

In our sequential optimization setting, the exponents β and c depend on problem dimension, so that $\beta = \frac{d}{d+k}$ for a small integer k. In this case, $-\log \beta = \log \frac{d+k}{d} = \log(1 + \frac{k}{d}) \approx \frac{k}{d}$. For simplicity in intuition, we may set $A = e$, and in the case to which our subsequent analysis applies, we take the sampling exponent $c = 1 - \beta$; the preceding requirement then becomes

$$t \geq \frac{\log \log n}{-\log \beta} \approx \frac{d}{k} \log \log n.$$

That is, after order $d \log \log n$ iterations, it is possible to achieve radius accuracy of order n^{-1}.

3.2. Lower bounds on optimality

Our lower bound construction begins with the familiar Le Cam's method for proving lower bounds in statistical and stochastic optimization (Tsybakov, 2009; Agarwal et al., 2012; Duchi, 2017). The idea is due to Agarwal et al.: given convex functions f and g, the separation between them is

$$\mathsf{d}_{\text{opt}}(f, g) := \inf_{x \in \mathbf{D}} \{f(x) + g(x)\} - \inf_{x^\star \in \mathbf{D}} f(x^\star) - \inf_{x^\star \in \mathbf{D}} g(x^\star).$$

The key to this construction is that if we have a point x optimizing f to accuracy $\frac{1}{2}\mathsf{d}_{\text{opt}}(f, g)$, that is, $f(x) \leq \inf_{x^\star \in \mathbf{D}} f(x^\star) + \frac{1}{2}\mathsf{d}_{\text{opt}}(f, g)$, then $g(x) \geq \inf_{x^\star \in \mathbf{D}} g(x^\star) + \frac{1}{2}\mathsf{d}_{\text{opt}}(f, g)$. Thus, an argument with Markov's inequality (see Agarwal et al. (2012, Eq. (18)) or Duchi (2017, Ch. 4.1)) implies that for any two distributions P_- and P_+, functions $f_-, f_+ \in \mathcal{F}$, and estimator $\widehat{x}$ based on observations from the distribution P_v, we have

$$\mathfrak{M}_M(\mathcal{F}, \mathcal{O}) \geq \max_{v \in \{-,+\}} \mathbb{E}_{P_v}[f_v(\widehat{x}) - f_v^\star] \geq \frac{\mathsf{d}_{\text{opt}}(f_-, f_+)}{4} \left(1 - \|P_- - P_+\|_{\text{TV}}\right), \tag{8}$$

where $\|P_- - P_+\|_{\text{TV}} = \sup_A |P_-(A) - P_+(A)|$ denotes the usual variation distance between distributions. For more details, please refer to Section B.1. Inequality (8) is the starting point of our strategy for lower bounds: we will construct a pair of functions f_- and f_+ for which d_{opt} is reasonably large, but for which the distributions of associated observations are close in $\|\cdot\|_{\text{TV}}$.

Our proof is somewhat delicate, as we must control the amount of interactivity allowed the optimization procedures. We construct functions at multiple scales, where each scale corresponds to a round or batch of data in the method being used for optimization. We do this by first constructing a nested sequence of packings of $\mathbf{D}$ that we use to define our "difficult" functions. For a given multiplier $0 < \eta < 1$ and values $\frac{1}{2} > \delta_1 \geq \delta_2 \geq \cdots \geq \delta_M$, each satisfying $\delta_t \leq \frac{\eta}{4}\delta_{t-1}$, we let the set $\mathcal{U}^{(1)}$ be a maximal $2\delta_1$-packing of the set $[\delta_1, 1 - \delta_1]^d$ with respect to the ℓ_p norm, meaning that for points $u, u' \in \mathcal{U}^{(1)}$, we have $\|u - u'\|_p \geq 2\delta_1$ whenever $u \neq u'$. Additionally, for any vector $u \in \mathbf{D}$, we define the set $\mathcal{U}_u^{(t)}$ as a maximal $2\delta_t$-packing of the ℓ_p ball $\mathbf{B}_u^p(\eta\delta_{t-1} - \delta_t)$ in ℓ_p-norm. Consider the collection of *all* sequences $u_1, u_2, \ldots, u_M$ defined recursively as elements in the chain

$$u_t \in \mathcal{U}_{u_{t-1}}^{(t)}. \tag{9}$$

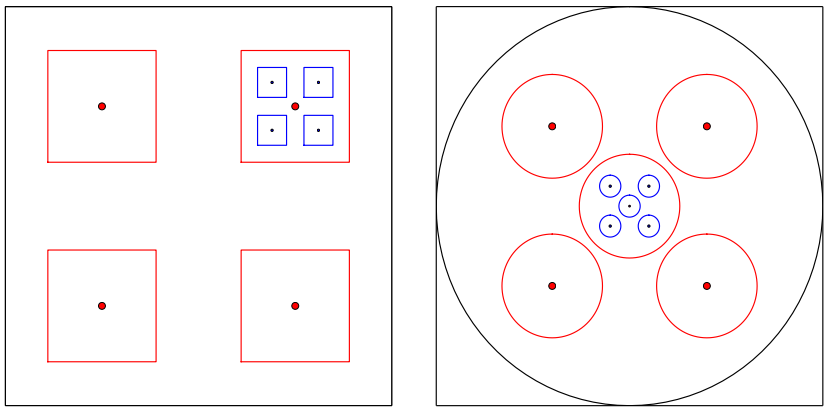


Figure 2: Recursively constructed packing sets $\mathcal{U}^{(1)}$ (red), $\mathcal{U}_{u_1}^{(2)}$ (blue). Left: packing in ℓ_∞-norm. Right: packing in ℓ_2-norm.

Because $\|u_t - u_t'\|_p \geq 2\delta_t$ for any pair $u_t \neq u_t' \in \mathcal{U}_{u_{t-1}}^{(t)}$, all of the balls $\mathbf{B}_{u_t}^p(\eta\delta_t)$ are disjoint, and a volume argument (Lemma 8) shows that the cardinality of $\mathcal{U}_{u_{t-1}}^{(t)}$ is at least $(\frac{\eta\delta_{t-1}}{4\delta_t})^d$. See Figure 2 for an illustration of this sequential construction.

Our idea, similar to one of Smith et al. (2017), is that for any path $u_{1:M}$ in the chain (9), we can construct a pair of functions $f_{u_{1:M}}^{-1}$ and $f_{u_{1:M}}^{+1}$ such that

$$f_{u_{1:M}}^{+1}(x) = f_{u_{1:M}}^{-1}(x) \text{ for all } x \notin \mathbf{B}_{u_M}^p(\delta_M), \text{ and } \mathsf{d}_{\text{opt}}(f_{u_{1:M}}^{+1}, f_{u_{1:M}}^{-1}) \asymp \delta_M^\kappa, \tag{10}$$

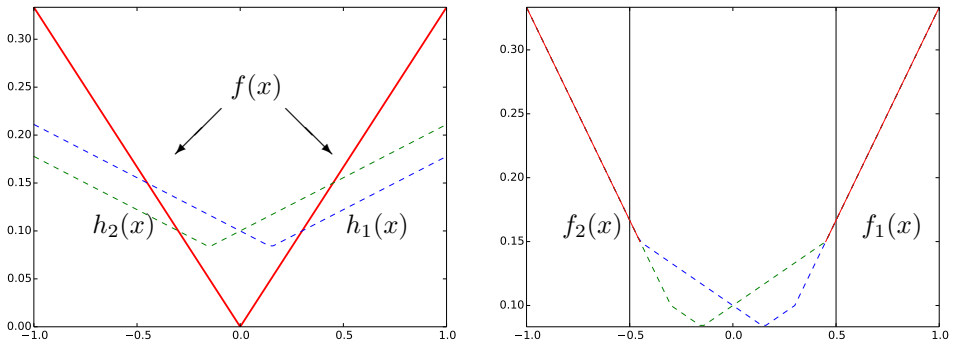
for a constant $\kappa \in \{1, 2\}$, depending on the function class that we consider. Additionally, we have that if $u_{1:M}$ and $\widetilde{u}_{1:M}$ are sequences in our construction (9) for which $u_1 = \widetilde{u}_1, \ldots, u_t = \widetilde{u}_t$ with $u_{t+1} \neq \widetilde{u}_{t+1}$, then the functions are equal except in a δ_t-sized region around u_t, the tth element in the chain, satisfying

$$f_{u_{1:M}}^{v}(x) = f_{\widetilde{u}_{1:M}}^{v'}(x) \text{ for } x \notin \mathbf{B}_{u_t}^p(\delta_t) \text{ and } v, v' \in \{\pm 1\}. \tag{11}$$

To construct the functions, we begin with the base function $f_{u_1}(x) := \|x - u_1\|_\infty$. Then, recursively, we define $h_{u_{1:t}}(x) = a_t \|x - u_t\|_\infty + b_t$ for appropriately chosen scalars α_t and β_t, defining $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\}$. See Figure 3 for an illustration. In the case that we desire our functions to be smooth and strongly convex, we instead begin with $f_{u_1}(x) = \frac{1}{2}\|x - u_1\|_2^2$, and then recurse via a "smoothed" maximum $f_{u_{1:t}}(x) := \textsc{sMax}\{f_{u_{1:t-1}}, a_t \|x - u_t\|_2^2 + b_t\}$. (See Figure 4 for an illustration, and Appendix C for details.) A careful calculation then shows that these functions satisfy our desired properties of function closeness, and even more,

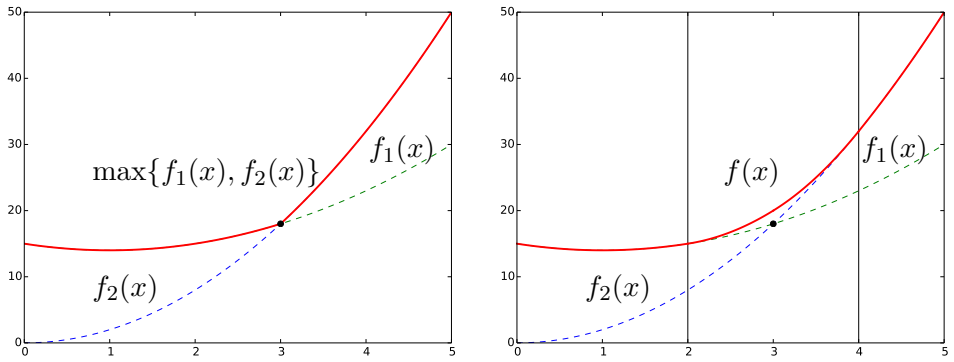
$$|f_{u_{1:M}}^{v}(x) - f_{\widetilde{u}_{1:M}}^{v'}(x)| \leq K_t \cdot \delta_t^\kappa \quad \text{for all } x \in \mathbf{D} \tag{12}$$

whenever $u_{1:t} = \widetilde{u}_{1:t}$, where K_t is a problem-dependant constant. For more details about function construction, please refer to Appendices B.2 and C.



(a) (b)

Figure 3: Construction of Lipschitz convex functions. (a) Function $f(x) = \frac{1}{3}|x|$, $h_1(x) = \frac{1}{9}|x - 0.15| + \frac{1}{12}$ and $h_2(x) = \frac{1}{9}|x + 0.15| + \frac{1}{12}$. All are Lipschitz, with Lipschitz constants $\frac{1}{3}$, $\frac{1}{9}$, and $\frac{1}{9}$, respectively. (b) Functions $f_1(x) = \max\{f(x), h_1(x)\}$ and $f_2(x) = \max\{f(x), h_2(x)\}$. Noticeably, the function $f_1(x)$ and $f_2(x)$ are different only within the region $x \in [-.5, .5]$. Functions f_1 and f_2 are indistinguishable based only on function value/gradient information calculated outside $[-.5, .5]$.



(a) (b)

Figure 4: The *smooth* technique for construction of strongly convex and smooth functions. (a) Function $f_1(x) = (x - 1)^2 + 14$ and $f_2(x) = 2x^2$. (b) A smoothed version of the maximum $\max\{f_1, f_2\}$, with gradients interpolated in the region $x \in [2, 4]$.

3.2.1. THE INFORMATION RECURSION

These functions are then hard to distinguish for iterative procedures: suppose a procedure, by querying the function $f_{u_{1:M}}$, has "identified" $u_{1:t}$, but is oblivious to $u_{t+1:M}$. Then, given batch of n points at which to compute function information, it is possible to distinguish two different functions only if one samples a point near u_{t+1}, which has exponentially small probability. Let us extend this heuristic a bit to give intuition for the lower bounds we prove. Consider a batch-based algorithm, querying n points in computational round t, attempting to distinguish functions $f_{u_{1:t}, u_{t+1}}$ and $f_{u_{1:t}, \widetilde{u}_{t+1}}$. As the functions are identical outside of $\mathbf{B}^p_{u_t}(\delta_t)$, we may consider sampling procedures that without loss of generality sample only in the ball $\mathbf{B}^p_{u_t}(\delta_t)$. Now, consider the amount of information that function evaluation queries can release when function values are perturbed by (say)

mean-zero Gaussian noise. In this case, we know that the difference in function values scales as δ_t^κ by inequality (12), and the KL-divergence

$$D_{\rm kl}\left(\mathsf{N}(f_{u_{1:t},u_{t+1}}(x),1)\|\mathsf{N}(f_{u_{1:t},\widetilde{u}_{t+1}}(x),1)\right)=\frac{1}{2}(f_{u_{1:t},u_{t+1}}(x)-f_{u_{1:t},\widetilde{u}_{t+1}}(x))^2\lesssim\delta_t^{2\kappa}, \quad (13)$$

where $\kappa\in\{1,2\}$ corresponds to the case in which we optimize a Lipschitz convex function ($\kappa=1$) or strongly convex and smooth convex function ($\kappa=2$).

By a careful argument we do not detail here, we can actually consider allocation of points to the slightly larger region $\mathbf{B}_{u_t}(\delta_{t-1})$, dividing $\mathbf{B}_{u_t}(\delta_{t-1})$ into sub-balls of radius δ_t; the number of such regions is $R_t=(\frac{\delta_{t-1}}{\delta_t})^d$ by a volume argument. By the pigeonhole principle, in at least one of these regions, the procedure can collect a sample size of at most n/R_t. In the typical proofs of information-theoretic lower bounds (Tsybakov, 2009; Agarwal et al., 2012; Duchi, 2017), the goal is to choose the separation between distributions, $\|P_0-P_1\|_{\rm TV}$ in Le Cam's method (8), to be a constant so as to guarantee a reasonable lower bound. In this case, recalling the KL-bound (13), we see that the "information" released in round t of a sequential sampling procedure is constant over the least-sampled region whenever

$$\delta_t^{2\kappa}\cdot\frac{n}{R_t}=1.$$

Now, we use our volume argument to note that $R_t=(\delta_{t-1}/\delta_t)^d$, and substituting above, this yields the "information" bound we iterate in our argument, that is,

$$\delta_t^{2\kappa}\cdot\frac{n\delta_t^d}{\delta_{t-1}^d}=1 \;\;\text{or}\;\; \delta_t=n^{-\frac{1}{d+2\kappa}}\delta_{t-1}^{\frac{d}{d+2k}}.$$

By inspection, beginning from $\delta_0=1$, this recursion has the solution

$$\delta_M=n^{-\frac{1}{2\kappa}\left(1-\left(\frac{d}{d+2\kappa}\right)^M\right)}. \quad (14)$$

Of course, this iteration requires very delicate conditioning arguments, which we perform in Appendix B.3. Finally, to prove a lower bound, we require that the functions themselves are separated according to our optimization distance. With that in mind, we also show that our construction satisfies $\mathsf{d}_{\rm opt}(f^{+1}_{u_{1:M}},f^{-1}_{u_{1:M}})\geq\delta_M^\kappa$, where as usual, $\kappa\in\{1,2\}$ corresponds to the Lipschitz or strongly convex case. Thus, we can find the optimum of f to accuracy only δ_M using M rounds, and the function error must scale as δ_M^κ, which is our desired result.

4. Lower Bounds

With our sketches and "big ideas" implemented, we turn to formal statements of our results. We begin with the lower bounds on the minimax risk (5). We defer the proof of Theorem 2 to appendices A through E, with apologies for the extraordinary length.

Theorem 2 *Consider the case when the domain* $\mathbf{D}=[0,1]^d$.

1. *When the function class* $\mathcal{F}=\mathcal{F}_L$ *and* $M\leq\log\log n/\log\left(1+\frac{2}{d}\right)$, *then there exist constants* $c_1,c_2>0$ *depending solely on* d,σ, *and* L *such that*

$$\mathfrak{M}_M(\mathcal{F}_L,\mathcal{O}_0)\geq c_1 n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^M\right)}e^{-\sqrt{2\log n}}\log^{-c_2}n.$$

2. *When the function class $\mathcal{F} = \mathcal{F}_{H,\lambda}$ and $M \le \log\log n / \log\left(1 + \frac{4}{d}\right)$, then there exist constants $c_1, c_2 > 0$ depending solely on d, σ, H, and λ such that*

$$\mathfrak{M}_M(\mathcal{F}_{H,\lambda}, \mathcal{O}_0) \ge c_1 n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+4}\right)^M\right)} e^{-\sqrt{2\log n}} \log^{-c_2} n$$

3. *When the function class $\mathcal{F} = \mathcal{F}_{H,\lambda}$ and $M \le \log\log n / \log\left(1 + \frac{2}{d}\right)$, then there exist constants $c_1, c_2 > 0$ depending solely on d, σ, H, and λ such that*

$$\mathfrak{M}_M(\mathcal{F}_{H,\lambda}, \mathcal{O}_1) \ge c_1 n^{-\left(1-\left(\frac{d}{d+2}\right)^M\right)} e^{-\sqrt{8\log n}} \log^{-c_2} n.$$

We provide some contextualizing remarks on this result.

1. In Theorem 2, the lower bounds using n observations for M rounds take the form $\tilde{\Omega}(n^{-\gamma})$, where

$$\gamma = \frac{\kappa}{2(\kappa-\zeta)}\left(1 - \left(\frac{d}{d+2(\kappa-\zeta)}\right)^M\right), \tag{15}$$

and $\tilde{\Omega}$ hides leading dimension-dependent constants and sub-polynomial terms in n. The constant κ corresponds to the function class, with $\kappa = 1$ for $\mathcal{F}_L$ and $\kappa = 2$ for $\mathcal{F}_{H,\lambda}$, while $\zeta \in \{0, 1\}$ indicates the order of the information oracle. The theorem shows that the rate is worse than the fully-sequential rate $n^{-\frac{\kappa}{2(\kappa-\zeta)}}$ for M smaller than $\log\log n / \log\left(1 + 2(\kappa-\zeta)/d\right)$.

2. It is also interesting to see how the rate of algorithm depends on d for fixed M. As d increases to infinity, $\frac{d}{d+2(\kappa-\zeta)} \to 1$, meaning that the minimax rates show *exponential degradation* as d increases. We cannot observe this phenomenon in fully sequential problems. (Our lower bounds have dimension-dependent leading constants, which we lave as an important open question.)

3. Our proof technique gives a generic approach of achieving these lower bounds whenever we can construct functions satisfying recursive closeness properties similar to equations (10), (11), and (12). (Condition 5 in Appendix B makes this precise.) If we can construct the set of functions satisfying these, we immediately obtain a lower bound of the form (15).

5. Upper Bounds

We now present our upper bounds for the stochastic *batched* convex optimization problem. We note that, while we present convergence guarantees, these algorithms are impractical, so it remains of substantial interest to understand *practical* but low-round optimization schemes. We can construct algorithms that achieve the lower bounds established in Theorem 2 over the functions $f \in \mathcal{F}$ to within constants, which depend on the dimension d in possibly onerous ways, and sub-polynomial factors in the sample size n. We defer the proof of Theorem 3 into appendices F through J.

Theorem 3 *Consider the case when the domain $\mathbf{D} = [0, 1]^d$. Fix $\delta > 0$.*

1. *When $\mathcal{F} = \mathcal{F}_{H,\lambda}$ and $\mathcal{O} = \mathcal{O}_1$, then there exists an algorithm (detailed in section G) and constants $C_1, C_2 > 0$ depending solely on $d, \sigma, H, \lambda, \delta$ such that, for all $\phi \in \mathcal{O}_1$, $f \in \mathcal{F}_{H,\lambda}$, and $M \le \log\log n / \log(1 + \frac{2}{d}) - C_2 \log\log\log n$, the output $\widehat{X}$ of the algorithm satisfies*

$$\mathbb{P}_{f,\phi}\left(f(\widehat{X}) - f^\star \le C_1 n^{-\left(1-\left(\frac{d}{d+2}\right)^M\right)} \log(n)\right) \ge 1 - \delta.$$

2. *When $\mathcal{F} = \mathcal{F}_{H,\lambda}$ and $\mathcal{O} = \mathcal{O}_0$, then there exists an algorithm (detailed in section H) and constants $C_1, C_2 > 0$ depending solely on $d, \sigma, H, \lambda, \delta$ such that, for all $\phi \in \mathcal{O}_0$, $f \in \mathcal{F}_{H,\lambda}$, and $M \le \log\log n / \log(1 + \frac{4}{d}) - C_2 \log\log\log n$, the output $\widehat{X}$ of the algorithm satisfies*

$$\mathbb{P}_{f,\phi}\left(f(\widehat{X}) - f^\star \le C_1 n^{-\frac{1}{2}\left(1 - \left(\frac{d}{d+4}\right)^M\right)} \log n\right) \ge 1 - \delta.$$

3. *When $\mathcal{F} = \mathcal{F}_L$ and $\mathcal{O} = \mathcal{O}_1$, then in the case when $d = 1$, there exists an algorithm (detailed in section I) and constant $C > 0$ depending solely on σ, L, δ such that, for all $\phi \in \mathcal{O}_1$, $f \in \mathcal{F}_L$, and $M \ge 1$, the output $\widehat{X}$ of the algorithm satisfies*

$$\mathbb{P}_{f,\phi}\left(f(\widehat{X}) - f^\star \le C n^{-\frac{1}{2}} \log n\right) \ge 1 - \delta.$$

4. *When $\mathcal{F} = \mathcal{F}_L$ and $\mathcal{O} = \mathcal{O}_0$, then,*

   (a) *in the case when $d = 1$, there exists an algorithm (detailed in section J) and constants $C_1, C_2 > 0$ depending solely on σ, L, δ such that, for all $\phi \in \mathcal{O}_0$, $f \in \mathcal{F}_L$, and $M \le \log\log n / \log 3 - C_2 \log\log\log n$, the output $\widehat{X}$ of the algorithm satisfies*

$$\mathbb{P}_{f,\phi}\left(f(\widehat{X}) - f^\star \le C_1 n^{-\frac{1}{2}\left(1 - \left(\frac{1}{3}\right)^M\right)} \log n\right) \ge 1 - \delta.$$

   (b) *in the case when $M = 1$ (but $d \ge 1$ arbitrary), there exists an algorithm (detailed in section J) and some constant $C > 0$ depending solely on d, σ, L, δ such that, for all $\phi \in \mathcal{O}_0$ and $f \in \mathcal{F}_L$, the output $\widehat{X}$ of the algorithm satisfies*

$$\mathbb{P}_{f,\phi}\left(f(\widehat{X}) - f^\star \le C n^{-\frac{1}{d+2}} (\log n)^{\frac{1}{d+2}}\right) \ge 1 - \delta.$$

We provide some remarks on this theorem before concluding. Theorems 2 and 3 together give a tight minimax rate (up to dimension-dependent constants and sub-polynomial factors of n) for the *batched* convex optimization problems for the strongly convex and smooth function class $\mathcal{F}_{H,\lambda}$ and the 1-dimensional Lipschitz convex class $\mathcal{F}_L$ (for both zeroth and first order oracles). In these settings, Theorems 2 and 3 imply that $M = \tilde{O}(d \log\log n)$ sequential rounds is necessary and sufficient to achieve the fully sequential rate in n.

The theorems also leave open a number of important questions. First, we do not have the tight rates for batched convex optimization on the Lipschitz function class $\mathcal{F}_L$ when $d, M > 1$; it is unclear whether $\tilde{O}(d \log\log n)$ rounds are sufficient to achieve the fully sequential rate. The construction of multi-round algorithms for the smooth and strongly convex function class $\mathcal{F}_{H,\lambda}$ uses the ideas we elaborate in Section 3: the algorithms are recursive algorithms based on ideas of grid search. However, the same idea does not apply to the Lipschitz function class $\mathcal{F}_L$ because local information is less useful for nonsmooth functions. Solving the general Lipschitz function case $\mathcal{F}_L$ calls for new ideas and techniques. Moreover, the algorithms for the proof of Theorem 3 require prior knowledge of the parameters σ, λ, H, L and δ; we have completely ignored questions of adaptivity (which still provide challenge even in fully sequential settings).

References

A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58 (5):3235–3249, 2012.

G. Ballard, J. Demmel, O. Holtz, and O. Schwartz. Minimizing communication in numerical linear algebra. *SIAM Journal on Matrix Analysis and Applications*, 32(3):866–901, 2011.

N. Cesa-Bianchi, O. Dekel, and O. Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems 26*, 2013a.

N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Regret minimization for reserve prices in second-price auctions. In *Proceedings of the Twenty-Fourth ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1190–1204, 2013b.

G. Dantzig. On the non-existence of tests of Student's hypothesis having power functions independent of σ. *The Annals of Mathematical Statistics*, 11(2):186–192, 1940.

O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

J. C. Duchi. Introductory lectures on stochastic convex optimization. In *Park City Mathematics Institute Graduate Summer School: Collected Lectures*. American Mathematical Society (forthcoming), 2017.

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

S. Fuller and L. Millett. *The Future of Computing Performance: Game Over or Next Level?* National Academies Press, 2011.

J. Hardwick and Q. F. Stout. Optimal few-stage designs. *Journal of Statistical Planning and Inference*, 104:121–145, 2002.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

F. Niu, B. Recht, C. Re, and S. Wright. Hogwild: a lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24*, 2011.

V. Perchet, P. Rigollet, S. Chassang, and E. Snowberg. Batched bandit problems. *Annals of Statistics*, 44(2):660681, 2016.

O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the Twenty Sixth Annual Conference on Computational Learning Theory*, pages 3–24, 2013.

A. Smith, A. Thakurta, and J. Upadhyay. Is interaction necessary for distributed private learning? In *IEEE Symposium on Security and Privacy*, 2017.

C. Stein. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16(3):243–258, 1945.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

Appendix A. Proof of Theorem 2: roadmap

In this section, we briefly explain how the proof of Theorem 2 is structured. The proof for the lower bounds is presented in Section B, whose subsections are arranged in the same order as Section 3.2 in the paper. Among them, in B.2, the construction of "difficult-to-distinguish" functions for smooth strongly convex functions is deferred to a separate section (Section C), because the explanation of smooth maximum of quadratic functions took too much space due to technicality.

Since the proof is highly involved, we focused on conveying the main idea of the proof while deferring technical details to separate sections. Sections D and E contain deferred technical proofs from Sections B and C, respectively.

The roadmap for the proof of upper bound (Theorem 3) can be found in F.

Notation for Sections B–E. Throughout Sections B–E, we use $\mathbf{B}_u^p(\delta)$ to denote an ℓ_p ball centered at $u \in \mathbb{R}^d$ with radius δ. As is standard, a δ-*packing* of S with respect to the metric ρ is a set $S' \subset S$ such that for $v, v' \in S'$ with $v \neq v'$, $\rho(v, v') \geq \delta$. A maximal δ-packing is any δ-packing S' of S with maximum cardinality. For integers $a \leq b$, let $a : b := \{a, a+1, \dots, b\}$. Let $\mathbf{1} \in \mathbb{R}^d$ be a d-dimensional vector full with ones, and $\mathbf{e}_j$ be the jth standard basis vector. Given two functions $f, g : \mathbb{R}^d \to \mathbb{R}$, we define the their *intersection set* by

$$\mathbf{Its}(f, g) := \{x \in \mathbb{R}^d \mid f(x) = g(x)\}.$$

For $A \subset \mathbb{R}^d$, $\operatorname{cl} A$ and $\operatorname{int} A$ denote its closure and interior, respectively. Let $\|P_- - P_+\|_{\mathrm{TV}} = \sup_G |P_-(G) - P_+(G)|$ be the usual total variation distance between two distributions P_- and P_+.

Appendix B. Proof of Theorem 2

B.1. The outline

Recall that the minimax error (5), which we want to bound from below, is

$$\mathfrak{M}_M(\mathcal{F}, \mathcal{O}) := \sup_{\phi \in \mathcal{O}} \inf_{\mathsf{A} \in \mathcal{A}_M} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[f(\widehat{X}) - f^\star \right].$$

For proof of lower bound, we are going to consider only oracles $\phi_{\mathsf{N}} \in \mathcal{O}$ with independent Gaussian noise with variance σ^2. With these i.i.d. Gaussian ϕ_{N}, we are going to provide a lower bound for $\sup_{f \in \mathcal{F}} \mathbb{E}_f[f(\widehat{X}) - f^\star]$ that holds for *any* algorithm A, which also is a lower bound for $\mathfrak{M}_M(\mathcal{F}, \mathcal{O})$. To this end, we reduce $\mathcal{F}$ to a subset of two functions $\{f_-, f_+\} \subset \mathcal{F}$. With any such subset and *any* probabilistic event G,

$$\begin{aligned} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[f(\widehat{X}) - f^\star \right] &\geq \max_{v \in \{-,+\}} \mathbb{E}_v \left[(f_v(\widehat{X}) - f_v^\star) \mathbb{I}\{G\} \right] \\ &= \max_{v \in \{-,+\}} \mathbb{P}_v(G) \mathbb{E}_v \left[f_v(\widehat{X}) - f_v^\star \mid G \right], \end{aligned} \tag{16}$$

where $\mathbb{P}_\pm$ and $\mathbb{E}_\pm$ are the probability and expectation given that the true function is $f_\pm$.

As mentioned in Section 3.2, our lower bound construction begins with the Le Cam technique for proving lower bounds in statistical and stochastic optimization problems (Tsybakov, 2009; Agarwal et al., 2012; Duchi, 2017). The idea is due to Agarwal et al.: given two convex functions f_- and f_+, the separation between them is

$$\mathsf{d}_{\mathrm{opt}}(f_-, f_+) := \inf_{x \in \mathbf{D}} \{f_-(x) + f_+(x)\} - \inf_{x^\star \in \mathbf{D}} f_-(x^\star) - \inf_{x^\star \in \mathbf{D}} f_+(x^\star).$$

The key to this construction is that if we have a point x optimizing f_- to accuracy $\frac{1}{2}\mathsf{d}_{\rm opt}(f_-, f_+)$, that is, $f_-(x) \le \inf_{x^\star \in \mathbf{D}} f_-(x^\star) + \frac{1}{2}\mathsf{d}_{\rm opt}(f_-, f_+)$, then $f_+(x) \ge \inf_{x^\star \in \mathbf{D}} f_+(x^\star) + \frac{1}{2}\mathsf{d}_{\rm opt}(f_-, f_+)$. Thus, an argument with Markov's inequality (see Agarwal et al. (2012, Eq. (18)) or Duchi (2017, Ch. 4.1)) gives the following lemma that reduces the optimization problem to a hypothesis testing problem:

Lemma 4 *Let G be any probabilistic event. Then,*

$$\max_{v \in \{-,+\}} \mathbb{E}_v \left[f_v(\widehat{X}) - f_v^\star \mid G \right] \ge \frac{\mathsf{d}_{\rm opt}(f_-, f_+)}{4} \left(1 - \|P_-(\cdot \mid G) - P_+(\cdot \mid G)\|_{\rm TV}\right).$$

This is a standard technique for proving minimax lower bounds, and we will defer the proof of Lemma 4 to the Appendix D.1.

Eq (16) and Lemma 4 is the starting point of our strategy for lower bounds. For any given algorithm A and function class $\mathcal{F}$, we will develop methods that can construct a pair of functions $f_-, f_+ \in \mathcal{F}$ (for which $\mathsf{d}_{\rm opt}(f_-, f_+)$ is reasonably large) such that there exists an event G with

$$\mathbb{P}_v(G) \ge \frac{1}{2 \cdot 4^M} \text{ for } v \in \{-,+\}, \quad \text{and} \quad \|P_-(\cdot \mid G) - P_+(\cdot \mid G)\|_{\rm TV} \le \frac{1}{2}, \tag{17}$$

under the sampling strategies defined by A.

In the following subsections, we will describe in details the methods outlined in Sections 3.2 and 3.2.1. In Section B.2, we provide how to construct pairs of functions $f_-, f_+ \in \mathcal{F}$ at multiple scales, where each scale corresponds to a round or batch of data in the method being used for optimization. Using these construction methods, in Section B.3 we provide a delicate induction argument based on pigeonhole principle and properties of constructed functions that enable us to choose the "difficult" f_- and f_+, and prove Eq (17) for some G. After these are done, we will come back to this point and finish the proof in Section B.4.

B.2. Function construction

For construction of functions that are necessary for our proof, we start by constructing a nested sequence of maximal packings of $\mathbf{D}$ that we use to define our "difficult" functions. Then using the points in the packings as parameters, we construct functions that have desirable properties.

Consider $\delta_1, \delta_2, \ldots, \delta_M$, which are real valued functions n which satisfy $\lim_{n\to\infty} \delta_1(n) = 0$ and $\lim_{n\to\infty} \frac{\delta_t(n)}{\delta_{t-1}(n)} = 0$. Given such δ_t's, a multiplier $0 < \eta < 1$, and a norm ℓ_p, we recursively define a hierarchy of maximal packings as follows:

1. Let $\mathcal{U}^{(1)}$ to be any maximal $2\delta_1$-packing of $[\delta_1, 1-\delta_1]^d$ with respect to ℓ_p norm.

2. For any $u_1 \in \mathcal{U}^{(1)}$, let $\mathcal{U}^{(2)}_{u_1}$ to be any maximal $2\delta_2$-packing of $\mathbf{B}^p_{u_1}(\eta\delta_1 - \delta_2)$ w.r.t. ℓ_p norm.

3. For any $u_2 \in \mathcal{U}^{(2)}_{u_1}$, let $\mathcal{U}^{(3)}_{u_2}$ to be any maximal $2\delta_3$-packing of $\mathbf{B}^p_{u_2}(\eta\delta_2 - \delta_3)$ w.r.t. ℓ_p norm.

   $\vdots$

M. For any $u_{M-1} \in \mathcal{U}^{(M-1)}_{u_{M-2}}$, let $\mathcal{U}^{(M)}_{u_{M-1}}$ to be any maximal $2\delta_M$-packing of $\mathbf{B}^p_{u_{M-1}}(\eta\delta_{M-1} - \delta_M)$ w.r.t. ℓ_p norm.

Let us explain in words what is going on. In the first stage, we construct a set of points $\mathcal{U}^{(1)}$ so that any $u_1 \in \mathcal{U}^{(1)}$ satisfies $\mathbf{B}^p_{u_1}(\delta_1) \subset \mathbf{D}$. Starting from second stage, we construct maximal $2\delta_t$ packings for *all* points in the previous stage; for example, $\mathcal{U}^{(2)}_{u_1}$ is defined for all $u_1 \in \mathcal{U}^{(1)}$ and $\mathcal{U}^{(2)}_{u_1}$ depends on which u_1 we choose. We continue in this recursive way until we define $\mathcal{U}^{(M)}_{u_{M-1}}$. After that stage, we define the final set $\mathcal{V} := \{\pm 1\}$. By construction we have $\mathbf{B}^p_{u_t}(\delta_t) \subset \mathbf{B}^p_{u_{t-1}}(\eta\delta_{t-1})$ for $t \in 2:M$. Also notice that for any $t \in 2:M$, $\mathbf{B}^p_{u_t}(\eta\delta_t) \cap \mathbf{B}^p_{\tilde{u}_t}(\eta\delta_t) = \emptyset$, for different $u_t, \tilde{u}_t \in \mathcal{U}^{(t)}_{u_{t-1}}$. This means that a point $u_t \in \mathcal{U}^{(t)}_{u_{t-1}}$ uniquely maps to all their "ancestors" $u_{t-1}, u_{t-2}, \ldots, u_1$ in the chain, because the neighborhoods of their ancestors never overlap with the other ancestors at the corresponding level.

Given these maximal packings, we can choose $u_1 \in \mathcal{U}^{(1)}$, and then a chain of parameters

$$u_t \in \mathcal{U}^{(t)}_{u_{t-1}} \tag{9}$$

for $t \in 2:M$, and choose $v \in \mathcal{V}$. For this set of parameters $u_{1:M}$ and v, we will define a corresponding function $f^v_{u_{1:M}}(x)$. The functions we construct has the property that a pair of functions with "similar parameters" have the same value outside a small set, while they can only differ inside the set. This construction is crucial for proof of Eq (17), as it makes functions with similar parameter values difficult to distinguish them.

More concretely, we summarize the list of desired properties that $f^v_{u_{1:M}}(x)$ must satisfy, in Condition 5. As we will show, Condition 5 holds for both function classes of interest—L-Lipschitz convex functions ($\mathcal{F}_L$) and H-smooth λ-strongly convex functions ($\mathcal{F}_{H,\lambda}$)—with a class-dependent set of constants $(C, \alpha, \eta, \beta, \kappa, p)$. If the two functions share the same parameter values from u_1 up to u_{t-1} ($t \in 2:M$), but their parameter deviated from each other after t, say $u_t, \ldots, u_M, v$ and $\tilde{u}_t, \ldots, \tilde{u}_M, \tilde{v}$, then the two functions $f^v_{u_{1:M}}(x)$ and $f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)$ are completely identical far away from u_{t-1}, and differ only at points near u_{t-1}. Similar thing happens when parameter values are the same up to level M ($u_{1:M}$) but v and $\tilde{v}$ differ.

Condition 5 *For a given class of functions $\mathcal{F}$, there exist constants $(C, \alpha, \eta, \beta, \kappa, p)$, where $C > 0$, $0 < \alpha < 1$, $0 < \beta < 1$, $0 < \eta < 1$, $\kappa \in \mathbb{N}$, $p \in [2, \infty]$, that satisfy the following statements: Construct the nested packing sets $\mathcal{U}^{(1)}, \ldots, \mathcal{U}^{(M)}_{u_{M-1}}$ w.r.t. ℓ_p norm and choose chains of parameters. For $t \in 2:M$, if we have two chains of parameters $(u_{1:t-1}, u_{t:M}, v)$ and $(u_{1:t-1}, \tilde{u}_{t:M}, \tilde{v})$, the corresponding functions $f^v_{u_{1:M}}$ and $f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}$ satisfy*

1. $f^v_{u_{1:M}}(x) = f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x),\ \forall x \notin \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$.
2. $|f^v_{u_{1:M}}(x) - f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)| \le C(1-\beta)\alpha^{t-2}\delta^\kappa_{t-1},\ \forall x \in \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$.
3. $\left\|\nabla f^v_{u_{1:M}}(x) - \nabla f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)\right\|_2 \le 2\kappa C\alpha^{t-2}\delta^{\kappa-1}_{t-1},\ \forall x \in \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$.

Similarly, for a chain of parameters $u_{1:M}$,

4. $f^{-1}_{u_{1:M}}(x) = f^{+1}_{u_{1:M}}(x),\ \forall x \notin \mathbf{B}^p_{u_M}(\delta_M)$.
5. $|f^{-1}_{u_{1:M}}(x) - f^{+1}_{u_{1:M}}(x)| \le C(1-\beta)\alpha^{M-1}\delta^\kappa_M,\ \forall x \in \mathbf{B}^p_{u_M}(\delta_M)$.
6. $\left\|\nabla f^{-1}_{u_{1:M}}(x) - \nabla f^{+1}_{u_{1:M}}(x)\right\|_2 \le 2\kappa C\alpha^{M-1}\delta^{\kappa-1}_M,\ \forall x \in \mathbf{B}^p_{u_M}(\delta_M)$.
7. $\mathsf{d}_{\mathrm{opt}}(f^{-1}_{u_{1:M}}, f^{+1}_{u_{1:M}}) = 2C\alpha^M\eta^\kappa\delta^\kappa_M$.

Algorithm 1 Construction of Lipschitz convex $f^v_{u_{1:M}}$.

Given parameters $u_1, u_2, \ldots, u_M, v$, size $\delta_1, \ldots, \delta_M$, and Lipschitzness parameter L,
Start with $f_{u_1}(x) = h_{u_1}(x) := L \left\|x - u_1\right\|_\infty$.
for $t = 2$ to M **do**
  $h_{u_{1:t}}(x) := \frac{L}{3^{t-1}} \left\|x - u_t\right\|_\infty + \sum_{m=1}^{t-1} \frac{L\delta_m}{2 \cdot 3^{m-1}}$.
  $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\}$.
end for
$h^v_{u_{1:M}}(x) := \frac{L}{3^M} \left\|x - u_M - \frac{v\delta_M}{2}\mathbf{1}\right\|_\infty + \sum_{m=1}^{M} \frac{L\delta_m}{2 \cdot 3^{m-1}}$.
Return $f^v_{u_{1:M}}(x) := \max\{f_{u_{1:M}}(x), h^v_{u_{1:M}}(x)\}$.

Here, whenever the function $f^v_{u_{1:M}}$ is non-differentiable at x, we replace $\nabla f^v_{u_{1:M}}(x)$ with any sub-gradient $g^v_{u_{1:M}} \in \partial f^v_{u_{1:M}}(x)$.

It now remains to show that we can construct functions $f^v_{u_{1:M}}$ that satisfy Condition 5, for both $\mathcal{F}_L$ and $\mathcal{F}_{H,\lambda}$. We present the construction $\mathcal{F}_L$ in the following subsection. Construction for $\mathcal{F}_{H,\lambda}$ is done in essentially the same way, but *smoothing* the maximum of quadratic functions require additional technicality. To help the proof run smoothly, we separate the construction for $\mathcal{F}_{H,\lambda}$ to a different section (Section C).

B.2.1. FUNCTION CONSTRUCTION: LIPSCHITZ CONVEX FUNCTIONS

For L-Lipschitz convex functions, construct nested maximal packing sets $\mathcal{U}^{(1)}, \ldots, \mathcal{U}^{(M)}_{u_{M-1}}$ w.r.t. ℓ_∞ norm. For any chain $u_{1:M}$ from the packings and $v \in \mathcal{V}$, Algorithm 1 takes those parameters as input and returns the corresponding $f^v_{u_{1:M}}$. We begin with the base function $f_{u_1}(x) := L \left\|x - u_1\right\|_\infty$. Then, recursively, we define $h_{u_{1:t}}(x) = a_t \left\|x - u_t\right\|_\infty + b_t$ for appropriate scalars a_t and b_t, defining $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\}$. For the scalars in Algorithm 1, we can show that the functions $f^v_{u_{1:M}}(x)$ satisfy Condition 5:

Lemma 6 *The functions constructed by Algorithm 1 satisfy Condition 5 with*

$$(C, \alpha, \eta, \beta, \kappa, p) = \left(L, \frac{1}{3}, \frac{1}{2}, \frac{1}{2}, 1, \infty\right).$$

The scalars a_t and b_t are carefully chosen so that in $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\}$, the max operation can only change the function for $x \in \mathbf{B}^\infty_{u_{t-1}}(\delta_{t-1})$; from this, we can show Condition 5.1. Also, since $f^v_{u_{1:M}}(x)$ and $f^{\tilde{v}}_{u_{1:t-1}, \tilde{u}_{t:M}}(x)$ can only differ in a ball of radius δ_{t-1}, their difference should be proportional to δ_{t-1}, implying Condition 5.2. The rest of the proof of Lemma 6 is deferred to Appendix D.2.

We also need to check whether the functions $f^v_{u_{1:M}}(x)$ we are constructing are indeed in our function class of interest: L-Lipschitz convex functions. The following lemma addresses this points, whose simple proof is deferred to Appendix D.3.

Lemma 7 *The functions constructed with Algorithm 1 are L-Lipschitz and convex.*

B.2.2. FUNCTION CONSTRUCTION: SMOOTH STRONGLY CONVEX FUNCTIONS

We can show that for $\mathcal{F}_{H,\lambda}$, we can construct functions $f^v_{u_{1:M}}$ that satisfy Condition 5, with $(\kappa, p) = (2, 2)$. For details, please refer to Section C.

B.3. The information recursion

The proof depends on whether the oracle is zeroth-order ($\mathcal{O}_0$) or first-order ($\mathcal{O}_1$). Let ζ denote the *order of the oracle*: $\zeta = 0$ whenever we are using zeroth-order oracles, $\zeta = 1$ for first-order. So, with κ and ζ, we can express the three cases presented in Theorem 2 into three tuples: $(\kappa, \zeta) = (1, 0)$ for Lipschitz convex/zeroth-order, $(2, 0)$ for smooth strongly convex/zeroth-order, and $(2, 1)$ for smooth strongly convex/first-order. Our proof strategy assumes $\kappa > \zeta$; this means that the proof *does not* apply to $(\kappa, \zeta) = (1, 1)$, which corresponds to Lipschitz convex functions with first-order oracles. Also, let $\mathbb{P}^v_{u_{1:M}}$ and $\mathbb{E}^v_{u_{1:M}}$ denote the probability of an event and expectation of any quantity, respectively, based on the event that the true objective function is $f^v_{u_{1:M}}$.

The rest of our proof is based solely on the assumption that construction of functions $f^v_{u_{1:M}} \in \mathcal{F}$ that satisfy Condition 5 is possible with some constants $(C, \alpha, \eta, \beta, \kappa, p)$ for the function class $\mathcal{F}$. This means that our analysis that follows can be used *universally* for any other function classes once we can construct functions $f^v_{u_{1:M}}$ that satisfy Condition 5.

For the information recursion step, the crucial argument is the pigeonhole principle, i.e., if there are n samples and R disjoint subsets of $\mathbf{D}$, there must be at least one *scarce-sampled subset* that contains at most n/R samples. For any constructed nested packing sets $\mathcal{U}^{(1)}, \ldots, \mathcal{U}^{(M)}_{u_{M-1}}$ and given algorithm A, we will show that we can inductively find a particular chain of parameters $u_{1:M}$ such that all $\mathbf{B}^p_{u_1}(\delta_1), \ldots, \mathbf{B}^p_{u_M}(\delta_M)$ are scarce-sampled subsets (each $\mathbf{B}^p_{u_t}(\delta_t)$ contains at most $n/|\mathcal{U}^{(t)}_{u_{t-1}}|$ sample points) with constant probability. For that particular $u_{1:M}$, it is difficult for A to distinguish between $f^{-1}_{u_{1:M}}$ and $f^{+1}_{u_{1:M}}$, hence leading to small total variation distance between $\mathbb{P}^{-1}_{u_{1:M}}$ and $\mathbb{P}^{+1}_{u_{1:M}}$. This corresponds to proving Eq (17), for the event G that all $\mathbf{B}^p_{u_1}(\delta_1), \ldots, \mathbf{B}^p_{u_M}(\delta_M)$ are scarce-sampled.

B.3.1. MINIMAX RATES: INFORMATION-THEORETIC INTUITION

Let us first provide a rather heuristic argument to give intuition for the "rates" (in n) of lower bounds we prove. Suppose a procedure A, by querying the true function $f^v_{u_{1:M}}$, has "identified" $u_{1:t-1}$, but is oblivious to $u_{t:M}$. Then, given a batch of n points at which to compute function information, it is possible to distinguish two different functions $f^v_{u_{1:M}}$ and $f^{\tilde{v}}_{u_{1:t-1}, \tilde{u}_{t:M}}$ only if one samples a point near u_{t-1}. Consider a batch-based algorithm, querying n points in computational round t, attempting to find the right value for u_t. As the functions are identical outside of $\mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$, we may consider a sampling scheme that without loss of generality sample n points only in the ball $\mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$. By the pigeonhole principle, for at least one u'_t, the procedure can collect a sample size of at most $n/|\mathcal{U}^{(t)}_{u_{t-1}}|$ in $\mathbf{B}^p_{u'_t}(\delta_t)$. Suppose for the sake of worst-case analysis that the true u_t is equal to the scarce-sampled ball u'_t. Now, with n samples in round t the procedure identified u_t, and given $n/|\mathcal{U}^{(t)}_{u_{t-1}}|$ samples in $\mathbf{B}^p_{u_t}(\delta_t)$ one must seek to find the next u_{t+1}.

Now, consider the amount of information about u_{t+1} that function evaluation queries in round t can release when function values and/or (sub)gradients are perturbed by i.i.d mean-zero Gaussian noise (recall our assumption that we consider ϕ_{N}). In this case, by Condition 5.2, we know that the difference in function values scales as δ_t^κ, and the KL-divergence

$$D_{\mathrm{kl}}\left(\mathsf{N}(f^v_{u_{1:M}}(x), \sigma^2) \| \mathsf{N}(f^{\tilde{v}}_{u_{1:t}, \tilde{u}_{t+1:M}}(x), \sigma^2)\right) \asymp (f^v_{u_{1:M}}(x) - f^{\tilde{v}}_{u_{1:t}, \tilde{u}_{t+1:M}}(x))^2 \lesssim \delta_t^{2\kappa}, \qquad (13)$$

when the oracle is zeroth-order ($\zeta = 0$). Similarly, we can show from Condition 5.3 that the information we get in the case of the first-order Gaussian-noise oracle ($\zeta = 1$) is $O(\delta_t^{2(\kappa-1)})$.

In the typical proofs of information-theoretic lower bounds (Tsybakov, 2009; Agarwal et al., 2012; Duchi, 2017), the goal is to choose the separation between distributions, $\|\mathbb{P}^{v}_{u_{1:M}} - \mathbb{P}^{\tilde{v}}_{u_{1:t},\tilde{u}_{t+1:M}}\|_{\mathrm{TV}}$ in Le Cam's method (8), to be a constant so as to guarantee a reasonable lower bound. In this case, recalling the KL-bound (13), we see that the "information" about u_{t+1} revealed in round t of a sequential sampling procedure is constant over the least-sampled region whenever

$$\delta_t^{2(\kappa-\zeta)} \cdot \frac{n}{|\mathcal{U}^{(t)}_{u_{t-1}}|} = \frac{n\delta_t^{d+2(\kappa-\zeta)}}{\delta_{t-1}^{d}} = 1, \;\; \text{or} \;\; \delta_t = n^{-\frac{1}{d+2(\kappa-\zeta)}} \delta_{t-1}^{\frac{d}{d+2(\kappa-\zeta)}}, \tag{18}$$

where $|\mathcal{U}^{(t)}_{u_{t-1}}| = (\frac{\delta_{t-1}}{\delta_t})^d$ was from a volume argument. By inspection, beginning from $\delta_0 = 1$, this recursion (18) has the solution

$$\delta_M = n^{-\frac{1}{2(\kappa-\zeta)}\left(1-\left(\frac{d}{d+2(\kappa-\zeta)}\right)^M\right)}. \tag{19}$$

Note, from Lemma 5.7, that our construction satisfies $\mathsf{d}_{\mathrm{opt}}(f^{-1}_{u_{1:M}}, f^{+1}_{u_{1:M}}) \asymp \delta_M^{\kappa}$. Using Eqs (16), (17), and (19), together with Lemmas 4 and 5.7, we can check the calculated rates agree with Theorem 2 up to sub-polynomial factors, as desired.

In the following subsections, we introduce a set of new notation in Section B.3.2, and then, in Section B.3.3, provide the formal inductive argument that proves Eq (17) for some G.

B.3.2. NOTATION

Using ζ, for example, we can define

$$\tilde{C}_\zeta := C(1-\beta)^{1-\zeta}(2\kappa)^{\zeta}, \tag{20}$$

with which we can simplify the notations quite a bit. For example, Condition 5.2–3 can be now written as

$$\begin{aligned} |f^{v}_{u_{1:M}}(x) - f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)| &\le \tilde{C}_0 \alpha^{t-2}\delta_{t-1}^{\kappa}, \; \forall x \in \mathbf{B}^p_{u_{t-1}}(\delta_{t-1}), \\ \left\|\nabla f^{v}_{u_{1:M}}(x) - \nabla f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)\right\|_2 &\le \tilde{C}_1 \alpha^{t-2}\delta_{t-1}^{\kappa-1}, \; \forall x \in \mathbf{B}^p_{u_{t-1}}(\delta_{t-1}). \end{aligned}$$

Next, define the "exponent" constants (as we saw from the intuition)

$$\gamma_t := \frac{1}{d+2(\kappa-\zeta)} \sum_{m=0}^{t-1} \left(\frac{d}{d+2(\kappa-\zeta)}\right)^m = \frac{1}{2(\kappa-\zeta)}\left(1-\left(\frac{d}{d+2(\kappa-\zeta)}\right)^t\right), \tag{21}$$

for $t \in 0:M$. Using this exponent constants, define $\delta_1, \ldots, \delta_M$ as

$$\begin{aligned} \delta_t &:= D_t n^{-\gamma_t} \exp\left(-2\sqrt{2}\gamma_{t-1}\sqrt{\log n}\right) && \text{for } t = 1:M-1, \\ \delta_M &:= D_M n^{-\gamma_M} \exp\left(-2\sqrt{2}\gamma_{M-1}\sqrt{\log n}\right) \log^{-\nu/\kappa} n, \end{aligned} \tag{22}$$

where $\nu > 0$ is any arbitrarily small number. Note that δ_t's have the same rates in n as seen in Eq (18) up to sub-polynomial factors. Some sub-polynomial factors appear in the process of the proof; whether they are artifact of our proof technique or not is still an open question.

The leading constants are defined recursively as

$$\begin{aligned} D_1 &:= \left(\frac{\sigma^2}{2\cdot 8^d \tilde{C}_\zeta^2}\right)^{\frac{1}{d+2(\kappa-\zeta)}} \\ D_t &:= \left(\frac{D_{t-1}\eta}{4}\right)^{\frac{d}{d+2(\kappa-\zeta)}} \left(\frac{\sigma^2}{8e\tilde{C}_\zeta^2\alpha^{2t-2}}\right)^{\frac{1}{d+2(\kappa-\zeta)}} \qquad \text{for } t \in 2:M. \end{aligned} \tag{23}$$

Solving the recurrence of D_t gives

$$D_t = D_1^{\left(\frac{d}{d+2(\kappa-\zeta)}\right)^{t-1}} \left(\frac{\eta^d\sigma^2}{8e\cdot 4^d\tilde{C}_\zeta^2}\right)^{\frac{1}{d}\sum_{m=1}^{t-1}\left(\frac{d}{d+2(\kappa-\zeta)}\right)^m} \alpha^{-\frac{2}{d}\sum_{m=1}^{t-1}(t-m)\left(\frac{d}{d+2(\kappa-\zeta)}\right)^m}. \tag{24}$$

Also, define

$$h_t := \left(\frac{\sigma}{\tilde{C}_\zeta\alpha^{t-1}D_t^{\kappa-\zeta}}\right)^2 n^{2(\kappa-\zeta)\gamma_t} \exp\left(4\sqrt{2}(\kappa-\zeta)\gamma_{t-1}\sqrt{\log n}\right) \text{ for } t = 1:M, \tag{25}$$

so that

$$\begin{aligned} \tilde{C}_\zeta^2\alpha^{2t-2}h_t\delta_t^{2(\kappa-\zeta)} &= \sigma^2 && \text{for } t \in 1:M-1, \\ \tilde{C}_\zeta^2\alpha^{2M-2}h_M\delta_M^{2(\kappa-\zeta)} &= \sigma^2\log^{-\frac{2\nu(\kappa-\zeta)}{\kappa}} n. \end{aligned} \tag{26}$$

Note that h_t is an increasing quantity as n grows. Also, the exponent of n in h_t is $1-(\frac{d}{d+2(\kappa-\zeta)})^t$, so we have $h_1,\dots,h_M \le n$ for large enough n.

Now that we have h_m's, define events $\Lambda_{u_m}^{(m)}$ as

$$\Lambda_{u_m}^{(m)} := \left\{\sum_{i=1}^n \mathbb{I}\left\{X_i^{(m)} \in \mathbf{B}_{u_m}^p(\delta_m)\right\} \le h_m\right\} \qquad \text{for } m = 1:M.$$

These are probabilistic events that, in the pigeonhole principle argument, this event corresponds to the case that this particular hole $\mathbf{B}_{u_m}^p(\delta_m)$ around u_m has small number of "pigeons" in it. Given the definition of $\Lambda_{u_m}^{(m)}$, we can see from Eq (18) that h_m corresponds to $n/|\mathcal{U}_{u_{m-1}}^{(m)}|$, and Eq (26) is a scaled version of $\delta_t^{2(\kappa-\zeta)} \times$ (# of samples) $= 1$.

For a fixed true function $f_{u_{1:M}}^v$, $\Lambda_{u_m}^{(m)}$ is an event that only a small number h_m of sampled points $X_i^{(m)}$ during the m-th round are in the region $\mathbf{B}_{u_m}^p(\delta_m)$, which contains the global minimum of $f_{u_{1:M}}^v$. So, if this occurs, the amount of information to distinguish between $f_{u_{1:M}}^v$ and other functions $f_{u_{1:m},\tilde{u}_{m+1:M}}^{\tilde{v}}$ is small, so it is hard to optimize $f_{u_{1:M}}^v$ to global optimality. Recall that $\mathbb{P}_{u_{1:M}}^v$ is the probability measure when underlying true function is $f_{u_{1:M}}^v$. So, if $\mathbb{P}_{u_{1:M}}^v(\Lambda_{u_m}^{(m)})$ happens with constant probability, it means that there is some chance for sampling strategy $Q^{(m)} \in \mathsf{A}$ to fail to sample good enough amount of informative sample points.

B.3.3. THE INDUCTIVE ARGUMENT

The key of this part is to show that, for any algorithm A, there *exist* $u_{1:M}$ such that Eq (17) is satisfied when we substitute $\mathbb{P}_- \leftarrow \mathbb{P}_{u_{1:M}}^{-1}$, $\mathbb{P}_+ \leftarrow \mathbb{P}_{u_{1:M}}^{+1}$, and $G \leftarrow \bigcap_{m=1}^M \Lambda_{u_m}^{(m)}$. This is proved in Lemma 11 at the end of a careful inductive argument. This means that if the true objective function is either

one of $f^{+1}_{u_{1:M}}$ or $f^{-1}_{u_{1:M}}$, with constant probability we do not have enough informative samples needed to distinguish between $f^{+1}_{u_{1:M}}$ and $f^{-1}_{u_{1:M}}$, which are reasonably separated with respect to $\mathsf{d}_{\mathrm{opt}}$ (by Condition 5.7), hence leading to non-trivial error in optimization.

Before jumping into the induction, we present a lemma that bounds the packing numbers of sets $\mathcal{U}^{(1)}, \mathcal{U}^{(2)}_{u_1}, \ldots, \mathcal{U}^{(M)}_{u_{M-1}}$. This will prove useful later in the induction.

Lemma 8 *For large enough n, the cardinality of maximal packings satisfy*

$$|\mathcal{U}^{(1)}| \geq \left(\frac{1}{8\delta_1}\right)^d, \text{ and } |\mathcal{U}^{(t)}_{u_{t-1}}| \geq \left(\frac{\eta\delta_{t-1}}{4\delta_t}\right)^d \text{ for } t \in 2:M,$$

regardless of the choice u_{t-1}.

The proof of Lemma 8 is a simple volumetric argument, which is provided in Appendix D.5.

The proof is done by mathematical induction. First we fix any M-stage procedure A (hence the distributions $Q^{(1)}, \ldots, Q^{(M)}$). Starting from $t = 1$ to $t = M$, we prove the statement

> ST_t : There exists a chain of parameters $u_1 \in \mathcal{U}^{(1)}, u_2 \in \mathcal{U}^{(2)}_{u_1}, \ldots, u_M \in \mathcal{U}^{(M)}_{u_{M-1}}$ and $v \in \mathcal{V}$ such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^t \Lambda^{(m)}_{u_m}) \geq \frac{1}{4^t}$ for sufficiently large n.

Base case ($t = 1$). Since the first stage observations are sampled without any information about the true function, the sampling strategies are all identical regardless of the true function $f^v_{u_{1:M}}$. This is, in other words, $\mathbb{P}^v_{u_{1:M}}(X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)) = Q^{(1)}(X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1))$, for any $u_{1:M}$ and v. Recall that interiors of balls $\mathrm{int}(\mathbf{B}^p_{u_1}(\delta_1))$ are disjoint for different $u_1 \in \mathcal{U}^{(1)}$. So, for any fixed $Q^{(1)}$,

$$\sum_{u_1 \in \mathcal{U}^{(1)}} \sum_{i=1}^n \mathbb{P}^v_{u_{1:M}}(X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)) = \sum_{i=1}^n \sum_{u_1 \in \mathcal{U}^{(1)}} Q^{(1)}(X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)) \leq n. \tag{27}$$

Now by the pigeonhole principle, there must exist at least one $u_1 \in \mathcal{U}^{(1)}$ such that

$$\sum\nolimits_{i=1}^n Q^{(1)}(X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)) \leq \tfrac{n}{|\mathcal{U}^{(1)}|}.$$

Now recall $|\mathcal{U}^{(1)}| \geq \left(\frac{1}{8\delta_1}\right)^d$ from Lemma 8. Given u_1, choose the next parameters $u_{2:M}$ and v arbitrarily. Then,

$$\sum_{i=1}^n Q^{(1)}(X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)) = \mathbb{E}^v_{u_{1:M}}\left[\sum_{i=1}^n \mathbb{I}\left\{X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)\right\}\right] \leq (8\delta_1)^d n.$$

Then, for those chosen $u_{1:M}$ and v, by Markov's inequality,

$$\begin{aligned}\mathbb{P}^v_{u_{1:M}}\left((\Lambda^{(1)}_{u_1})^c\right) &= \mathbb{P}^v_{u_{1:M}}\left(\sum_{i=1}^n \mathbb{I}\left\{X^{(1)}_i \in \mathbf{B}^p_{u_1}(\delta_1)\right\} > h_1\right) \leq \frac{(8\delta_1)^d n}{h_1} \\ &= \frac{8^d \tilde{C}^2_\zeta}{\sigma^2}\delta_1^{d+2(\kappa-\zeta)} n = \frac{8^d \tilde{C}^2_\zeta}{\sigma^2} D_1^{d+2(\kappa-\zeta)} n^{-\gamma_1(d+2(\kappa-\zeta))} n = \frac{1}{2},\end{aligned}$$

by definition of D_1 (23) and γ_1 (21). This implies ST_1. In words, for the first-stage sampling strategy $Q^{(1)}$, there exist parameter $u_{1:M}$ and v such that, no more than h_1 sample points are in $\mathbf{B}^p_{u_1}(\delta_1)$ (the region where global optimum lies) with probability at least 1/4.

Inductive step ($2 \le t \le M$)**.** At step t, by the induction hypothesis ST_{t-1} we know that there exist $u_{1:M}$ and v such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m}) \ge \frac{1}{4^{t-1}}$. This means that for the sampling strategies $Q^{(1)}, Q^{(2)}, \ldots, Q^{(t-1)}$, there is constant probability for them all to fail to sample sufficient amount of samples in $\mathbf{B}^p_{u_1}(\delta_1), \mathbf{B}^p_{u_2}(\delta_2), \ldots, \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$. Note that these are the balls containing the global minimizer of $f^v_{u_{1:M}}$.

Given the chain $u_{1:M}$ and v as defined by ST_{t-1}, consider re-choosing the parameters from level t and onwards. That is, we leave the first $t-1$ parameters $u_{1:t-1}$ unchanged, and arbitrarily re-choose the rest of them to define another chain of parameters, say $\tilde{u}_t \in \mathcal{U}^{(t)}_{u_{t-1}}, \tilde{u}_{t+1} \in \mathcal{U}^{(t+1)}_{\tilde{u}_t}, \ldots, \tilde{u}_M \in \mathcal{U}^{(M)}_{\tilde{u}_{M-1}}$ and $\tilde{v} \in \mathcal{V}$. Then, note by Condition 5.1 that $f^v_{u_{1:M}}(x) = f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)$ for all $x \notin \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$. Since the two functions $f^v_{u_{1:M}}(x)$ and $f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)$ look exactly the same outside $\mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$ and it is known that $Q^{(1)}, \ldots, Q^{(t-1)}$ are likely to fail to sample sufficient amount of samples in $\mathbf{B}^p_{u_1}(\delta_1), \ldots, \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$, it is plausible to conjecture that the similar thing might happen to $f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)$ as well. The next lemma formalizes and proves this idea:

Lemma 9 *Suppose there exist $u_{1:M}$ and v such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m}) \ge \frac{1}{4^{t-1}}$ for sufficiently large n (ST_{t-1}). For* **any** *$\tilde{u}_{t:M}$ and $\tilde{v}$ re-chosen as above, the following inequality holds:*

$$\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}\left(\bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m}\right) \ge \frac{1}{2 \cdot 4^{t-1}}, \tag{28}$$

for sufficiently large n.

Notice that this lemma holds for *any* new choice of $\tilde{u}_{t:M}$ and $\tilde{v}$. The proof of Lemma 9 is presented in Appendix D.6.

Next, consider the conditional probability of $\Lambda^{(t)}_{\tilde{u}_t}$ given $\bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m}$, for any re-chosen $\tilde{u}_t \in \mathcal{U}^{(t)}_{u_{t-1}}$. Recall that the sampling strategy $Q^{(t)}$ was fixed before we start our proof, so the conditional distribution $Q^{(t)}(X^{(t)}_{1:n} \mid \bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m})$ is also a fixed probability distribution. Then, by the pigeonhole principle, there exists at least one $\tilde{u}_t \in \mathcal{U}^{(t)}_{u_{t-1}}$ such that $\mathbf{B}^p_{\tilde{u}_t}(\delta_t)$ are scarce-sampled, i.e., have at most h_t sample points in $\mathbf{B}^p_{\tilde{u}_t}(\delta_t)$, with a constant probability. The next lemma formalizes and proves this idea.

Lemma 10 *Suppose there exist $u_{1:M}$ and v such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m}) \ge \frac{1}{4^{t-1}}$ for sufficiently large n (ST_{t-1}). Then, there* **exist** *re-chosen parameters $\tilde{u}_{t:M}$ and $\tilde{v}$ such that the following lower bound is satisfied:*

$$\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}\left(\Lambda^{(t)}_{\tilde{u}_t} \mid \bigcap_{m=1}^{t-1} \Lambda^{(m)}_{u_m}\right) \ge \frac{1}{2} \tag{29}$$

The proof of Lemma 10 is in Appendix D.7. By Lemma 10, there exists a function $f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}$ such that Eq (29) holds. Combining this with Eq (28), we finish the proof of ST_t.

Final step. The proof of the final step is similar to the inductive step. From ST_M, let $u_{1:M}$ and v be the parameter values satisfying $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}) \ge \frac{1}{4^M}$. We can state another lemma, which actually is our goal (17):

Lemma 11 *Suppose there exist $u_{1:M}$ and v such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}) \ge \frac{1}{4^M}$ for sufficiently large n. Then, for $\tilde{v} \ne v$, the following lower bound is satisfied for n large enough:*

$$\mathbb{P}^{\tilde{v}}_{u_{1:M}}\left(\bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}\right) \ge \frac{1}{2 \cdot 4^M}. \tag{30}$$

Also, the total variation distance between two conditional probability $\mathbb{P}^{-1}_{u_{1:M}}$ *and* $\mathbb{P}^{+1}_{u_{1:M}}$ *given the event* $\bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}$ *satisfies*

$$\left\| \mathbb{P}^{-1}_{u_{1:M}} \left(\cdot \mid \bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m} \right) - \mathbb{P}^{+1}_{u_{1:M}} \left(\cdot \mid \bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m} \right) \right\|_{\mathrm{TV}} \le \frac{1}{2}, \tag{31}$$

for sufficiently large n.

The proof is very similar to those of Lemma 9 and 10, and is also deferred to Appendix D.8.

B.4. Finishing the proof

Now, we are left with the final step of our proof. Recall from Eq (16), Lemmas 4 and 11, Condition 5.7, and substituting $G \leftarrow \bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}$ that

$$\begin{aligned}
\mathfrak{M}_M(\mathcal{F}, \mathcal{O}) &\ge \frac{1}{2 \cdot 4^M} \max_{v \in \mathcal{V}} \mathbb{E}^v_{u_{1:M}} [f^v_{u_{1:M}}(\widehat{X}) - (f^v_{u_{1:M}})^\star \mid \Lambda_{1:M}] \\
&\ge \frac{\mathsf{d}_{\mathrm{opt}}(f^{-1}_{u_{1:M}}, f^{+1}_{u_{1:M}})}{2 \cdot 4^{M+1}} \left(1 - \left\| \mathbb{P}^{-1}_{u_{1:M}}(\cdot \mid \bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}) - \mathbb{P}^{+1}_{u_{1:M}} + (\cdot \mid \bigcap_{m=1}^{M} \Lambda^{(m)}_{u_m}) \right\|_{\mathrm{TV}} \right) \\
&\ge \frac{C \alpha^M \eta^\kappa D_M^\kappa}{2 \cdot 4^{M+1}} n^{-\kappa \gamma_M} \exp\left(-2\sqrt{2} \kappa \gamma_{M-1} \sqrt{\log n} \right) \log^{-\nu} n,
\end{aligned} \tag{32}$$

where the last inequality used

$$\begin{aligned}
&\mathsf{d}_{\mathrm{opt}}(f^{-1}_{u_{1:M}}, f^{+1}_{u_{1:M}}) = 2C\alpha^M \eta^\kappa \delta_M^\kappa, \text{ and} \\
&\delta_M := D_M n^{-\gamma_M} \exp\left(-2\sqrt{2} \gamma_{M-1} \sqrt{\log n} \right) \log^{-\nu/\kappa} n.
\end{aligned}$$

The sub-polynomial factors such as $\exp\left(-2\sqrt{2}\kappa\gamma_{M-1}\sqrt{\log n}\right)$ and $\log^{-\nu} n$ unfortunately appear during the course of analysis (while proving lemmas), and whether they are artifacts of analysis or inevitable factors is not clear at this moment. To simplify the expressions a bit, note from Eq (21) that

$$\gamma_{M-1} = \frac{1}{2(\kappa - \zeta)} \left(1 - \left(\frac{d}{d + 2(\kappa - \zeta)} \right)^{M-1} \right),$$

and that the term

$$\exp\left(\frac{\sqrt{2}\kappa}{\kappa - \zeta} \left(\frac{d}{d + 2(\kappa - \zeta)} \right)^{M-1} \sqrt{\log n} \right) \log^{-\nu} n$$

in Eq (32) is an increasing but sub-polynomial factor in n. Thus, in presenting the lower bound we can safely discard those factors, resulting in a bound

$$\mathfrak{M}_M(\mathcal{F}, \mathcal{O}) \ge \frac{C \alpha^M \eta^\kappa D_M^\kappa}{2 \cdot 4^{M+1}} n^{-\kappa \gamma_M} \exp\left(-\frac{\sqrt{2}\kappa \sqrt{\log n}}{\kappa - \zeta} \right). \tag{33}$$

Bounding the leading constant. From the leading constant $\frac{C\alpha^M \eta^\kappa D_M^\kappa}{2 \cdot 4^{M+1}}$ that appeared in Eq (33) and the definition of D_M in Eq (24), we can check that the leading constant is dependent on M, d, σ, C, α, η, β, κ, and ζ, and has a lower bound of the form $c_1 \cdot r^M$, where $c_1 > 0$ and $0 < r < 1$ are independent of M. Recall that our theorem statements are for $M \le \log\log n / \log(1 + \frac{2(\kappa-\zeta)}{d})$. This implies that

$$\log(r^M) \ge \log\left(r^{\log\log n/\log(1+\frac{2(\kappa-\zeta)}{d})}\right) = \frac{\log r}{\log(1+\frac{2(\kappa-\zeta)}{d})} \log\log n = \log\left((\log n)^{-c_2}\right),$$

where $c_2 := -\log r / \log(1 + \frac{2(\kappa-\zeta)}{d}) > 0$. Thus, for small M, the leading constant can be bounded below by $c_1 \cdot \log^{-c_2} n$, where $c_1, c_2 > 0$ depend on d, σ, C, α, η, β, κ, and ζ, *not* M.

Now recall from Condition 5 and Lemmas 6 and 17 that C, α, η, β, κ are defined according to function classes, and ζ depends on oracle order. So, in Lipschitz convex/zeroth-order case ($\mathcal{F}_L$), the constants c_1 and c_2 depend only on d, σ, L. In smooth strongly convex case ($\mathcal{F}_{H,\lambda}$), the constants c_1, c_2 (for $\zeta = 0, 1$) depend only on d, σ, H, and λ.

Proving each cases. Given the general bound (33), and definition of γ_M (21), let us now substitute κ and ζ to finish the proof for each case presented in Theorem 2. Recall that $\kappa = 1$ for Lipschitz convex functions, and $\kappa = 2$ for smooth strongly convex functions. Also, $\zeta = 0$ for zeroth order oracle, and $\zeta = 1$ for first order oracle.

In the case of L-Lipschitz and zeroth-order $(\kappa, \zeta) = (1, 0)$, for $M \le \log\log n / \log(1 + \frac{2}{d})$,

$$\mathfrak{M}_M(\mathcal{F}_L, \mathcal{O}_0) \ge c_1 n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^M\right)} e^{-\sqrt{2\log n}} \log^{-c_2} n,$$

where $c_1, c_2 > 0$ depend only on d, σ, L. The cases of $(\kappa, \zeta) = (2, 0)$ and $(\kappa, \zeta) = (2, 1)$ can be treated in similar ways.

Appendix C. Function construction for smooth strongly convex functions

This section handles the construction of smooth strongly convex functions ($\mathcal{F}_{H,\lambda}$). The high-level idea is the same as the Lipschitz case, but taking max of two smooth functions f and g can break smoothness on the *intersection set* $\mathbf{Its}(f, g) := \{x \in \mathbb{R}^d \mid f(x) = g(x)\}$, so this case requires considerably more involved treatment. On the vicinity of the intersection set, we *interpolate* the gradients of two functions to smooth the boundary, using suprema of infinitely many hyperplanes. Section C.1 shows a simple 1-d example that illustrates the idea of "smoothing" maximum of two quadratic functions, and Section C.2 presents the algorithm SMAX that calculates smooth maximum of two multi-dimensional quadratic functions. Finally, Section C.3 describes recursive multi-stage construction of the function (with ℓ_2 balls at this time) by a similar way as in Section B.2.1. We also present that with suitable parameter choices, the constructed functions are in $\mathcal{F}_{H,\lambda}$ and satisfy Condition 5.

C.1. Smooth interpolation of two quadratic functions: a 1D example

Before we describe all the complicated details, let us start with an easy 1D example that illustrates our key approach. In Lipschitz convex case, the goal of getting functions that have the same values outside a certain set while having different values in that set was simply achieved by taking max

operations. We want to do the same for smooth strongly convex functions, but the difficulty is that the maximum of two functions f and g is not smooth on the intersection set $\mathbf{Its}(f, g)$. To remedy this problem, we can *interpolate* or *smooth* the functions near the intersection set using *suprema of infinitely many hyperplanes.*

To illustrate the idea, let us start with a simple 1D example; consider taking the maximum of two quadratic functions $f_1(x) = 2x^2$ and $f_2(x) = (x-1)^2 + 14$ defined on $[0, \infty)$. Their intersection set is $\mathbf{Its}(f_1, f_2) = \{3\}$, which contains the only non-smooth point of $\max\{f_1, f_2\}$. Note that $\max\{f_1, f_2\} = f_1$ for $x \geq 3$ and $\max\{f_1, f_2\} = f_2$ for $x \leq 3$. Smoothing is done by linearly interpolating the gradient in the vicinity of the non-smooth point, for example $x \in [2, 4]$, while leaving the function values outside $[2, 4]$ the same. We define constants $\dot{g}_-, \dot{g}_0, \dot{g}_+$:

$$\dot{g}_- := \dot{f}_2(2) = 2,\ \dot{g}_0 := \frac{\dot{f}_2(3) + \dot{f}_1(3)}{2} = \frac{4+12}{2} = 8,\ \dot{g}_+ := \dot{f}_1(4) = 16,$$

and then define the linearly interpolated gradients

$$\begin{aligned} \dot{h}_-(x) &:= \dot{g}_- + (\dot{g}_0 - \dot{g}_-)(x-2) = 6x - 10 && \text{for } x \in [2, 3], \\ \dot{h}_+(x) &:= \dot{g}_+ - (\dot{g}_0 - \dot{g}_+)(x-4) = 8x - 16 && \text{for } x \in [3, 4]. \end{aligned}$$

Since the gradient in $[2, 4]$ changed, also calculate interpolated function value accordingly:

$$\begin{aligned} h_-(x) &:= f_2(2) + \int_2^x \dot{h}_-(t)dt = 3x^2 - 10x + 23 && \text{for } x \in [2, 3], \\ h_+(x) &:= f_1(4) - \int_x^4 \dot{h}_+(t)dt = 4x^2 - 16x + 32 && \text{for } x \in [3, 4]. \end{aligned}$$

Note that $f_2(2) = h_-(2) = 15$, $h_-(3) = h_+(3) = 20$, and $h_+(4) = f_1(4) = 32$, so the functions f_2, h_-, h_+, f_1 can be "connected" to make a continuous function. Lastly define an infinite number of affine functions using previously calculated $\dot{h}_-$, $\dot{h}_+$, h_- and h_+:

$$\begin{aligned} f_-^\rho(x) &:= \dot{h}_-(\rho)(x - \rho) + h_-(\rho) \text{ for } \rho \in [2, 3], \\ f_+^\rho(x) &:= \dot{h}_+(\rho)(x - \rho) + h_+(\rho) \text{ for } \rho \in [3, 4]. \end{aligned}$$

Here, we are defining one affine function $f_-^\rho(x)$ for *each* value of $\rho \in [2, 3]$. Same applies to f_+^ρ for $\rho \in [3, 4]$.

Now, we can define the interpolated function, which is the supremum of the original functions f_1 and f_2, and affine functions f_-^ρ and f_+^ρ:

$$f(x) := \max\left\{ f_1(x), f_2(x), \sup_{\rho \in [2,3]} f_-^\rho(x), \sup_{\rho \in [3,4]} f_+^\rho(x) \right\}. \tag{34}$$

We can prove that this $f(x)$, defined as the maximum of many functions, is actually a smooth interpolation of maximum of f_1 and f_2. We state this in the following lemma.

Lemma 12 *The function $f(x)$ defined in Eq (34) satisfies the following:*

$$f(x) = \begin{cases} f_2(x) & \text{if } 0 \leq x \leq 2 \\ h_-(x) & \text{if } 2 \leq x \leq 3 \\ h_+(x) & \text{if } 3 \leq x \leq 4 \\ f_1(x) & \text{if } x \geq 4, \end{cases} \text{ and } \dot{f}(x) = \begin{cases} \dot{f}_2(x) & \text{if } 0 \leq x \leq 2 \\ \dot{h}_-(x) & \text{if } 2 \leq x \leq 3 \\ \dot{h}_+(x) & \text{if } 3 \leq x \leq 4 \\ \dot{f}_1(x) & \text{if } x \geq 4. \end{cases}$$

Also, $f(x)$ is a 8-smooth and 2-strongly convex function.

The proof of Lemma 12 is a special case of Lemmas 14 and 15, which we will omit.

By Lemma 12, we saw that the maximum of two quadratic functions can be smoothly interpolated by using infinite number of hyperplanes. Also note that this function is 8-smooth and 2-strongly convex, whereas f_1 and f_2 are in the class of 4-smooth and 2-strongly convex functions; the interpolation cause the "increase" of the smoothness constant, which we will observe in the multi-dimensional example as well.

C.2. Smooth interpolation of two quadratic functions: multi-dimension

Extending to multi-dimension. Now, we extend the domain to d-dimension, and consider taking the maximum of two quadratic functions $f_1(x)$ and $f_2(x)$, each minimized at x_1 and x_2, respectively, where $x_2 \in \mathbf{B}_{x_1}^2(\eta\delta)$:

$$f_1(x) := \|x - x_1\|_2^2 \text{ and } f_2(x) := \alpha \|x - x_2\|_2^2 + \beta\delta^2,$$

where the parameters satisfy $0 < \eta < 1$, $\delta > 0$, $0 < \alpha < 1$, and $0 < \beta < 1$.

By solving $f_1(x) = f_2(x)$, we can see that their intersection set $\mathbf{Its}(f_1, f_2) := \{x \in \mathbb{R}^d \mid f_1(x) = f_2(x)\}$ is a sphere:

$$\mathbf{Its}(f_1, f_2) = \{x \mid \|x - c\|_2^2 = r^2\},$$

where

$$c = \frac{1}{1-\alpha}x_1 - \frac{\alpha}{1-\alpha}x_2 = x_1 - \frac{\alpha}{1-\alpha}(x_2 - x_1), \tag{35}$$

$$r = \sqrt{\frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{\beta\delta^2}{1-\alpha}}. \tag{36}$$

Since $\frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{\beta\delta^2}{1-\alpha} > 0$ by assumptions on parameters, this sphere exists.

As seen in the 1D example, we need some "margin" for interpolation near non-smooth points. In the 1D example the "margin" or what we call "interpolation set" was the interval $[2, 4]$, on which we alter the function values to do smooth interpolation. So, after taking the maximum between f_1 and f_2, we will smooth the non-smooth points in the intersection set $\mathbf{Its}(f_1, f_2)$ by linearly interpolating the gradients on a set $\mathbf{Itp}$ called "interpolation set":

$$\mathbf{Itp} := \{x \mid (1-\theta)r \le \|x - c\|_2 \le (1+\theta)r\},$$

where $0 < \theta < 1$ will be chosen shortly.

Choosing the right parameters. We now state a lemma that chooses the parameters in the "right" way that makes our construction of smooth maximum easier.

Lemma 13 *Recall the conditions $0 < \eta < 1$, $\delta > 0$, $0 < \alpha < 1$, $0 < \beta < 1$, and $0 < \theta < 1$ on parameters of f_1, f_2, and $\mathbf{Itp}$. Choose parameters that satisfy*

$$\eta + \alpha + \alpha\eta < 1, \tag{37}$$

Algorithm 2 Algorithm $\textsc{sMax}(f_1, f_2, \alpha, \eta, \delta)$.

Assume $0 < \alpha < 1, 0 < \eta < 1, \eta + \alpha + \alpha\eta < 1, \delta > 0$. Let $\beta := \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}$.

Assume $f_1(x) = s\left\|x - x_1\right\|_2^2 + t$, $f_2(x) = s\alpha\left\|x - x_2\right\|_2^2 + s\beta\delta^2 + t$, $s > 0$, $x_2 \in \mathbf{B}_{x_1}^2(\eta\delta)$

1: $\theta := \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta}$, $c := x_1 - \frac{\alpha}{1-\alpha}(x_2 - x_1)$, $r := \sqrt{\frac{\alpha}{(1-\alpha)^2}\left\|x_1 - x_2\right\|_2^2 + \frac{\beta\delta^2}{1-\alpha}}$.

2: For all $\rho \in [(1-\theta)r, r]$ and all unit vectors $\left\|w\right\|_2 = 1$,

$$\begin{aligned}
\dot{h}_-(\rho, w) :=& \frac{2\alpha}{1-\alpha}(x_1 - x_2) - \frac{(1-\alpha)(1-\theta)r}{\theta}w + \left(\frac{1-\alpha}{\theta} + 2\alpha\right)\rho w, \\
h_-(\rho, w) :=& \frac{\alpha}{(1-\alpha)^2}\left\|x_1 - x_2\right\|_2^2 + \beta\delta^2 + \frac{(1-\alpha)(1-\theta)^2 r^2}{2\theta} \\
&+ \left(\frac{2\alpha}{1-\alpha}\langle x_1 - x_2, w\rangle - \frac{(1-\alpha)(1-\theta)r}{\theta}\right)\rho + \left(\frac{1-\alpha}{2\theta} + \alpha\right)\rho^2, \\
f_-^{\rho,w}(x) :=& s\langle \dot{h}_-(\rho, w), x - (c + \rho w)\rangle + s h_-(\rho, w) + t.
\end{aligned}$$

3: For all $\rho \in [r, (1+\theta)r]$ and all unit vectors $\left\|w\right\|_2 = 1$,

$$\begin{aligned}
\dot{h}_+(\rho, w) :=& \frac{2\alpha}{1-\alpha}(x_1 - x_2) - \frac{(1-\alpha)(1+\theta)r}{\theta}w + \left(\frac{1-\alpha}{\theta} + 2\right)\rho w, \\
h_+(\rho, w) :=& \frac{\alpha^2}{(1-\alpha)^2}\left\|x_1 - x_2\right\|_2^2 + \frac{(1-\alpha)(1+\theta)^2 r^2}{2\theta} \\
&+ \left(\frac{2\alpha}{1-\alpha}\langle x_1 - x_2, w\rangle - \frac{(1-\alpha)(1+\theta)r}{\theta}\right)\rho + \left(\frac{1-\alpha}{2\theta} + 1\right)\rho^2, \\
f_+^{\rho,w}(x) :=& s\langle \dot{h}_+(\rho, w), x - (c + \rho w)\rangle + s h_+(\rho, w) + t.
\end{aligned}$$

4: Return $f(x) := \max\left\{f_1(x), f_2(x), \sup_{\rho\in[(1-t)r,r],\left\|w\right\|_2=1} f_-^{\rho,w}(x), \sup_{\rho\in[r,(1+t)r],\left\|w\right\|_2=1} f_+^{\rho,w}(x)\right\}$.

$$\beta = \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}, \tag{38}$$

$$\theta = \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta}. \tag{39}$$

Then, the following statements hold:

$$\mathbf{Itp} \subset \mathrm{cl}(\mathbf{B}_{x_1}^2(\eta\delta)^c \cap \mathbf{B}_{x_1}^2(\delta)) \text{ for any } x_2 \in \mathbf{B}_{x_1}^2(\eta\delta), \tag{40}$$

The proof of Lemma 13 is provided in Appendix E.1. With the parameters satisfying Eqs (37)–(39), we can ensure that the interpolation set is a subset of $\mathrm{cl}(\mathbf{B}_{x_1}^2(\eta\delta)^c \cap \mathbf{B}_{x_1}^2(\delta))$, so any point $x \notin \mathrm{cl}(\mathbf{B}_{x_1}^2(\eta\delta)^c \cap \mathbf{B}_{x_1}^2(\delta))$ will not be affected by the interpolation.

Smoothing the maximum of two quadratic functions. We now describe how the smooth interpolation of $\max\{f_1, f_2\}$ is done in $\mathbf{Itp}$. Notice that $\mathbf{Itp}$ can be expressed in a "polar" form:

$$\mathbf{Itp} := \{x \mid (1-\theta)r \le \left\|x - c\right\|_2 \le (1+\theta)r\} = \{c + \rho w \mid (1-\theta)r \le \rho \le (1+\theta)r, \left\|w\right\|_2 = 1\},$$

and we will specify the new interpolated gradient and function values for all $(1-\theta)r \le \rho \le (1+\theta)r$ and $\|w\|_2 = 1$. For each fixed direction w, interpolated gradients $\dot{h}_-(\rho, w)$ and $\dot{h}_+(\rho, w)$ are obtained by linearly interpolating the gradients along w. After that, we obtain interpolated function values $h_-(\rho, w)$ and $h_+(\rho, w)$ by integrating directional derivatives along w, starting from $f_2(c + (1-\theta)r)$ and $f_1(c + (1+\theta)r)$, respectively.

For each fixed w, we define

$$\begin{aligned}\dot{g}_-(w) &:= \nabla f_2(c + (1-\theta)rw),\\ \dot{g}_0(w) &:= \frac{\nabla f_2(c + rw) + \nabla f_1(c + rw)}{2},\\ \dot{g}_+(w) &:= \nabla f_1(c + (1+\theta)rw).\end{aligned}$$

and then linearly interpolate the gradients along each direction w:

$$\begin{aligned}\dot{h}_-(\rho, w) &:= \dot{g}_-(w) + \frac{\dot{g}_0(w) - \dot{g}_-(w)}{\theta r}(\rho - (1-\theta)r) && \text{for } \rho \in [(1-\theta)r, r],\\ \dot{h}_+(\rho, w) &:= \dot{g}_+(w) - \frac{\dot{g}_0(w) - \dot{g}_+(w)}{\theta r}(\rho - (1+\theta)r) && \text{for } \rho \in [r, (1+\theta)r].\end{aligned}$$

Function values are obtained by integrating the directional derivatives along the direction w: The function values after interpolation is calculated by integrating the directional derivatives, i.e.,

$$\begin{aligned}h_-(\rho, w) &:= f_2(c + (1-\theta)rw) + \int_{(1-\theta)r}^{\rho} \langle \dot{h}_-(t, w), w\rangle dt && \text{for } \rho \in [(1-\theta)r, r],\\ h_+(\rho, w) &:= f_1(c + (1+\theta)rw) - \int_{\rho}^{(1+\theta)r} \langle \dot{h}_+(t, w), w\rangle dt && \text{for } \rho \in [r, (1+\theta)r].\end{aligned}$$

Using $\dot{h}_-$, $\dot{h}_+$, h_-, and h_+ defined as above, we can define infinite number of hyperplanes corresponding to each point $c + \rho w$ in $\mathbf{Itp}$,

$$\begin{aligned}f_-^{\rho,w}(x) &:= \langle \dot{h}_-(\rho, w), x - (c + \rho w)\rangle + h_-(\rho, w) && \text{for } \rho \in [(1-\theta)r, r], \|w\|_2 = 1,\\ f_+^{\rho,w}(x) &:= \langle \dot{h}_+(\rho, w), x - (c + \rho w)\rangle + h_+(\rho, w) && \text{for } \rho \in [r, (1+\theta)r], \|w\|_2 = 1.\end{aligned}$$

Finally, we define the smoothed function

$$f(x) := \max\left\{ f_1(x), f_2(x), \sup_{\rho \in [(1-t)r, r], \|w\|_2 = 1} f_-^{\rho,w}(x), \sup_{\rho \in [r, (1+t)r], \|w\|_2 = 1} f_+^{\rho,w}(x) \right\}.$$

For more details of the calculation, please refer to Appendix E.2.

We summarize the construction in Algorithm 2. The equations written in Algorithm 2 are just explicit calculation of $\dot{h}_-$, $\dot{h}_+$, h_-, and h_+, using the parameters as defined in Lemma 13. Algorithm 2 presents the process of getting the "smooth maximum" of two quadratic functions, for a slightly more general case where $f_1(x)$ and $f_2(x)$ are defined in the form

$$f_1(x) := s\|x - x_1\|_2^2 + t \;\text{ and }\; f_2(x) := s\alpha\|x - x_2\|_2^2 + s\beta\delta^2 + t,$$

where $s > 0$ and $t \in \mathbb{R}$. That is, s and t are scale and translation in the range space.

Correctness of smooth maximum. We now prove that the output of Algorithm 2 is indeed the smooth interpolation of $\max\{f_1, f_2\}$, and the interpolated function values attain the maximum/suprimum as originally intended. We prove this by the following lemma, whose technical proof is deferred to Appendix E.3.

Lemma 14 *Let*

$$f_1(x) := \|x - x_1\|_2^2 \;\; \textit{and} \;\; f_2(x) := \alpha \|x - x_2\|_2^2 + \beta\delta^2,$$

where parameters satisfy $0 < \alpha < 1$, $0 < \eta < 1$, $\eta + \alpha + \alpha\eta < 1$, $\delta > 0$, $\beta = \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}$, *and* $x_2 \in \mathbf{B}_{x_1}^2(\eta\delta)$. *Then, the output* $f(x)$ *of* $\textsc{sMax}(f_1, f_2, \alpha, \eta, \delta)$ *satisfies, for all* $\rho \geq 0$ *and* $\|w\|_2 = 1$,

$$f(c+\rho w) = \begin{cases} f_2(c+\rho w) & \textit{if } \rho \in [0, (1-\theta)r] \\ h_-(\rho, w) & \textit{if } \rho \in [(1-\theta)r, r] \\ h_+(\rho, w) & \textit{if } \rho \in [r, (1+\theta)r] \\ f_1(c+\rho w) & \textit{if } \rho \in [(1+\theta)r, \infty) \end{cases}$$

$$\nabla f(c+\rho w) = \begin{cases} \nabla f_2(c+\rho w) & \textit{if } \rho \in [0, (1-\theta)r] \\ \dot{h}_-(\rho, w) & \textit{if } \rho \in [(1-\theta)r, r] \\ \dot{h}_+(\rho, w) & \textit{if } \rho \in [r, (1+\theta)r] \\ \nabla f_1(c+\rho w) & \textit{if } \rho \in [(1+\theta)r, \infty). \end{cases}$$

Given Lemma 14, we showed that the interpolated function $f(x)$ has function values and gradients as specified in the lemma, which agrees with our intended construction in Appendix E.2. Note that positive scaling and translation does not hurt the correctness of interpolation.

Now, we prove the smoothness and strong convexity constants of $f(x)$.

Lemma 15 *Under the same setting as Lemma 14, the output* $f(x)$ *of* $\textsc{sMax}(f_1, f_2, \alpha, \eta, \delta)$ *is* $\left(2 + \frac{1-\alpha}{\theta}\right)$*-smooth and* 2α*-strongly convex.*

The proof is in Appendix E.4. Note that, as in the 1D case, the smoothness constant increased after interpolation while the strong convexity constant stayed the same.

We end this subsection with a lemma on the range of $f(x)$, which will prove useful for multi-stage construction as well as the reader's comprehension of the interpolation. The proof is deferred to Appendix E.5.

Lemma 16 *Let*

$$f_1(x) := s\|x - x_1\|_2^2 + t \;\; \textit{and} \;\; f_2(x) := s\alpha\|x - x_2\|_2^2 + s\beta\delta^2 + t,$$

where parameters satisfy $0 < \alpha < 1$, $0 < \eta < 1$, $\eta + \alpha + \alpha\eta < 1$, $\delta > 0$, $\beta = \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}$, $s > 0$ *and* $x_2 \in \mathbf{B}_{x_1}^2(\eta\delta)$. *Then, the output* $f(x)$ *of* $\textsc{sMax}(f_1, f_2, \alpha, \eta, \delta)$ *satisfies*

1. $f(x) = \begin{cases} f_1(x) & \forall x \in \mathrm{cl}(\mathbf{B}_{x_1}^2(\delta)^c), \\ f_2(x) & \forall x \in \mathbf{B}_{x_1}^2(\eta\delta), \end{cases}$

Algorithm 3 Construction of smooth strongly convex $f^v_{u_{1:M}}$.

Given parameters $u_1, u_2, \ldots, u_M, v$, size $\delta_1, \ldots, \delta_M$,
$\alpha, \eta \in (0,1)$ satisfying $\eta + \alpha + \alpha\eta < 1$, and $C > 0$,
1: Let $\beta := \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}$
2: Start with $f_{u_1}(x) = h_{u_1}(x) := C\|x - u_1\|_2^2$.
3: **for** $t = 2$ to M **do**
4: $\quad h_{u_{1:t}}(x) := C\alpha^{t-1}\|x - u_t\|_2^2 + C\beta\sum_{m=1}^{t-1}\alpha^{m-1}\delta_m^2$
5: $\quad g_{u_{1:t}}(x) := \textsc{sMax}(h_{u_{1:t-1}}, h_{u_{1:t}}, \alpha, \eta, \delta_{t-1})$.
6: $\quad f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), g_{u_{1:t}}(x)\}$.
7: **end for**
8: $h^v_{u_{1:M}}(x) := C\alpha^M\|x - u_M - v\eta\delta_M\mathbf{e}_1\|_2^2 + C\beta\sum_{m=1}^{M}\alpha^{m-1}\delta_m^2$.
9: $g^v_{u_{1:M}}(x) := \textsc{sMax}(h_{u_{1:M}}, h^v_{u_{1:M}}, \alpha, \eta, \delta_M)$.
10: Return $f^v_{u_{1:M}}(x) := \max\{f_{u_{1:M}}(x), g^v_{u_{1:M}}(x)\}$.

2. $\nabla f(x) = \begin{cases} \nabla f_1(x) & \forall x \in \mathrm{cl}(\mathbf{B}^2_{x_1}(\delta)^c), \\ \nabla f_2(x) & \forall x \in \mathbf{B}^2_{x_1}(\eta\delta), \end{cases}$

3. $s\beta\delta^2 + t \le f(x) \le s\delta^2 + t \quad \forall x \in \mathbf{B}^2_{x_1}(\delta)$.

4. $\|\nabla f(x)\|_2 \le 2s\delta \quad \forall x \in \mathbf{B}^2_{x_1}(\delta)$.

C.3. Multi-stage recursive construction

Now let us consider applying $\textsc{sMax}(\cdot)$ many times in a recursive way, as done in the Lipschitz convex case; we will iteratively apply $\textsc{sMax}$ while zooming into narrower regions of the domain. The outline of the construction is the same, except a bit of difference in details.

As seen in Section B.2, we construct nested maximal packings. For $\mathcal{F}_{H,\lambda}$ this is done with ℓ_2 norm. Given the maximal packings, we can recursively choose chain of parameters $u_1, \ldots, u_M$ and v. Algorithm 3 constructs a function $f^v_{u_{1:M}}(x)$ that corresponds to the specific choice of $u_1, \ldots, u_M$ and v. Algorithm 3 defines a series of quadratic functions $h_{u_{1:t}}(x)$ and repeatedly takes smooth maximum with previous ones to get the final function.

With Algorithm 3, we can show that the outputs of Algorithm 3 satisfy Condition 5, as desired.

Lemma 17 *The functions constructed by Algorithm 1 satisfy Condition 5 with*

$$(C, \alpha, \eta, \beta, \kappa, p) = \left(C, \alpha, \eta, \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}, 2, 2\right),$$

if $\alpha, \eta \in (0,1)$, $\eta + \alpha + \alpha\eta < 1$, *and* $C > 0$.

The proof of Lemma 17 is deferred to Appendix E.6.

We also want to check whether the functions $f^v_{u_{1:M}}(x)$ we are constructing are indeed smooth and strongly convex. Especially, one might wonder if the $\max$ operations at Lines 6 and 10 in Algorithm 3 can hurt the smoothness. The following lemma addresses this points, whose proof is deferred to Appendix E.7.

Lemma 18 *The functions constructed with Algorithm 3 are* $\left(C\left(2+\frac{1-\alpha}{\theta}\right)\right)$*-smooth and* $2C\alpha^M$*-strongly convex, where* $\theta := \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta}$ *as in Eq (39). Moreover, when* $H/5 \geq \lambda$*, with the parameter choice*

$$\alpha = \left(\frac{1}{2}\right)^{\frac{1}{M}}, \ \eta = \frac{1-\alpha}{2}, \ C = \frac{H}{5}$$

the constructed functions are H*-smooth and* λ*-strongly convex. For* $H/5 < \lambda < H$*, there also exists choice of parameters that returns* H*-smooth and* λ*-strongly convex functions, although a bit more complicated.*

Appendix D. Technical Proofs for Section B

D.1. Proof of Lemma 4

Consider a hypothesis testing problem, where $v \in \{-,+\}$ is sampled uniformly at random by the nature and v is not known to us. We have to estimate the v using a hypothesis test Ψ. Observe that if $f_v(\widehat{X}) - f_v^\star \leq \mathsf{d}_{\mathrm{opt}}(f_-, f_+)/2$ for some $v \in \{-,+\}$, then for $v' \neq v$,

$$\begin{aligned} &\mathsf{d}_{\mathrm{opt}}(f_-, f_+) \leq f_v(\widehat{X}) + f_{v'}(\widehat{X}) - f_v^\star - f_{v'}^\star \leq \tfrac{\mathsf{d}_{\mathrm{opt}}(f_-,f_+)}{2} + f_{v'}(\widehat{X}) - f_{v'}^\star \\ \implies &\tfrac{\mathsf{d}_{\mathrm{opt}}(f_-,f_+)}{2} \leq f_{v'}(\widehat{X}) - f_{v'}^\star. \end{aligned}$$

so only a single $v \in \{-,+\}$ may satisfy $f_v(\widehat{X}) - f_v^\star \leq \mathsf{d}_{\mathrm{opt}}(f_-, f_+)/2$. From this observation, we can define our test $\widehat{\Psi}$ using our optimization estimate $\widehat{X}$, $\widehat{\Psi} = \operatorname{argmin}_{v\in\{-,+\}} f_v(\widehat{X}) - f_v^\star$, where ties are broken arbitrarily. Notice that for any v, $\widehat{\Psi} \neq v$ implies $f_v(\widehat{X}) - f_v^\star \geq \frac{\mathsf{d}_{\mathrm{opt}}(f_-,f_+)}{2}$. Then, using Markov's inequality,

$$\begin{aligned} &\max_{v\in\{-,+\}} \mathbb{E}_v\left[f_v(\widehat{X}) - f_v^\star \mid G\right] \geq \frac{1}{2}\sum_{v\in\{-,+\}} \mathbb{E}_v\left[f_v(\widehat{X}) - f_v^\star \mid G\right] \\ \geq &\frac{\mathsf{d}_{\mathrm{opt}}(f_-, f_+)}{4} \sum_{v\in\{-,+\}} \mathbb{P}_v\left(f_{u_{1:M}}^v(\widehat{X}) - (f_{u_{1:M}}^v)^\star \geq \tfrac{\mathsf{d}_{\mathrm{opt}}(f_-,f_+)}{2} \mid G\right) \\ \geq &\frac{\mathsf{d}_{\mathrm{opt}}(f_-, f_+)}{4} \sum_{v\in\{-,+\}} \mathbb{P}_v(\widehat{\Psi} \neq v \mid G) \geq \frac{\mathsf{d}_{\mathrm{opt}}(f_-, f_+)}{4} \inf_{\Psi} \sum_{v\in\{-,+\}} \mathbb{P}_v(\Psi \neq v \mid G). \end{aligned}$$

where last the infimum is taken over all possible tests. By a classical inequality on hypothesis testing and total variation distance,

$$\inf_{\Psi} \sum_{v\in\{-,+\}} \mathbb{P}_v(\Psi \neq v \mid G) \geq 1 - \left\|\mathbb{P}_-(\cdot \mid G) - \mathbb{P}_+(\cdot \mid G)\right\|_{\mathrm{TV}}.$$

D.2. Proof of Lemma 6

We start by showing the following technical lemma, which illustrates how the functions in the max operation are placed above or below one another. Its proof is deferred to Appendix D.4.

Lemma 19 *For any set of parameters* $u_1, u_2, \ldots, u_M, v$ *chosen by* $u_1 \in \mathcal{U}^{(1)}$, $u_t \in \mathcal{U}_{u_{t-1}}^{(t)}$ *for* $t \in 2:M$*, and* $v \in \mathcal{V}$*, run Algorithm 1 and get* $f_{u_{1:M}}^v(x)$*. Then, for any* $t \in 2:M$*, we have:*

1. $f_{u_{1:t}}(x) = \begin{cases} f_{u_{1:t-1}}(x) & \forall x \notin \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}), \\ h_{u_{1:t}}(x) & \forall x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}/2), \end{cases}$

2. $\sum_{m=1}^{t-1} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f_{u_{1:t}}(x) \le \sum_{m=1}^{t-2} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{t-1}}{3^{t-2}}$ *for all* $x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$.

Also, at the final step,

3. $f^v_{u_{1:M}}(x) = \begin{cases} f_{u_{1:M}}(x) & \forall x \notin \mathbf{B}^{\infty}_{u_M}(\delta_M), \\ h^v_{u_{1:M}}(x) & \forall x \in \mathbf{B}^{\infty}_{u_M}(\delta_M/2), \end{cases}$

4. $\sum_{m=1}^{M} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f^v_{u_{1:M}}(x) \le \sum_{m=1}^{M-1} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_M}{3^{M-1}}$ *for all* $x \in \mathbf{B}^{\infty}_{u_M}(\delta_M)$.

We prove Lemma 6.1 and 6.4 using simple argument that max operations done in Algorithm 1 only changes limited parts of the domain. Recall the definition $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\}$. From Lemma 19.1, note that whenever we have $f_{u_{1:t-1}}(x)$ and take max operation with $h_{u_{1:t}}(x)$ to construct $f_{u_{1:t}}(x)$, any point $\forall x \notin \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$ does not change its value. This means that the max operation can only change function values in $\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$. Also, later iterations of the algorithm do not change that the function values at $x \notin \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$, because $\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}) \supset \mathbf{B}^{\infty}_{u_t}(\delta_t) \supset \cdots \supset \mathbf{B}^{\infty}_{u_M}(\delta_M)$. From this argument, we can see that $f^v_{u_{1:M}}(x) = f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x) = f_{u_{1:t-1}}(x)$ for all $x \notin \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$, therefore proving Lemma 6.1. Similarly, from Lemma 19.3, the final line $f^v_{u_{1:M}}(x) := \max\{f_{u_{1:M}}(x), h^v_{u_{1:M}}(x)\}$ in Algorithm 1 can only change function values in $\mathbf{B}^{\infty}_{u_M}(\delta_M)$, so $f^{-1}_{u_{1:M}}(x) = f^{+1}_{u_{1:M}}(x) = f_{u_{1:M}}(x)$ for all $x \notin \mathbf{B}^{\infty}_{u_M}(\delta_M)$, proving Lemma 6.4.

Lemma 6.5 can be implied directly by Lemma 19.4. In order to prove Lemma 6.2, note the following facts from Lemma 19.4 and 19.2:

$$\sum_{m=1}^{M} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f^v_{u_{1:M}}(x) \le \sum_{m=1}^{M-1} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_M}{3^{M-1}} \quad \text{for all } x \in \mathbf{B}^{\infty}_{u_M}(\delta_M),$$
$$\sum_{m=1}^{M-1} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f_{u_{1:M}}(x) \le \sum_{m=1}^{M-2} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{M-1}}{3^{M-2}} \quad \text{for all } x \in \mathbf{B}^{\infty}_{u_{M-1}}(\delta_{M-1}).$$

Note from Lemma 19.3 that $f^v_{u_{1:M}}(x) = f_{u_{1:M}}(x)$ for all $x \notin \mathbf{B}^{\infty}_{u_M}(\delta_M)$, and that, for all $x \in \mathbf{B}^{\infty}_{u_M}(\delta_M)$,

$$f^v_{u_{1:M}}(x) \le \sum_{m=1}^{M-1} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_M}{3^{M-1}} \le \sum_{m=1}^{M-2} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{M-1}}{3^{M-2}}.$$

The last inequality is because $\frac{3\delta_{M-1}}{2} \ge \delta_M$ holds for large enough n by assumption that $\delta_M = o(\delta_{M-1})$. From these observations, we have

$$\sum_{m=1}^{M-1} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f^v_{u_{1:M}}(x) \le \sum_{m=1}^{M-2} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{M-1}}{3^{M-2}} \quad \text{for all } x \in \mathbf{B}^{\infty}_{u_{M-1}}(\delta_{M-1}).$$

Again note that, for any $x \notin \mathbf{B}^{\infty}_{u_{M-1}}(\delta_{M-1})$ we also have $x \notin \mathbf{B}^{\infty}_{u_M}(\delta_M)$, so $f^v_{u_{1:M}}(x) = f_{u_{1:M}}(x) = f_{u_{1:M-1}}(x)$. We can repeat a similar argument and obtain

$$\sum_{m=1}^{M-2} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f^v_{u_{1:M}}(x) \le \sum_{m=1}^{M-3} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{M-2}}{3^{M-3}} \quad \text{for all } x \in \mathbf{B}^{\infty}_{u_{M-2}}(\delta_{M-2}).$$

For any $t \in 2:M$, we can repeat this argument until $\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$, so that we get

$$\sum_{m=1}^{t-1} \frac{L\delta_m}{2\cdot 3^{m-1}} \le f^v_{u_{1:M}}(x) \le \sum_{m=1}^{t-2} \frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{t-1}}{3^{t-2}} \quad \text{for all } x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}),$$

which directly implies Lemma 6.2 that we are after.

In order to prove Lemma 6.3 and 6.6, we first need to show that the function value $f^v_{u_{1:M}}(x)$ in $\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$ can be expressed as

$$f^v_{u_{1:M}}(x) = \max\left\{ \max_{k\in t-1:M} \left\{h_{u_{1:k}}(x)\right\}, h^v_{u_{1:M}}(x)\right\}, \text{ for all } x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}). \tag{41}$$

Notice from Lemma 19.1 that $f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ for all $x \in \mathbf{B}^{\infty}_{u_{t-2}}(\delta_{t-2}/2)$. Recall that $\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}) \subset \mathbf{B}^{\infty}_{u_{t-2}}(\delta_{t-2}/2)$, so $f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ in $\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$. After this point, $f^v_{u_{1:M}}(x)$ is obtained from max operations with $h_{u_{1:t}}, \ldots, h_{u_{1:M}}, h^v_{u_{1:M}}$. This proves Eq (41). Now notice that the subgradient of

$$h_{u_{1:t}}(x) := \frac{L}{3^{t-1}}\left\|x - u_t\right\|_\infty + \sum_{m=1}^{t-1} \frac{L\delta_m}{2\cdot 3^{m-1}}$$

always has ℓ_1 norm exactly $\frac{L}{3^{t-1}}$. From Eq (41), we can observe that for any point $x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$, any subgradient $g^v_{u_{1:M}} \in \partial f^v_{u_{1:M}}(x)$ has $\left\|g^v_{u_{1:M}}\right\|_1 \le \frac{L}{3^{t-2}}$. Since this holds for any set of parameters $u_{t:M}$, v, $\tilde{u}_{t:M}$, and $\tilde{v}$, we get Lemma 6.3. From a similar argument as Eq (41), we have

$$f^v_{u_{1:M}}(x) = \max\left\{h_{u_{1:M}}(x), h^v_{u_{1:M}}(x)\right\}, \text{ for all } x \in \mathbf{B}^{\infty}_{u_M}(\delta_M),$$

whereby we can prove Lemma 6.6.

Finally, we have to show Lemma 6.7. To do so, we first show that, for any choice of $u_1, u_2, \ldots, u_M$ and v,

$$\inf_x f^v_{u_{1:M}}(x) = \sum_{m=1}^{M} \frac{L\delta_m}{2\cdot 3^{m-1}} \tag{42}$$

In fact, from Lemma 19.3, we have $f^v_{u_{1:M}}(x) = h^v_{u_{1:M}}(x)$ for all $x \in \mathbf{B}^{\infty}_{u_M}(\delta_M/2)$. Also, $h^v_{u_{1:M}}(x)$ is minimized at $u_M + \frac{v\delta_M}{2}\mathbf{1} \in \mathbf{B}^{\infty}_{u_M}(\delta_M/2)$, whose minimum value is the RHS of Eq (42). So, for any $x \in \mathbf{D}$,

$$f^v_{u_{1:M}}(x) \ge h^v_{u_{1:M}}(x) \ge h^v_{u_{1:M}}\left(u_M + \frac{v\delta_M}{2}\mathbf{1}\right) = \sum_{m=1}^{M} \frac{L\delta_m}{2\cdot 3^{m-1}},$$

proving Eq (42).

Next, we show that

$$\inf_x (f^{+1}_{u_{1:M}}(x) + f^{-1}_{u_{1:M}}(x)) = \sum_{m=1}^{M} \frac{L\delta_m}{3^{m-1}} + \frac{L\delta_M}{3^M}. \tag{43}$$

Again note that $f^v_{u_{1:M}}(x) = h^v_{u_{1:M}}(x)$ for all $x \in \mathbf{B}^{\infty}_{u_M}(\delta_M/2)$. That is, for $x \in \mathbf{B}^{\infty}_{u_M}(\delta_M/2)$, we have $f^{+1}_{u_{1:M}}(x) = h^{+1}_{u_{1:M}}(x)$ and $f^{-1}_{u_{1:M}}(x) = h^{-1}_{u_{1:M}}(x)$. Therefore, for any $x \in \mathbf{B}^{\infty}_{u_M}(\delta_M/2)$,

$$f^{+1}_{u_{1:M}}(x) + f^{-1}_{u_{1:M}}(x) = h^{+1}_{u_{1:M}}(x) + h^{-1}_{u_{1:M}}(x)$$

$$= \frac{L}{3^M}\left(\left\|x - u_M - \tfrac{\delta_M}{2}\mathbf{1}\right\|_\infty + \left\|x - u_M + \tfrac{\delta_M}{2}\mathbf{1}\right\|_\infty\right) + \sum_{m=1}^{M}\frac{L\delta_m}{3^{m-1}}.$$

By triangle inequality, we have

$$\left\|x - u_M - \tfrac{\delta_M}{2}\mathbf{1}\right\|_\infty + \left\|x - u_M + \tfrac{\delta_M}{2}\mathbf{1}\right\|_\infty \geq \left\|\left(u_M + \tfrac{\delta_M}{2}\mathbf{1}\right) - \left(u_M - \tfrac{\delta_M}{2}\mathbf{1}\right)\right\|_\infty = \delta_M$$

Note also that $x = u_M$ in fact attains this lower bound. So, for any $x \in \mathbf{D}$,

$$f^{+1}_{u_{1:M}}(x) + f^{-1}_{u_{1:M}}(x) \geq h^{+1}_{u_{1:M}}(x) + h^{-1}_{u_{1:M}}(x) \geq h^{+1}_{u_{1:M}}(u_M) + h^{-1}_{u_{1:M}}(u_M) = \sum_{m=1}^{M}\frac{L\delta_m}{3^{m-1}} + \frac{L\delta_M}{3^M},$$

thus proving Eq (43). Now, Lemma 6.7 follows from Eq (42) and Eq (43).

D.3. Proof of Lemma 7

Consider the functions constructed in Algorithm 1. For any $t \in 1 : M$, the function $h_{u_{1:t}}(x)$ is convex and L-Lipschitz with respect to ℓ_p norm for any $p \geq 1$. Hence, $f_{u_{1:M}}(x) := \max_{1\leq t\leq M}\{h_{u_{1:t}}(x)\}$ is also convex and L-Lipschitz. With the same reasoning, $h^v_{u_{1:M}}(x)$ is convex and L-Lipshcitz, and so $f^v_{u_{1:M}}(x) := \max\{f_{u_{1:M}}(x), h^v_{u_{1:M}}(x)\}$ is.

D.4. Proof of Lemma 19

We demonstrate in details the proof for Lemma 19 below, which is based on an induction argument.

Base case $t = 2$. In the base case, for any $u_1 \in \mathcal{U}^{(1)}$ and $u_2 \in \mathcal{U}^{(2)}_{u_1}$, we want to show that

$$f_{u_1}(x) \geq h_{u_{1:2}}(x) \text{ for all } x \notin \mathbf{B}^\infty_{u_1}(\delta_1) \text{ and } f_{u_1}(x) \leq h_{u_{1:2}}(x) \text{ for any } x \in \mathbf{B}^\infty_{u_1}(\delta_1/2),$$

which correspond to Lemma 19.1. Recall the definitions that

$$f_{u_1}(x) = h_{u_1}(x) := L\left\|x - u_1\right\|_\infty \text{ and } h_{u_{1:2}}(x) := \frac{L}{3}\left\|x - u_2\right\|_\infty + \frac{L\delta_1}{2}.$$

Indeed, note that, since by definition $u_2 \in \mathbf{B}^\infty_{u_1}(\frac{\delta_1}{2} - \delta_2)$, we have $\|u_2 - u_1\|_\infty \leq \delta_1/2$. Therefore, by triangle inequality, we have $\|x - u_2\|_\infty \leq \|x - u_1\|_\infty + \delta_1/2$. Thus, for any $x \notin \mathbf{B}^\infty_{u_1}(\delta_1)$, we have

$$\begin{aligned} f_{u_1}(x) - h_{u_{1:2}}(x) &= L\left(\|x - u_1\|_\infty - \frac{\|x - u_2\|_\infty}{3} - \frac{\delta_1}{2}\right) \\ &\geq L\left(\|x - u_1\|_\infty - \frac{\|x - u_1\|_\infty}{3} - \frac{2\delta_1}{3}\right) \geq 0. \end{aligned}$$

On the other hand, by triangle inequality $\|x - u_2\|_\infty + \delta_1/2 \geq \|x - u_1\|_\infty$. Thus, for $x \in \mathbf{B}^\infty_{u_1}(\delta_1/2)$,

$$f_{u_1}(x) - h_{u_{1:2}}(x) = L\left(\|x - u_1\|_\infty - \frac{\|x - u_2\|_\infty}{3} - \frac{\delta_1}{2}\right)$$

$$\leq L\left(\|x-u_1\|_\infty - \frac{\|x-u_1\|_\infty}{3} - \frac{\delta_1}{3}\right) \leq 0,$$

which proves Lemma 19.1.

We are left with the inequalities below:

$$\frac{L\delta_1}{2} \leq f_{u_{1:2}}(x) \leq L\delta_1 \text{ for all } x \in \mathbf{B}^\infty_{u_1}(\delta_1). \tag{44}$$

Recall the definition $f_{u_{1:2}}(x) = \max\{f_{u_1}(x), h_{u_{1:2}}(x)\}$. Indeed, the LHS of Eq (44) follows from the fact that $L\delta_1/2 \leq h_{u_{1:2}}(x) \leq f_{u_{1:2}}(x)$. To show the RHS of inequality (44), we note first that for any $x \in \mathbf{B}^\infty_{u_1}(\delta_1)$, $f_{u_1}(x) = L\|x-u_1\|_\infty \leq L\delta_1$. In addition to that, since $\|u_1-u_2\|_\infty \leq \delta_1/2$, by triangle inequality, we also have that, for any $x \in \mathbf{B}^\infty_{u_1}(\delta_1)$,

$$h_{u_{1:2}}(x) := \frac{L}{3}\|x-u_2\|_\infty + \frac{L\delta_1}{2} \leq \frac{L}{3}\left(\|x-u_1\|_\infty + \frac{\delta_1}{2}\right) + \frac{L\delta_1}{2} \leq L\delta_1$$

Thus, we have, for any $x \in \mathbf{B}^\infty_{u_1}(\delta_1)$, $f_{u_{1:2}}(x) = \max\{f_{u_1}(x), h_{u_{1:2}}(x)\} \leq L\delta_1$. Together, we prove the desired inequality (44), which corresponds to Lemma 19.2 for $t=2$.

Inductive case $2 < t \leq M$**.** In the first step, we show our first claim in Lemma 19.1:

$f_{u_{1:t-1}}(x) \geq h_{u_{1:t}}(x)$ for any $x \notin \mathbf{B}^\infty_{u_{t-1}}(\delta_{t-1})$ and $f_{u_{1:t-1}}(x) \leq h_{u_{1:t}}(x)$ for any $x \in \mathbf{B}^\infty_{u_{t-1}}(\delta_{t-1}/2)$.

Recall the definitions that

$$\begin{aligned}
h_{u_{1:t-1}}(x) &:= \frac{L}{3^{t-2}}\|x-u_{t-1}\|_\infty + \sum_{m=1}^{t-2}\frac{L\delta_m}{2\cdot 3^{m-1}},\\
h_{u_{1:t}}(x) &:= \frac{L}{3^{t-1}}\|x-u_t\|_\infty + \sum_{m=1}^{t-1}\frac{L\delta_m}{2\cdot 3^{m-1}},\\
f_{u_{1:t-1}}(x) &:= \max\{f_{u_{1:t-2}}(x), h_{u_{1:t-1}}(x)\} \geq h_{u_{1:t-1}}(x),\\
f_{u_{1:t}}(x) &:= \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\} \geq h_{u_{1:t}}(x).
\end{aligned}$$

Indeed, note that, since by definition $u_t \in \mathbf{B}^\infty_{u_{t-1}}(\frac{\delta_{t-1}}{2} - \delta_t)$, we have $\|u_t - u_{t-1}\|_\infty \leq \delta_{t-1}/2$. Hence, by triangle inequality, we have $\|x-u_t\|_\infty \leq \|x-u_{t-1}\|_\infty + \delta_{t-1}/2$. Also, by definition $f_{u_{1:t-1}}(x) \geq h_{u_{1:t-1}}(x)$ for all x. Thus, for any $x \notin \mathbf{B}^\infty_{u_{t-1}}(\delta_{t-1})$, we have

$$\begin{aligned}
f_{u_{1:t-1}}(x) - h_{u_{1:t}}(x) \geq h_{u_{1:t-1}}(x) - h_{u_{1:t}}(x) &= \frac{L}{3^{t-2}}\left(\|x-u_{t-1}\|_\infty - \frac{\|x-u_t\|_\infty}{3} - \frac{\delta_{t-1}}{2}\right)\\
&\geq \frac{L}{3^{t-2}}\left(\|x-u_{t-1}\|_\infty - \frac{\|x-u_{t-1}\|_\infty}{3} - \frac{2\delta_{t-1}}{3}\right) \geq 0.
\end{aligned}$$

On the other hand, recall that $\mathbf{B}^\infty_{u_{t-1}}(\delta_{t-1}) \subset \mathbf{B}^\infty_{u_{t-2}}(\delta_{t-2}/2)$. When $x \in \mathbf{B}^\infty_{u_{t-1}}(\delta_{t-1}/2) \subset \mathbf{B}^\infty_{u_{t-2}}(\delta_{t-2}/2)$, Observe that by inductive hypothesis $f_{u_{1:t-2}}(x) \leq h_{u_{1:t-1}}(x)$. By definition of $f_{u_{1:t-1}}(x) := \max\{f_{u_{1:t-2}}(x), h_{u_{1:t-1}}(x)\}$, we can see that $f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ for $x \in$

$\mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}/2)$. By triangle inequality, $\|x-u_t\|_\infty + \delta_{t-1}/2 \geq \|x-u_{t-1}\|_\infty$. Thus, for $x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}/2)$,

$$f_{u_{1:t-1}}(x) - h_{u_{1:t}}(x) = h_{u_{1:t-1}}(x) - h_{u_{1:t}}(x) = \frac{L}{3^{t-2}}\left(\|x-u_{t-1}\|_\infty - \frac{\|x-u_t\|_\infty}{3} - \frac{\delta_{t-1}}{2}\right)$$
$$\leq \frac{L}{3^{t-2}}\left(\|x-u_{t-1}\|_\infty - \frac{\|x-u_{t-1}\|_\infty}{3} - \frac{\delta_{t-1}}{3}\right) \leq 0,$$

which proves Lemma 19.1.

We are left with Lemma 19.2, which we restate below:

$$\sum_{m=1}^{t-1}\frac{L\delta_m}{2\cdot 3^{m-1}} \leq f_{u_{1:t}}(x) \leq \sum_{m=1}^{t-2}\frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{t-1}}{3^{t-2}} \quad \text{for all } x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}).$$

Recall the definition $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), h_{u_{1:t}}(x)\}$. Indeed, the LHS of Lemma 19.2 follows from the fact that $\sum_{m=1}^{t-1}\frac{L\delta_m}{2\cdot 3^{m-1}} \leq h_{u_{1:t}}(x) \leq f_{u_{1:t}}(x)$. To show the RHS of Lemma 19.2, recall first that if $x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1}) \subset \mathbf{B}^{\infty}_{u_{t-2}}(\delta_{t-2}/2)$, $f_{u_{1:t-2}}(x) \leq h_{u_{1:t-1}}(x)$ by induction hypothesis, so $f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$. Thus, for all $x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$,

$$f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x) := \frac{L}{3^{t-2}}\|x-u_{t-1}\|_\infty + \sum_{m=1}^{t-2}\frac{L\delta_m}{2\cdot 3^{m-1}} \leq \sum_{m=1}^{t-2}\frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{t-1}}{3^{t-2}}.$$

Also, $\|u_{t-1}-u_t\|_\infty \leq \delta_{t-1}/2$, and by triangle inequality, $\|x-u_t\|_\infty \leq \|x-u_{t-1}\|_\infty + \delta_{t-1}/2$. Using this, for all $x \in \mathbf{B}^{\infty}_{u_{t-1}}(\delta_{t-1})$

$$h_{u_{1:t}}(x) := \frac{L}{3^{t-1}}\|x-u_t\|_\infty + \sum_{m=1}^{t-1}\frac{L\delta_m}{2\cdot 3^{m-1}} \leq \frac{L}{3^{t-1}}\|x-u_{t-1}\|_\infty + \frac{L\delta_{t-1}}{2\cdot 3^{t-1}} + \sum_{m=1}^{t-1}\frac{L\delta_m}{2\cdot 3^{m-1}}$$
$$\leq \frac{L\delta_{t-1}}{3^{t-1}} + \frac{L\delta_{t-1}}{2\cdot 3^{t-1}} + \sum_{m=1}^{t-1}\frac{L\delta_m}{2\cdot 3^{m-1}} = \sum_{m=1}^{t-2}\frac{L\delta_m}{2\cdot 3^{m-1}} + \frac{L\delta_{t-1}}{3^{t-2}}.$$

This finishes showing the RHS of Lemma 19.2.

Final Case. It is left to prove Lemma 19.3–4. Their proof can be done in a similar way as Lemma 19.1–2 for the inductive cases, hence omitted.

D.5. Proof of Lemma 8

Since $\delta_1 = o(1)$, for large enough n we have $\delta_1 \leq 1/8$, so

$$|\mathcal{U}^{(1)}| = 2\delta_1\text{-packing number (w.r.t. } \ell_p) \text{ of } [\delta_1, 1-\delta_1]^d$$
$$\geq 2\delta_1\text{-packing number (w.r.t. } \ell_p) \text{ of } \left[\frac{1}{4}, \frac{3}{4}\right]^d$$
$$\geq 2\delta_1\text{-covering number (w.r.t. } \ell_p) \text{ of } \mathbf{B}^p_0(1/4)$$
$$\geq \frac{\mathrm{Vol}(\mathbf{B}^p_0(1/4))}{\mathrm{Vol}(\mathbf{B}^p_0(2\delta_1))} = \left(\frac{1}{8\delta_1}\right)^d.$$

Similarly, by $\delta_t = o(\delta_{t-1})$, for large enough n we have $\delta_t \le \eta\delta_{t-1}/4$.

$$\begin{aligned}|\mathcal{U}^{(t)}_{u_{t-1}}| &= 2\delta_t\text{-packing number (w.r.t. } \ell_p) \text{ of } \mathbf{B}^p_{u_{t-1}}(\eta\delta_{t-1} - \delta_t)\\ &\ge 2\delta_t\text{-packing number (w.r.t. } \ell_p) \text{ of } \mathbf{B}^p_{u_{t-1}}(\eta\delta_{t-1}/2)\\ &\ge \frac{\mathrm{Vol}(\mathbf{B}^p_{u_{t-1}}(\eta\delta_{t-1}/2))}{\mathrm{Vol}(\mathbf{B}^p_0(2\delta_t))} = \left(\frac{\eta\delta_{t-1}}{4\delta_t}\right)^d.\end{aligned}$$

D.6. Proof of Lemma 9

In this section and the following two, for simplicity in notation, let

$$\Lambda_{1:k} = \bigcap\nolimits_{m=1}^{k} \Lambda^{(m)}_{u_m}.$$

The proofs of Lemmas 9 and 10 rely heavily on the following lemma which uses likelihood ratio and concentration inequality arguments to prove bounds between different probability measures defined by "similar" functions.

Lemma 20 *For $t \in 2 : M$, let two sets of parameters share the same value $u_{1:t-1}$ and then they differ afterwards: $u_{t:M}$, v and $\tilde{u}_{t:M}$, $\tilde{v}$. Consider any probabilistic event G that is a function of the random variables*

$$\begin{aligned}&X^{(1:k)}_{1:n}, Y^{(1:k)}_{1:n}, X^{(k+1)}_{1:n} && \text{if oracle is zeroth-order, or}\\ &X^{(1:k)}_{1:n}, Y^{(1:k)}_{1:n}, Z^{(1:k)}_{1:n}, X^{(k+1)}_{1:n} && \text{if oracle is first-order,}\end{aligned}$$

where $k \in 1 : (t-1)$. Then, the following holds for any choice of $u_{1:t-1}$, $u_{t:M}$, v and $\tilde{u}_{t:M}$, $\tilde{v}$:

$$\mathbb{P}^v_{u_{1:M}}(G \cap \Lambda_{1:k}) \le K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G \cap \Lambda_{1:k}) + K_2,$$

where the quantities K_1, K_2 are

$$\begin{aligned}K_1 &:= \exp\left(\frac{\sum_{j=0}^{\zeta}\sum_{m=1}^{k}\xi^{(j)}_m}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)})\sum_{m=1}^{k}h_m}{2\sigma^2}\right),\\ K_2 &:= 2\sum_{m=1}^{k}\sum_{j=0}^{\zeta}\exp\left(-\frac{(\xi^{(j)}_m)^2}{2\sigma^2 h_m \tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)}}\right),\end{aligned}$$

where $\xi^{(j)}_m$ for $m \in 1 : k$ and $j \in 0 : \zeta$ is any positive quantity we can choose.

The proof of Lemma 20 is deferred to Appendix D.9.

By assumption we have $u_{1:M}$ and v such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{t-1}\Lambda^{(m)}_{u_m}) \ge \frac{1}{4^{t-1}}$ for sufficiently large n. Then, re-choose any $\tilde{u}_{t:M}$ and $\tilde{v}$, and apply Lemma 20, with

$$k = t-2, G = \Lambda^{(t-1)}_{u_{t-1}}, \text{ and } \xi^{(j)}_m = h_m^{\frac{1}{4}}\delta_{t-1}^{\frac{\kappa-j}{2}} \text{ for } m \in 1 : t-2, j \in 0 : \zeta.$$

Then, we get

$$\mathbb{P}^v_{u_{1:M}}(\Lambda_{1:t-1}) \le K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(\Lambda_{1:t-1}) + K_2,$$

where the quantities K_1 and K_2 are

$$K_1 := \exp\left(\frac{(\sum_{j=0}^{\zeta}\delta_{t-1}^{\frac{\kappa-j}{2}})\sum_{m=1}^{t-2}h_m^{\frac{1}{4}}}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)})\sum_{m=1}^{t-2}h_m}{2\sigma^2}\right),$$

$$K_2 := 2\sum_{m=1}^{t-2}\sum_{j=0}^{\zeta}\exp\left(-\frac{1}{2\sigma^2 h_m^{\frac{1}{2}}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{\kappa-j}}\right).$$

By definitions of γ_t (21), δ_t (22), and h_t (25), whenever $m < t-1$, $h_m\delta_{t-1}^{2(\kappa-j)}$ polynomially decays to 0 as n increases. So, $K_1 \downarrow 1$ and $K_2 \downarrow 0$ as $n \to \infty$. Therefore, for large enough n, $K_1 \le \sqrt{2}$ and $K_2 \le \left(1-\frac{1}{\sqrt{2}}\right)\frac{1}{4^{t-1}}$, resulting in

$$\mathbb{P}^{v}_{u_{1:M}}(\Lambda_{1:t-1}) \le \sqrt{2}\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(\Lambda_{1:t-1}) + \left(1-\frac{1}{\sqrt{2}}\right)\frac{1}{4^{t-1}}. \tag{45}$$

From the assumption, we have $\mathbb{P}^{v}_{u_{1:M}}(\Lambda_{1:t-1}) \ge \frac{1}{4^{t-1}}$. Using this with Eq (45), we finish the proof of Lemma 9:

$$\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(\Lambda_{1:t-1}) \ge \frac{1}{2\cdot 4^{t-1}},$$

for any $\tilde{u}_{t:M}$ and $\tilde{v}$, as desired.

D.7. Proof of Lemma 10

For the proof of Lemmas 10 and 11, the following elementary lemma is useful.

Lemma 21 *If $a_1 \le ca_1' + b$ and $a_2' \le ca_2 + b$, then*

$$\frac{a_1}{a_1+a_2} \le \frac{c^2a_1'}{a_1'+a_2'} + \frac{cb}{a_1'+a_2'}$$

for any $a_1, a_2, a_1', a_2', b \ge 0$ and $a_1+a_2 > 0$, $a_1'+a_2' > 0$ and $c \ge 1$.

Proof The proof of the above lemma is straightforward calculation. Note that, the function $f(x) := \frac{x}{x+y}$ is an increasing function of $x > 0$ for any $y > 0$, so we have

$$\frac{a_1}{a_1+a_2} \le \frac{ca_1'+b}{ca_1'+a_2+b} \le \frac{ca_1'+b}{a_1'+a_2+b},$$

where in the last inequality, we use the assumption that $c \ge 1$. Now, notice that, the function $f(x) = \frac{y}{x+z}$ is a decreasing function of $x > 0$ for any positive number y, z, so we have,

$$\frac{c^2a_1'+cb}{a_1'+a_2'} \ge \frac{c^2a_1'+cb}{a_1'+ca_2+b} = \frac{ca_1'+b}{(a_1'+b)/c+a_2} \ge \frac{ca_1'+b}{a_1'+a_2+b}$$

where in the last inequality, we use the assumption that $c \ge 1$. Combining the two elementary inequalities above, we have shown the claim in the lemma. ■

By assumption we have $u_{1:M}$ and v such that $\mathbb{P}^v_{u_{1:M}}(\bigcap_{m=1}^{t-1}\Lambda^{(m)}_{u_m}) \geq \frac{1}{4^{t-1}}$ for sufficiently large n. For any re-chosen $\tilde{u}_{t:M}$ and $\tilde{v}$, let $E^{(t)}_{i,\tilde{u}_t}$ be an event on which the i-th sample of t-th round is in $\mathbf{B}^p_{\tilde{u}_t}(\delta_t)$:

$$E^{(t)}_{i,\tilde{u}_t} := \{X^{(t)}_i \in \mathbf{B}^p_{\tilde{u}_t}(\delta_t)\}, \text{ for } i \in 1:n.$$

Then, apply Lemma 20, with

$$\begin{aligned}
&k = t-1,\ G = (E^{(t)}_{i,\tilde{u}_t})^c,\\
&\xi^{(j)}_m = h_m^{\frac{1}{4}}\delta_{t-1}^{\frac{\kappa-j}{2}} \text{ for } m \in 1:t-2, j \in 0:\zeta,\\
&\xi^{(\zeta)}_{t-1} = \sqrt{2}\sigma^2\sqrt{\log n},\ \xi^{(0)}_{t-1} = h_{t-1}^{\frac{1}{4}}\delta_{t-1}^{\frac{\kappa}{2}} \text{ if } \zeta = 1.
\end{aligned}$$

Then, we get

$$\mathbb{P}^v_{u_{1:M}}((E^{(t)}_{i,\tilde{u}_t})^c \cap \Lambda_{1:t-1}) \leq K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}((E^{(t)}_{i,\tilde{u}_t})^c \cap \Lambda_{1:t-1}) + K_2,$$

where the quantities K_1 and K_2 are

$$\begin{aligned}
K_1 :=& \exp\left(\frac{(\sum_{j=0}^{\zeta}\delta_{t-1}^{\frac{\kappa-j}{2}})\sum_{m=1}^{t-2}h_m^{\frac{1}{4}}}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)})\sum_{m=1}^{t-2}h_m}{2\sigma^2}\right)\\
&\times \exp\left(\frac{\sqrt{2}\sigma^2\sqrt{\log n} + \mathbb{I}\{\zeta=1\}\,\delta_{t-1}^{\frac{\kappa}{2}}h_{t-1}^{\frac{1}{4}}}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)})h_{t-1}}{2\sigma^2}\right)\\
K_2 :=& 2\sum_{m=1}^{t-2}\sum_{j=0}^{\zeta}\exp\left(-\frac{1}{2\sigma^2 h_m^{\frac{1}{2}}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{\kappa-j}}\right)\\
&+ 2\exp\left(-\frac{\sigma^2\log n}{h_{t-1}\tilde{C}_\zeta^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-\zeta)}}\right) + \mathbb{I}\{\zeta=1\}\exp\left(-\frac{1}{2\sigma^2 h_{t-1}^{\frac{1}{2}}\tilde{C}_0^2\alpha^{2t-4}\delta_{t-1}^{\kappa}}\right).
\end{aligned}$$

As seen the proof of Lemma 9, the first multiplicative term in K_1 and the first additive term in K_2 go down to 1 and down to 0, respectively. Moreover, if $\zeta = 1$, then $h_t\delta_t^{2\kappa}$ also polynomially decays to zero. So, if $\zeta = 1$, the terms

$$\exp\left(\frac{\mathbb{I}\{\zeta=1\}\,\delta_{t-1}^{\frac{\kappa}{2}}h_{t-1}^{\frac{1}{4}}}{\sigma^2} + \frac{(\tilde{C}_0^2\alpha^{2t-4}\delta_{t-1}^{2\kappa})h_{t-1}}{2\sigma^2}\right) \downarrow 1, \text{ and } \exp\left(-\frac{1}{2\sigma^2 h_m^{\frac{1}{2}}\tilde{C}_0^2\alpha^{2t-4}\delta_{t-1}^{\kappa}}\right) \downarrow 0.$$

Also, by noting the identity $\tilde{C}_\zeta^2\alpha^{2t-4}h_t\delta_{t-1}^{2(\kappa-\zeta)} = \sigma^2$ from Eq (26),

$$\exp\left(\frac{\tilde{C}_\zeta^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-\zeta)}h_{t-1}}{2\sigma^2}\right) = \sqrt{e},\ 2\exp\left(-\frac{\sigma^2\log n}{h_{t-1}\tilde{C}_\zeta^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-\zeta)}}\right) = \frac{2}{n}.$$

Summarizing all these observations, for large enough n, we have

$$\mathbb{P}^v_{u_{1:M}}((E^{(t)}_{i,\tilde{u}_t})^c \cap \Lambda_{1:t-1}) \leq K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}((E^{(t)}_{i,\tilde{u}_t})^c \cap \Lambda_{1:t-1}) + K_2, \tag{46}$$

where

$$1 \le K_1 \le \sqrt{3e}\exp(\sqrt{2}\sqrt{\log n}),\ K_2 = \frac{2}{n} + O(\exp(-n^\tau)). \tag{47}$$

for some $\tau > 0$.

Note that, we can switch between $u_{t:M}$, v and $\tilde{u}_{t:M}$, $\tilde{v}$, and apply Lemma 20 again, this time for $G = E^{(t)}_{i,\tilde{u}_t}$. This time, the equation we get is

$$\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(E^{(t)}_{i,\tilde{u}_t} \cap \Lambda_{1:t-1}) \le K_1 \mathbb{P}^{v}_{u_{1:M}}(E^{(t)}_{i,\tilde{u}_t} \cap \Lambda_{1:t-1}) + K_2, \tag{48}$$

with the same K_1 and K_2. Now, apply lemma Lemma 21 for Eqs (48) and (46), then

$$\begin{aligned}\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(E^{(t)}_{i,\tilde{u}_t} \mid \Lambda_{1:t-1}) &\le K_1^2 \mathbb{P}^{v}_{u_{1:M}}(E^{(t)}_{i,\tilde{u}_t} \mid \Lambda_{1:t-1}) + \frac{K_1 K_2}{\mathbb{P}^{v}_{u_{1:M}}(\Lambda_{1:t-1})} \\ &\le K_1^2 \mathbb{P}^{v}_{u_{1:M}}(E^{(t)}_{i,\tilde{u}_t} \mid \Lambda_{1:t-1}) + 4^{t-1} K_1 K_2. \end{aligned} \tag{49}$$

Now we are ready to use the pigeonhold principle argument for this lemma. We then sum up both sides of Eq (49) for all possible values of $\tilde{u}_{t:M}$, $\tilde{v}$, and $i = 1 : n$. For simplicity in notation, let us denote the summation

$$\sum_{\tilde{u}_{t+1} \in \mathcal{U}^{(t+1)}_{\tilde{u}_t}} \cdots \sum_{\tilde{u}_M \in \mathcal{U}^{(M)}_{\tilde{u}_{M-1}}} \sum_{\tilde{v} \in \mathcal{V}} \equiv \sum_{\tilde{u}_{t+1:M},\tilde{v}}.$$

Also, recall that there are $2|\mathcal{U}^{(t)}_{u_{t-1}}| \prod_{m=t+1}^{M} |\mathcal{U}^{(m)}_{\tilde{u}_{m-1}}|$ possible values of $\tilde{u}_{t:M}$ and $\tilde{v}$. After summing up, we get

$$\begin{aligned}&\sum_{\tilde{u}_t} \sum_{\tilde{u}_{t+1:M},\tilde{v}} \sum_{i=1}^{n} \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(E^{(t)}_{i,\tilde{u}_t} \mid \Lambda_{1:t-1}) \\ =&\sum_{\tilde{u}_t} \sum_{\tilde{u}_{t+1:M},\tilde{v}} \mathbb{E}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}} \left(\sum_{i=1}^{n} \mathbb{I}\left\{ X_i^{(t)} \in \mathbf{B}^p_{\tilde{u}_t}(\delta_t) \right\} \mid \Lambda_{1:t-1} \right) \\ \le& K_1^2 \sum_{i=1}^{n} \sum_{\tilde{u}_t} \sum_{\tilde{u}_{t+1:M},\tilde{v}} \mathbb{P}^{v}_{u_{1:M}}(E^{(t)}_{i,\tilde{u}_t} \mid \Lambda_{1:t-1}) + 4^{t-1} K_1 K_2 \cdot 2n |\mathcal{U}^{(t)}_{u_{t-1}}| \prod_{m=t+1}^{M} |\mathcal{U}^{(m)}_{\tilde{u}_{m-1}}| \\ \le& K_1^2 \cdot 2n \prod_{m=t+1}^{M} |\mathcal{U}^{(m)}_{\tilde{u}_{m-1}}| + 4^{t-1} K_1 K_2 \cdot 2n |\mathcal{U}^{(t)}_{u_{t-1}}| \prod_{m=t+1}^{M} |\mathcal{U}^{(m)}_{\tilde{u}_{m-1}}|, \end{aligned} \tag{50}$$

where the last inequality used that $\sum_{\tilde{u}_t} \mathbb{P}^{v}_{u_{1:M}}(E^{(t)}_{i,\tilde{u}_t} \mid \Lambda_{1:t-1}) \le 1$.

Since there are $2|\mathcal{U}^{(t)}_{u_{t-1}}| \prod_{m=t+1}^{M} |\mathcal{U}^{(m)}_{\tilde{u}_{m-1}}|$ possible values of $\tilde{u}_{t:M}$ and $\tilde{v}$, Equation 50 implies that by the pigeonhole principle, there exists at least one set of parameters $\tilde{u}_{t:M}$ and $\tilde{v}$ that satisfies

$$\begin{aligned}&\mathbb{E}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}} \left(\sum_{i=1}^{n} \mathbb{I}\left\{ X_i^{(t)} \in \mathbf{B}^p_{\tilde{u}_t}(\delta_t) \right\} \mid \Lambda_{1:t-1} \right) \\ \le& \frac{nK_1^2}{|\mathcal{U}^{(t)}_{u_{t-1}}|} + 4^{t-1} n K_1 K_2 \le n K_1^2 \left(\frac{4\delta_t}{\eta \delta_{t-1}} \right)^d + 4^{t-1} n K_1 K_2, \end{aligned}$$

where Lemma 8 is used in the last inequality. This implies, by Markov's inequality,

$$\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(\Lambda^{(t)}_{\tilde{u}_t} \mid \Lambda_{1:t-1}) \geq 1 - \frac{1}{h_t}\mathbb{E}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}\left(\sum_{i=1}^{n}\mathbb{I}\left\{X_i^{(t)} \in \mathbf{B}^p_{\tilde{u}_t}(\delta_t)\right\} \mid \Lambda_{1:t-1}\right)$$
$$\geq 1 - \frac{nK_1^2}{h_t}\left(\frac{4\delta_t}{\eta\delta_{t-1}}\right)^d - \frac{4^{t-1}nK_1K_2}{h_t}.$$

To finish the proof of Lemma 10, it suffices to show that the RHS of the last inequality is greater than or equal to $\frac{1}{2}$, which is equivalent to

$$nK_1^2\left(\frac{4\delta_t}{\eta\delta_{t-1}}\right)^d + 4^{t-1}nK_1K_2 \leq \frac{h_t}{2}. \tag{51}$$

Recall from Eq (47) that $1 \leq K_1 \leq \sqrt{3e}\exp(\sqrt{2}\sqrt{\log n})$ and $K_2 = \frac{2}{n} + O(\exp(-n^\tau))$. Substituting these to the first term in the LHS of Eq (51) yields

$$nK_1^2\left(\frac{4\delta_t}{\eta\delta_{t-1}}\right)^d \leq 3en\exp\left(2\sqrt{2}\sqrt{\log n}\right)\left(\frac{4\delta_t}{\eta\delta_{t-1}}\right)^d.$$

From the definition of γ_t (21), δ_t (22), and D_t (23), we can get useful identities

$$\left(\frac{4D_t}{\eta D_{t-1}}\right)^d = \frac{\sigma^2}{8e\tilde{C}_\zeta^2\alpha^{2t-2}D_t^{2(\kappa-\zeta)}}, \text{ and}$$
$$1 - d(\gamma_t - \gamma_{t-1}) = 1 - \left(\frac{d}{d+2(\kappa-\zeta)}\right)^t = 2(\kappa-\zeta)\gamma_t,$$

thereby one can get

$$3en\exp\left(2\sqrt{2}\sqrt{\log n}\right)\left(\frac{4\delta_t}{\eta\delta_{t-1}}\right)^d$$
$$= \frac{3\sigma^2}{8C_\zeta^2\alpha^{2t-2}D_t^{2(\kappa-\zeta)}}n^{2(\kappa-\zeta)\gamma_t}\exp\left(4\sqrt{2}(\kappa-\zeta)\gamma_{t-1}\sqrt{\log n}\right)(\log n)^{-\mathbb{I}\{t=M\}d\nu/\kappa}.$$

comparing with h_t (25), we check that

$$nK_1^2\left(\frac{4\delta_t}{\eta\delta_{t-1}}\right)^d \leq \frac{3}{8}h_t. \tag{52}$$

Now, substitute $K_1 \leq \sqrt{3e}\exp(\sqrt{2}\sqrt{\log n})$ and $K_2 = \frac{2}{n} + O(\exp(-n^\tau))$ to the second term in the LHS of Eq (51), the we get:

$$4^{t-1}nK_1K_2 \leq 2\cdot 4^{t-1}\sqrt{3e}\exp\left(\sqrt{2}\sqrt{\log n}\right)\left(1 + O(n\exp(-n^\tau))\right).$$

Since $O(n\exp(-n^\tau)) \downarrow 0$ as $n \to \infty$, we can see that the first term in LHS of Eq (51) dominates the second one for sufficiently large n. With the observation in Eq (52), we can see Eq (51) holds for large enough n, so this finishes the proof of Lemma 10.

D.8. Proof of Lemma 11

For this lemma, we present a variant of Lemma 20, for fixed $u_{1:M}$ and different v and $\tilde{v}$. Lemma 22 is just a simple variant of Lemma 20 for the case $t = M + 1$ (final case), so its proof is omitted.

Lemma 22 *Pick any set of parameters $u_{1:M}$, and $v, \tilde{v} \in \mathcal{V}$ where $v \neq \tilde{v}$. Consider any probabilistic event G that is a function of the random variables*

$$\begin{aligned} &X_{1:n}^{(1:k)}, Y_{1:n}^{(1:k)}, X_{1:n}^{(k+1)} && \text{if oracle is zeroth-order, or} \\ &X_{1:n}^{(1:k)}, Y_{1:n}^{(1:k)}, Z_{1:n}^{(1:k)}, X_{1:n}^{(k+1)} && \text{if oracle is first-order,} \end{aligned}$$

where $k \in 1 : M-1$. Otherwise, the probabilistic event G could also be a function of $X_{1:n}^{(1:M)}, Y_{1:n}^{(1:M)}$, and/or $Z_{1:n}^{(1:M)}$, in which case we let $k = M$. Then, the following holds for any $u_{1:M}$ and $v \neq \tilde{v}$.

$$\mathbb{P}_{u_{1:M}}^{v}(G \cap \Lambda_{1:k}) \leq K_1 \mathbb{P}_{u_{1:M}}^{\tilde{v}}(G \cap \Lambda_{1:k}) + K_2,$$

where the quantities K_1, K_2 are

$$\begin{aligned} K_1 &:= \exp\left(\frac{\sum_{j=0}^{\zeta} \sum_{m=1}^{k} \xi_m^{(j)}}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta} \tilde{C}_j^2 \alpha^{2M-2} \delta_M^{2(\kappa-j)}) \sum_{m=1}^{k} h_m}{2\sigma^2} \right), \\ K_2 &:= 2 \sum_{m=1}^{k} \sum_{j=0}^{\zeta} \exp\left(- \frac{(\xi_m^{(j)})^2}{2\sigma^2 h_m \tilde{C}_j^2 \alpha^{2M-2} \delta_M^{2(\kappa-j)}} \right), \end{aligned}$$

where $\xi_m^{(j)}$ for $m \in 1 : k$ and $j \in 0 : \zeta$ is any positive quantity we can choose.

To prove the first statement of Lemma 11, we can repeat the same process as Lemma 9, this time with Lemma 22, and

$$k = M - 1, G = \Lambda_{u_M}^{(M)}, \text{ and } \xi_m^{(j)} = h_m^{\frac{1}{4}} \delta_{t-1}^{\frac{\kappa-j}{2}} \text{ for } m \in 1 : M-1, j \in 0 : \zeta.$$

Then, we get

$$\mathbb{P}_{u_{1:M}}^{v}(\Lambda_{1:M}) \leq K_1 \mathbb{P}_{u_{1:M}}^{\tilde{v}}(\Lambda_{1:M}) + K_2,$$

where the quantities $K_1 \downarrow 1$ and $K_2 \downarrow 0$ as $n \to \infty$.

Therefore, for large enough n, $K_1 \leq \sqrt{2}$ and $K_2 \leq \left(1 - \frac{1}{\sqrt{2}}\right) \frac{1}{4^{t-1}}$, resulting in

$$\mathbb{P}_{u_{1:M}}^{v}(\Lambda_{1:M}) \leq \sqrt{2} \mathbb{P}_{u_{1:M}}^{\tilde{v}}(\Lambda_{1:M}) + \left(1 - \frac{1}{\sqrt{2}}\right) \frac{1}{4^M}. \tag{53}$$

From the assumption, we have $\mathbb{P}_{u_{1:M}}^{v}(\Lambda_{1:M}) \geq \frac{1}{4^M}$. With these observations, we finish the proof of the first part of Lemma 11:

$$\mathbb{P}_{u_{1:M}}^{\tilde{v}}(\Lambda_{1:M}) \geq \frac{1}{2 \cdot 4^M},$$

for $\tilde{v} \neq v$, as desired.

Now, the last goal is to show that the total variation between $\mathbb{P}^{-1}_{u_{1:M}}$ and $\mathbb{P}^{+1}_{u_{1:M}}$ conditional on the event $\Lambda_{1:M}$ is small when the number of sample size n is sufficiently large. The technique is essentially the same as that appeared in the proof of lemma 10. We apply Lemma 22, with

$$k = M,\ G = \text{any } G,\ \xi_m^{(j)} = h_m^{\frac{1}{4}} \delta_M^{\frac{\kappa - j}{2}} \text{ for } m \in 1:M, j \in 0:\zeta.$$

Then, we get

$$\mathbb{P}^{v}_{u_{1:M}}(G \cap \Lambda_{1:M}) \le K_1 \mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \cap \Lambda_{1:M}) + K_2,$$

where the quantities K_1 and K_2 are

$$K_1 := \exp\left(\frac{(\sum_{j=0}^{\zeta} \delta_M^{\frac{\kappa-j}{2}}) \sum_{m=1}^{M} h_m^{\frac{1}{4}}}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta} \tilde{C}_j^2 \alpha^{2M-2} \delta_M^{2(\kappa-j)}) \sum_{m=1}^{M} h_m}{2\sigma^2} \right)$$

$$K_2 := 2 \sum_{m=1}^{M} \sum_{j=0}^{\zeta} \exp\left(- \frac{1}{2\sigma^2 h_m^{\frac{1}{2}} \tilde{C}_j^2 \alpha^{2M-2} \delta_M^{\kappa-j}} \right).$$

In this case, we can note that since $\tilde{C}_\zeta^2 \alpha^{2M-2} h_M \delta_M^{2(\kappa-\zeta)} = \sigma^2 \log^{-\frac{2\nu(\kappa-\zeta)}{\kappa}} n$ (26), any $h_m \delta_M^{2(\kappa-j)}$ decreases to zero with $n \to \infty$. So, we can see $K_1 \downarrow 1$ and $K_2 \downarrow 0$ as $n \to \infty$, which allows us to write

$$\mathbb{P}^{v}_{u_{1:M}}(G \cap \Lambda_{1:M}) \le \sqrt{\frac{5}{4}} \mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \cap \Lambda_{1:M}) + \frac{1}{\sqrt{5} \cdot 4^{M+1}}, \tag{54}$$

for n sufficiently large. Using exactly the same techniques, we can also get similar inequalities as follows:

$$\mathbb{P}^{\tilde{v}}_{u_{1:M}}(G^c \cap \Lambda_{1:M}) \le \sqrt{\frac{5}{4}} \mathbb{P}^{v}_{u_{1:M}}(G^c \cap \Lambda_{1:M}) + \frac{1}{\sqrt{5} \cdot 4^{M+1}} \tag{55}$$

$$\mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \cap \Lambda_{1:M}) \le \sqrt{\frac{5}{4}} \mathbb{P}^{v}_{u_{1:M}}(G \cap \Lambda_{1:M}) + \frac{1}{\sqrt{5} \cdot 4^{M+1}} \tag{56}$$

$$\mathbb{P}^{v}_{u_{1:M}}(G^c \cap \Lambda_{1:M}) \le \sqrt{\frac{5}{4}} \mathbb{P}^{\tilde{v}}_{u_{1:M}}(G^c \cap \Lambda_{1:M}) + \frac{1}{\sqrt{5} \cdot 4^{M+1}} \tag{57}$$

Now apply Lemma 21 to Eqs (54) and (55) to get

$$\mathbb{P}^{v}_{u_{1:M}}(G \mid \Lambda_{1:M}) \le \frac{5}{4} \mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \mid \Lambda_{1:M}) + \frac{1}{2 \cdot 4^{M+1} \mathbb{P}^{\tilde{v}}_{u_{1:M}}(\Lambda_{1:M})} \le \frac{5}{4} \mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \mid \Lambda_{1:M}) + \frac{1}{4}.$$

Similarly, from Eqs (56) and (57),

$$\mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \mid \Lambda_{1:M}) \le \frac{5}{4} \mathbb{P}^{v}_{u_{1:M}}(G \mid \Lambda_{1:M}) + \frac{1}{4}.$$

From these two equations, we have

$$\left| \mathbb{P}^{v}_{u_{1:M}}(G \mid \Lambda_{1:M}) - \mathbb{P}^{\tilde{v}}_{u_{1:M}}(G \mid \Lambda_{1:M}) \right| \le \frac{1}{2},$$

for any G and $v \neq \tilde{v}$. Thus, by definition of total variation distance, we have

$$\left\| \mathbb{P}^{-1}_{u_{1:M}}(\cdot \mid \Lambda_{1:M}) - \mathbb{P}^{+1}_{u_{1:M}}(\cdot \mid \Lambda_{1:M}) \right\|_{\mathrm{TV}} \le \frac{1}{2},$$

as desired.

D.9. Proof of Lemma 20

For $t \in 2:M$ and $k \in 1:t-1$, consider any event G that is a function of

$$\begin{aligned} &X_{1:n}^{(1:k)}, Y_{1:n}^{(1:k)}, X_{1:n}^{(k+1)} && \text{if oracle is zeroth-order, or} \\ &X_{1:n}^{(1:k)}, Y_{1:n}^{(1:k)}, Z_{1:n}^{(1:k)}, X_{1:n}^{(k+1)} && \text{if oracle is first-order.} \end{aligned}$$

In this proof, we develop a generic technique that provides an upper bound of the probability $G \cap \bigcap_{m=1}^{k} \Lambda_{u_m}^{(m)}$ under measure $\mathbb{P}_{u_{1:M}}^{v}$ in terms of the probability of the same event under another measure $\mathbb{P}_{u_{1:t-1},\tilde{u}_{t:M}}^{\tilde{v}}$ based on a different function. Recall that the two probability measure are defined by their corresponding functions $f_{u_{1:M}}^{v}$ and $f_{u_{1:t-1},\tilde{u}_{t:M}}^{\tilde{v}}$ that only differ in the interior of $\mathbf{B}_{u_{t-1}}^{p}(\delta_{t-1})$. Note that, the conclusion stated in the lemma provides relationship between the probabilities of an event under different measures and therefore the technique used in this proof can be of independent interest to the readers.

Notation. To give a clear illustration of how we make that bound happen, we need to introduce some notation, mainly to "partition" the events $\Lambda_{u_m}^{(m)}$. For $m \in 1:k$, we enumerate and index all the possible 2^n subsets of $\{1, 2, \ldots, n\}$ by S_{l_m}, where $l_m \in 1:2^n$. Also, for all $m \in 1:k$ and $i \in 1:n$, define

$$\begin{aligned} \epsilon_i^{(m,0)} &= Y_i^{(m)} - f_{u_{1:M}}^{v}(X_i^{(m)}), \\ \epsilon_i^{(m,1)} &= Z_i^{(m)} - \nabla f_{u_{1:M}}^{v}(X_i^{(m)}), \\ \Delta_i^{(m,0)} &= f_{u_{1:M}}^{v}(X_i^{(m)}) - f_{u_{1:t-1},\tilde{u}_{t:M}}^{\tilde{v}}(X_i^{(m)}), \\ \Delta_i^{(m,1)} &= \nabla f_{u_{1:M}}^{v}(X_i^{(m)}) - \nabla f_{u_{1:t-1},\tilde{u}_{t:M}}^{\tilde{v}}(X_i^{(m)}). \end{aligned}$$

Here, whenever the function is not differentiable at x, we replace $\nabla f_{u_{1:M}}^{v}(x)$ with any subgradient of $f_{u_{1:M}}^{v}$ at x. To interpret these quantities, based on the assumption that $f_{u_{1:M}}^{v}$ is the true function, $\epsilon_i^{(m,0)}$ is the (i.i.d. Gaussian) error of the zeroth order oracle at $X_i^{(m)}$, and $\Delta_i^{(m,0)}$ is the difference between two functions $f_{u_{1:M}}^{v}$ and $f_{u_{1:t-1},\tilde{u}_{t:M}}^{\tilde{v}}$ at $X_i^{(m)}$. Similarly, $\epsilon_i^{(m,1)}$ and $\Delta_i^{(m,1)}$ are the error and difference in the first order information (gradient values) at $X_i^{(m)}$. Note that $\epsilon_i^{(m,0)}$ and $\Delta_i^{(m,0)}$ are scalars while the other two are vectors. Note that, by Condition 5.1, $\Delta_i^{(m,0)}$ and $\Delta_i^{(m,1)}$ are zero when $X_i^{(m)} \notin \mathbf{B}_{u_{t-1}}^{p}(\delta_{t-1})$. Also note that, based on the assumption that $f_{u_{1:t-1},\tilde{u}_{t:M}}^{\tilde{v}}$ is the true function, then $\epsilon_i^{(m,0)} + \Delta_i^{(m,0)}$ and $\epsilon_i^{(m,1)} + \Delta_i^{(m,1)}$ are the errors of the oracle.

Next, for each set S_{l_m}, we introduce the following groups of events,

$$\begin{aligned} \Gamma_{u_m,l_m}^{(m)} &:= \left\{ X_i^{(m)} \in \mathbf{B}_{u_m}^{p}(\delta_m),\ \forall i \in S_{l_m} \right\} \bigcap \left\{ X_i^{(m)} \notin \mathbf{B}_{u_m}^{p}(\delta_m),\ \forall i \notin S_{l_m} \right\}, \quad \text{for } m \in 1:k, \\ \Xi_{u_m,l_m}^{(m)} &:= \bigcap_{j=0}^{\zeta} \Xi_{u_m,l_m}^{(m,j)}, \quad \text{for } m \in 1:k, \\ \Xi_{u_m,l_m}^{(m,j)} &:= \left\{ \left| \sum_{i \in S_{l_m}} \langle \epsilon_i^{(m,j)}, \Delta_i^{(m,j)} \rangle \right| \le \xi_m^{(j)} \right\}, \quad \text{for } m \in 1:k, j \in 0:1. \end{aligned}$$

The event $\Gamma_{u_m,l_m}^{(m)}$ occurs when $X_i^{(m)}$ is in the ball $\mathbf{B}_{u_m}^{p}(\delta_m)$ if and only if the index $i \in S_{l_m}$. Recall that ζ represents the order of the oracle. So if we are using zeroth-order oracle, $\Xi_{u_m,l_m}^{(m)} = \Xi_{u_m,l_m}^{(m,0)}$,

while $\Xi^{(m)}_{u_m,l_m} = \Xi^{(m,0)}_{u_m,l_m} \cap \Xi^{(m,1)}_{u_m,l_m}$ for first-order oracles. The event $\Xi^{(m,j)}_{u_m,l_m}$ is that the sum of noise introduced by the oracle is smaller than some positive quantity $\xi^{(j)}_m$ which we can choose. Recall that $\xi^{(j)}_m$ appeared in the statement of the lemma.

Partitioning $\Gamma^{(m)}_{u_m,l_m}$ into disjoint events. Notice the events $\Gamma^{(m)}_{u_m,l_m}$ are disjoint from one another for different $l_m \in 1:2^n$ because it can never happen at the same time with different values of l_m. From this observation, we can get a partition of the event $\bigcap_{m=1}^k \Lambda^{(m)}_{u_m}$, as the following equation suggests,

$$\bigcap_{m=1}^k \Lambda^{(m)}_{u_m} = \bigcap_{m=1}^k \left(\bigcup_{l_m} \left(\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \right) \right) = \bigcup_{l_1,l_2,\ldots,l_k} \left(\bigcap_{m=1}^k \left(\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \right) \right)$$

From the above equation, we note that in order to establish an upper bound of $\mathbb{P}^v_{u_{1:M}} \left(G \cap \bigcap_{m=1}^k \Lambda^{(m)}_{u_m} \right)$, it suffices to get an upper bound of $\mathbb{P}^v_{u_{1:M}} \left(G \cap \bigcap_{m=1}^k \left(\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \right) \right)$ for any fixed sequence $\{l_m\}_{m=1}^k \in \{1:2^n\}^k$ and then do the summation over all possible $\{l_m\}_{m=1}^k$. As we will see later, when restricted on the set $\bigcap_{m=1}^k \Gamma^{(m)}_{u_m,l_m}$, we can give an explicit form of the likelihood ratio between the two probability measures $\mathbb{P}^v_{u_{1:M}}$ and $\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}$, which greatly helps us analyze the relationship between the two probabilities that are computed under different measures.

Now *fix any* $\{l_m\}_{m=1}^k$. For the sake of simplicity in notation, let us denote

$$\Lambda_{1:k} = \bigcap_{m=1}^k \Lambda^{(m)}_{u_m}, \quad \Gamma_{1:k} = \bigcap_{m=1}^k \Gamma^{(m)}_{u_m,l_m}, \quad \Xi_{1:k} = \bigcap_{m=1}^k \Xi^{(m)}_{u_m,l_m}.$$

Expressed in this compact form, we want to get an upper bound for $\mathbb{P}^v_{u_{1:M}} (G \cap \Gamma_{1:k} \cap \Lambda_{1:k})$. Starting from this point, we have the following inequality:

$$\begin{aligned}
&\mathbb{P}^v_{u_{1:M}} (G \cap \Gamma_{1:k} \cap \Lambda_{1:k}) \\
=&\mathbb{P}^v_{u_{1:M}} (G \cap \Gamma_{1:k} \cap \Lambda_{1:k} \cap \Xi_{1:k}) + \mathbb{P}^v_{u_{1:M}} (G \cap \Gamma_{1:k} \cap \Lambda_{1:k} \cap (\Xi_{1:k})^c) \\
\le&\mathbb{P}^v_{u_{1:M}} (G \cap \Gamma_{1:k} \cap \Lambda_{1:k} \cap \Xi_{1:k}) + \sum_{j=0}^{\zeta} \sum_{m=1}^k \mathbb{P}^v_{u_{1:M}} \left(\Gamma_{1:k} \cap \Lambda_{1:m} \cap (\Xi^{(m,j)}_{u_m,l_m})^c \right). \qquad (58)
\end{aligned}$$

Note that the RHS of the Eq (58) consists of two parts. The rest of the proof consists of two parts: we first bound the first term of RHS of the Eq (58), and the bound the second term while summing up disjoint events together.

Bounding the first term of Eq (58). We first give an upper bound of the first term using the likelihood ratio between two different measures. Recall from Condition 5.1 that the functions $f^v_{u_{1:M}}$ and $f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}$ have the same values at all $x \notin \mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$. Given the event $\Gamma^{(m)}_{u_m,l_m}$, for $m \in 1:k$ and any $i \notin S_{l_m}$, we have

$$\begin{aligned}
p^v_{u_{1:M}}(y^{(m)}_i \mid x^{(m)}_i) &= p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_i \mid x^{(m)}_i) && \text{if } \zeta = 0, \\
p^v_{u_{1:M}}(y^{(m)}_i, z^{(m)}_i \mid x^{(m)}_i) &= p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_i, z^{(m)}_i \mid x^{(m)}_i) && \text{if } \zeta = 1.
\end{aligned}$$

Also recall the definition of $\tilde{C}_\zeta = C(1-\beta)^{1-\zeta}(2\kappa)^\zeta$ from Eq (20) and Condition 5.2–3 that

$$|f^v_{u_{1:M}}(x) - f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)| \le \tilde{C}_0 \alpha^{t-2}\delta_{t-1}^\kappa,\ \forall x \in \mathbf{B}^p_{u_{t-1}}(\delta_{t-1}) \subset \mathbf{B}^p_{u_m}(\delta_m)$$
$$\left\|\nabla f^v_{u_{1:M}}(x) - \nabla f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x)\right\|_2 \le \tilde{C}_1 \alpha^{t-2}\delta_{t-1}^{\kappa-1},\ \forall x \in \mathbf{B}^p_{u_{t-1}}(\delta_{t-1}) \subset \mathbf{B}^p_{u_m}(\delta_m).$$

For $i \in S_{l_m}$, the ratio between $p^v_{u_{1:M}}$ and $p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}$ is the ratio between two Gaussian distributions. Thus, if $\zeta = 0$ (zeroth-order oracle), we have

$$\begin{aligned}
&\mathbb{I}\left\{\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \cap \Xi^{(m)}_{u_m,l_m}\right\} \frac{p^v_{u_{1:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})}{p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})} \\
=&\mathbb{I}\left\{\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \cap \Xi^{(m)}_{u_m,l_m}\right\} \exp\left(\frac{1}{\sigma^2}\sum_{i\in S_{l_m}} \epsilon_i^{(m,0)}\Delta_i^{(m,0)} + \frac{1}{2\sigma^2}\sum_{i\in S_{l_m}}(\Delta_i^{(m,0)})^2\right) \\
\le&\mathbb{I}\left\{\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m}\right\} \exp\left(\frac{\xi_m^{(0)}}{\sigma^2} + \frac{h_m\tilde{C}_0^2\alpha^{2t-4}\delta_{t-1}^{2\kappa}}{2\sigma^2}\right),
\end{aligned} \tag{59}$$

where the last inequality used the definitions of $\Xi^{(m)}_{u_m,l_m}$ and $\Lambda^{(m)}_{u_m}$, and Condition 5.2. For $\zeta = 1$ (first-order oracle), note that

$$p^v_{u_{1:M}}(y^{(m)}_{1:n}, z^{(m)}_{1:n} \mid x^{(m)}_{1:n}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i\in S_{l_m}}(\epsilon_i^{(m,0)})^2 - \frac{1}{2\sigma^2}\sum_{i\in S_{l_m}}\left\|\epsilon_i^{(m,1)}\right\|_2^2\right)$$

By a similar argument as the zeroth-order case, we can get

$$\begin{aligned}
&\mathbb{I}\left\{\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \cap \Xi^{(m)}_{u_m,l_m}\right\} \frac{p^v_{u_{1:M}}(y^{(m)}_{1:n}, z^{(m)}_{1:n} \mid x^{(m)}_{1:n})}{p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_{1:n}, z^{(m)}_{1:n} \mid x^{(m)}_{1:n})} \\
=&\mathbb{I}\left\{\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \cap \Xi^{(m)}_{u_m,l_m}\right\} \exp\left(\frac{1}{\sigma^2}\sum_{i\in S_{l_m}}\sum_{j=0}^{1}\langle\epsilon_i^{(m,j)}, \Delta_i^{(m,j)}\rangle + \frac{1}{2\sigma^2}\sum_{i\in S_{l_m}}\sum_{j=0}^{1}\left\|\Delta_i^{(m,j)}\right\|_2^2\right) \\
\le&\mathbb{I}\left\{\Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m}\right\} \exp\left(\frac{\xi_m^{(0)} + \xi_m^{(1)}}{\sigma^2} + \frac{h_m\tilde{C}_0^2\alpha^{2t-4}\delta_{t-1}^{2\kappa} + h_m\tilde{C}_1^2\alpha^{2t-4}\delta_{t-1}^{2\kappa-2}}{2\sigma^2}\right),
\end{aligned} \tag{60}$$

where the last inequality used the definitions of $\Xi^{(m)}_{u_m,l_m}$ and $\Lambda^{(m)}_{u_m}$, and Condition 5.2–3.

Now consider again the zeroth-order case. For any event E that is a function of random variables

$$X^{(1:k)}_{1:n}, Y^{(1:k)}_{1:n}, X^{(k+1)}_{1:n},$$

the probability $\mathbb{P}^v_{u_{1:M}}(E)$ can be expressed as

$$\begin{aligned}
&\mathbb{P}^v_{u_{1:M}}(E) = \mathbb{E}^v_{u_{1:M}}[\mathbb{I}\{E\}] \\
=&\int \mathbb{I}\{E\}\, dQ^{(k+1)}(x^{(k+1)}_{1:n} \mid x^{(1:k)}_{1:n}, y^{(1:k)}_{1:n}) dP^v_{u_{1:M}}(y^{(k)}_{1:n} \mid x^{(k)}_{1:n}) dQ^{(k)}(x^{(k)}_{1:n} \mid x^{(1:k-1)}_{1:n}, y^{(1:k-1)}_{1:n})
\end{aligned}$$

$$
\begin{aligned}
&\times \cdots \times dP^{v}_{u_{1:M}}(y^{(1)}_{1:n} \mid x^{(1)}_{1:n}) dQ^{(1)}(x^{(1)}_{1:n}). \\
=&\int \mathbb{I}\{E\} \prod_{m=1}^{k} \frac{dP^{v}_{u_{1:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})}{dP^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})} dQ^{(k+1)}(x^{(k+1)}_{1:n} \mid x^{(1:k)}_{1:n}, y^{(1:k)}_{1:n}) dP^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(k)}_{1:n} \mid x^{(k)}_{1:n}) \\
&dQ^{(k)}(x^{(k)}_{1:n} \mid x^{(1:k-1)}_{1:n}, y^{(1:k-1)}_{1:n}) \times \cdots \times dP^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(1)}_{1:n} \mid x^{(1)}_{1:n}) dQ^{(1)}(x^{(1)}_{1:n}) \\
=&\mathbb{E}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}} \left[\mathbb{I}\{E\} \prod_{m=1}^{k} \frac{p^{v}_{u_{1:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})}{p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})} \right]. \qquad (61)
\end{aligned}
$$

Substituting $E = G \cap \Gamma_{1:k} \cap \Lambda_{1:k} \cap \Xi_{1:k}$ to Eq 61 gives

$$
\begin{aligned}
&\mathbb{P}^{v}_{u_{1:M}}(G \cap \Gamma_{1:k} \cap \Lambda_{1:k} \cap \Xi_{1:k}) \\
=&\mathbb{E}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}} \left[\mathbb{I}\{G\} \prod_{m=1}^{k} \left(\mathbb{I}\left\{ \Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \cap \Xi^{(m)}_{u_m,l_m} \right\} \frac{p^{v}_{u_{1:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})}{p^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(y^{(m)}_{1:n} \mid x^{(m)}_{1:n})} \right) \right] \\
\leq&\mathbb{E}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}} \left[\mathbb{I}\{G\} \prod_{m=1}^{k} \left(\mathbb{I}\left\{ \Gamma^{(m)}_{u_m,l_m} \cap \Lambda^{(m)}_{u_m} \right\} \exp\left(\frac{\xi^{(0)}_m}{\sigma^2} + \frac{h_m \tilde{C}_0^2 \alpha^{2t-4} \delta_{t-1}^{2\kappa}}{2\sigma^2} \right) \right) \right] \\
=&\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G \cap \Gamma_{1:k} \cap \Lambda_{1:k}) \exp\left(\frac{\sum_{m=1}^{k} \xi^{(0)}_m}{\sigma^2} + \frac{\tilde{C}_0^2 \alpha^{2t-4} \delta_{t-1}^{2\kappa} \sum_{m=1}^{k} h_m}{2\sigma^2} \right),
\end{aligned}
$$

where the inequality used Eq (59). We can get a similar upper bound for the first-order case using Eq (60), and in fact, we can express both cases into a unified form using ζ:

$$
\mathbb{P}^{v}_{u_{1:M}}(G \cap \Gamma_{1:k} \cap \Lambda_{1:k} \cap \Xi_{1:k}) \leq K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G \cap \Gamma_{1:k} \cap \Lambda_{1:k}), \qquad (62)
$$

where we define

$$
K_1 := \exp\left(\frac{\sum_{j=0}^{\zeta} \sum_{m=1}^{k} \xi^{(j)}_m}{\sigma^2} + \frac{(\sum_{j=0}^{\zeta} \tilde{C}_j^2 \alpha^{2t-4} \delta_{t-1}^{2(\kappa-j)}) \sum_{m=1}^{k} h_m}{2\sigma^2} \right),
$$

so simplify the notation. By this, we finished bounding the first term of RHS in Eq (58).

Bounding the second term of Eq (58) and summing disjoint events. Now, we need to deal with the second term. Note that

$$
\begin{aligned}
&\sum_{j=0}^{\zeta} \sum_{m=1}^{k} \mathbb{P}^{v}_{u_{1:M}} \left(\Gamma_{1:k} \cap \Lambda_{1:m} \cap (\Xi^{(m,j)}_{u_m,l_m})^c \right) \\
=&\sum_{j=0}^{\zeta} \sum_{m=1}^{k-1} \mathbb{P}^{v}_{u_{1:M}} \left(\Gamma_{1:k} \cap \Lambda_{1:m} \cap (\Xi^{(m,j)}_{u_m,l_m})^c \right) + \sum_{j=0}^{\zeta} \mathbb{P}^{v}_{u_{1:M}} \left(\Gamma_{1:k} \cap \Lambda_{1:k} \cap (\Xi^{(k,j)}_{u_k,l_k})^c \right) \\
=&\sum_{j=0}^{\zeta} \sum_{m=1}^{k-1} \mathbb{P}^{v}_{u_{1:M}} \left(\Gamma_{1:k} \cap \Lambda_{1:m} \cap (\Xi^{(m,j)}_{u_m,l_m})^c \right) \\
&+ \mathbb{P}^{v}_{u_{1:M}} (\Gamma_{1:k} \cap \Lambda_{1:k}) \sum_{j=0}^{\zeta} \mathbb{P}^{v}_{u_{1:M}} \left((\Xi^{(k,j)}_{u_k,l_k})^c \mid \Gamma_{1:k} \cap \Lambda_{1:k} \right) \qquad (63)
\end{aligned}
$$

We can use a concentration inequality of Gaussian random variables to bound above the conditional probability term. We should first note that once $\Gamma^{(k)}_{u_k,l_k}$ and $\Lambda^{(k)}_{u_k}$ are given, this means that at most h_k among $X^{(k)}_i$ are in $\mathbf{B}^p_{u_{t-1}}(\delta_{t-1})$, so

$$\sum_{i\in S_{l_k}} (\Delta^{(k,0)}_i)^2 \le h_k \tilde{C}_0^2 \alpha^{2t-4}\delta_{t-1}^{2\kappa} \text{ and } \sum_{i\in S_{l_k}} \left\|\Delta^{(k,1)}_i\right\|_2^2 \le h_k \tilde{C}_1^2 \alpha^{2t-4}\delta_{t-1}^{2(\kappa-1)}.$$

Since the observation noise is independent zero-mean Gaussian with variance σ^2, given that the true function is $f^v_{u_{1:M}}$ we have $\epsilon^{(m,0)}_i \sim \mathcal{N}(0,\sigma^2)$ and $\epsilon^{(m,1)}_i \sim \mathcal{N}(0,\sigma^2 I_d)$. Now, we can apply the concentration inequalities to get

$$\mathbb{P}^v_{u_{1:M}}\left((\Xi^{(k,j)}_{u_k,l_k})^c \mid \Gamma_{1:k}\cap\Lambda_{1:k}\right) \le 2\exp\left(-\frac{(\xi^{(j)}_k)^2}{2\sigma^2 h_k \tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)}}\right), \tag{64}$$

Substituting Eqs (62), (63), and (64) into the RHS of Eq (58), we have

$$\begin{aligned}
&\mathbb{P}^v_{u_{1:M}}(G\cap\Gamma_{1:k}\cap\Lambda_{1:k})\\
\le& K_1\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G\cap\Gamma_{1:k}\cap\Lambda_{1:k}) + \sum_{j=0}^{\zeta}\sum_{m=1}^{k-1}\mathbb{P}^v_{u_{1:M}}\left(\Gamma_{1:k}\cap\Lambda_{1:m}\cap(\Xi^{(m,j)}_{u_m,l_m})^c\right)\\
&+ 2\mathbb{P}^v_{u_{1:M}}(\Gamma_{1:k}\cap\Lambda_{1:k})\sum_{j=0}^{\zeta}\exp\left(-\frac{(\xi^{(j)}_k)^2}{2\sigma^2 h_k\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)}}\right).
\end{aligned}\tag{65}$$

Recall that the events $\Gamma^{(k)}_{u_k,l_k}$ are mutually disjoint for all possible values of $l_k \in 1:2^n$, and their union is the whole probability space. Using this, we can sum up both sides of Eq (65) over all possible values of l_k to eliminate $\Gamma^{(k)}_{u_k,l_k}$ and obtain

$$\begin{aligned}
&\mathbb{P}^v_{u_{1:M}}(G\cap\Gamma_{1:k-1}\cap\Lambda_{1:k})\\
\le& K_1\mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G\cap\Gamma_{1:k-1}\cap\Lambda_{1:k}) + \sum_{j=0}^{\zeta}\sum_{m=1}^{k-1}\mathbb{P}^v_{u_{1:M}}\left(\Gamma_{1:k-1}\cap\Lambda_{1:m}\cap(\Xi^{(m,j)}_{u_m,l_m})^c\right)\\
&+ 2\mathbb{P}^v_{u_{1:M}}(\Gamma_{1:k-1}\cap\Lambda_{1:k})\sum_{j=0}^{\zeta}\exp\left(-\frac{(\xi^{(j)}_k)^2}{2\sigma^2 h_k\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)}}\right).
\end{aligned}\tag{66}$$

In the second term of RHS in Eq (66), we can apply a similar concentration inequality to get

$$\begin{aligned}
&\sum_{j=0}^{\zeta}\mathbb{P}^v_{u_{1:M}}\left(\Gamma_{1:k-1}\cap\Lambda_{1:k-1}\cap(\Xi^{(k-1,j)}_{u_{k-1},l_{k-1}})^c\right)\\
\le& 2\mathbb{P}^v_{u_{1:M}}(\Gamma_{1:k-1}\cap\Lambda_{1:k-1})\sum_{j=0}^{\zeta}\exp\left(-\frac{(\xi^{(j)}_{k-1})^2}{2\sigma^2 h_{k-1}\tilde{C}_j^2\alpha^{2t-4}\delta_{t-1}^{2(\kappa-j)}}\right),
\end{aligned}$$

and substituting this bound to Eq (66) and summing over all possible $\Gamma^{(k-1)}_{u_{k-1},l_{k-1}}$ gives

$$\mathbb{P}^v_{u_{1:M}}(G\cap\Gamma_{1:k-2}\cap\Lambda_{1:k})$$

$$\leq K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G \cap \Gamma_{1:k-2} \cap \Lambda_{1:k}) + \sum_{j=0}^{\zeta} \sum_{m=1}^{k-2} \mathbb{P}^{v}_{u_{1:M}} \left(\Gamma_{1:k-2} \cap \Lambda_{1:m} \cap (\Xi^{(m,j)}_{u_m,l_m})^c \right)$$
$$+ 2 \sum_{m=k-1}^{k} \mathbb{P}^{v}_{u_{1:M}} (\Gamma_{1:k-2} \cap \Lambda_{1:m}) \sum_{j=0}^{\zeta} \exp\left(- \frac{(\xi^{(j)}_m)^2}{2\sigma^2 h_m \tilde{C}_j^2 \alpha^{2t-4} \delta_{t-1}^{2(\kappa-j)}} \right).$$

After repeating this process until we eliminate $\Gamma^{(1)}_{u_1,l_1}$, we get

$$\begin{aligned}
&\mathbb{P}^{v}_{u_{1:M}} (G \cap \Lambda_{1:k}) \\
\leq & K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G \cap \Lambda_{1:k}) + 2 \sum_{m=1}^{k} \mathbb{P}^{v}_{u_{1:M}} (\Lambda_{1:m}) \sum_{j=0}^{\zeta} \exp\left(- \frac{(\xi^{(j)}_m)^2}{2\sigma^2 h_m \tilde{C}_j^2 \alpha^{2t-4} \delta_{t-1}^{2(\kappa-j)}} \right) \\
\leq & K_1 \mathbb{P}^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(G \cap \Lambda_{1:k}) + 2 \sum_{m=1}^{k} \sum_{j=0}^{\zeta} \exp\left(- \frac{(\xi^{(j)}_m)^2}{2\sigma^2 h_m \tilde{C}_j^2 \alpha^{2t-4} \delta_{t-1}^{2(\kappa-j)}} \right) \qquad (67)
\end{aligned}$$

Now, we define

$$K_2 := 2 \sum_{m=1}^{k} \sum_{j=0}^{\zeta} \exp\left(- \frac{(\xi^{(j)}_m)^2}{2\sigma^2 h_m \tilde{C}_j^2 \alpha^{2t-4} \delta_{t-1}^{2(\kappa-j)}} \right).$$

Then we get the claim of the lemma.

Appendix E. Technical Proofs for Section C

E.1. Proof of Lemma 13

Originally we had requirements $0 < \eta < 1$, $\delta > 0$, $0 < \alpha < 1$, $0 < \beta < 1$, and $0 < \theta < 1$. Let the parameter values satisfy

$$\eta + \alpha + \alpha\eta < 1, \;\; \beta = \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}, \;\; \theta = \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta}.$$

First, we check that all the parameters are in the desired range. We start by noting that the inequalities $0 < \alpha < 1, 0 < \eta < 1, \eta + \alpha + \alpha\eta < 1$ are feasible. With these assumptions on α and η, it is easy to check $0 < \theta < 1$. Also,

$$\eta + \alpha + \alpha\eta < 1 \iff \eta < \frac{(1-\alpha)(1+\eta)}{2} \iff \eta^2 < \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha} = \beta,$$

so we can ensure $\beta > 0$. Also,

$$\beta = \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha} < \frac{(1-\alpha)(1+\eta)^2}{4} < 1.$$

Now consider the interpolation set,

$$\mathbf{Itp} := \{x \mid (1-\theta)r \leq \|x - c\|_2 \leq (1+\theta)r\},$$

where r and c are defined in Eqs (35) and (36), which we repeat below, and further evaluate with the assumption (38) on β:

$$c = \frac{1}{1-\alpha}x_1 - \frac{\alpha}{1-\alpha}x_2 = x_1 - \frac{\alpha}{1-\alpha}(x_2 - x_1),$$
$$r = \sqrt{\frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{\beta\delta^2}{1-\alpha}} = \sqrt{\frac{(1+\eta)^2\delta^2}{4} - \frac{\alpha}{(1-\alpha)^2}(\eta^2\delta^2 - \|x_1 - x_2\|_2^2)}.$$

We now show that, with the choice of θ in Eq (39),

$$\mathbf{Itp} \subset \mathrm{cl}(\mathbf{B}_{x_1}^2(\eta\delta)^c \cap \mathbf{B}_{x_1}^2(\delta)) \text{ for any } x_2 \in \mathbf{B}_{x_1}^2(\eta\delta),$$

which is the goal of this lemma, Eq (40). Consider $x_2 = x_1 + \rho\delta\mathbf{e}_1$, where $0 \le \rho \le \eta$. This choice of x_2 is a representative of all other x_2 that are $\rho\delta$ away from x_1. The center point c and radius r of intersection set, and the interpolation set $\mathbf{Itp}$ can be represented as functions of ρ:

$$c(\rho) = x_1 - \frac{\alpha\rho\delta}{1-\alpha}\mathbf{e}_1,$$
$$r(\rho) = \sqrt{\frac{(1+\eta)^2\delta^2}{4} - \frac{\alpha}{(1-\alpha)^2}(\eta^2 - \rho^2)\delta^2},$$
$$\mathbf{Itp}(\rho) = \{x \mid (1-\theta)r(\rho) \le \|x - c(\rho)\|_2 \le (1+\theta)r(\rho)\}.$$

Define $L_{\max}(\rho)$ and $L_{\min}(\rho)$ that are farthest and closest distance of points in $\mathbf{Itp}(\rho)$ to x_1, defined as follows:

$$L_{\max}(\rho) := \sup_{x\in\mathbf{Itp}(\rho)} \|x - x_1\|_2, \quad L_{\min}(\rho) := \inf_{x\in\mathbf{Itp}(\rho)} \|x - x_1\|_2.$$

The desired condition $\mathbf{Itp} \in \mathbf{B}_{x_1}^2(\eta\delta)^c \cap \mathbf{B}_{x_1}^2(\delta)$ can now be written as $L_{\max}(\rho) \le \delta$ and $L_{\min}(\rho) \ge \eta\delta$ for all $\rho \in [0, \eta]$.

Since x_1 and c are away by $\frac{\alpha\rho\delta}{1-\alpha}$, the farthest distance $L_{\max}(\rho)$ can be calculated as

$$L_{\max}(\rho) = (1+\theta)r(\rho) + \frac{\alpha\rho\delta}{1-\alpha} = (1+\theta)\sqrt{\frac{(1+\eta)^2\delta^2}{4} - \frac{\alpha}{(1-\alpha)^2}(\eta^2 - \rho^2)\delta^2} + \frac{\alpha\rho\delta}{1-\alpha}$$

and it is clearly an increasing function of ρ, so we only need to check that $L_{\max}(\eta) \le \delta$.

$$L_{\max}(\eta) = (1+\theta)\sqrt{\frac{(1+\eta)^2\delta^2}{4}} + \frac{\alpha\eta\delta}{1-\alpha} = \left(\frac{2(1-\alpha-\alpha\eta)}{1+\eta-\alpha-\alpha\eta}\right)\frac{(1+\eta)\delta}{2} + \frac{\alpha\eta\delta}{1-\alpha}$$
$$= \frac{(1-\alpha-\alpha\eta)\delta}{1-\alpha} + \frac{\alpha\eta\delta}{1-\alpha} = \delta.$$

For $L_{\min}(\rho)$, it requires a bit more thought. As long as $(1-\theta)r(\rho) > \frac{\alpha\rho\delta}{1-\alpha}$, we have $x_1 \notin \mathbf{Itp}(\rho)$ and

$$L_{\min}(\rho) = (1-\theta)r(\rho) - \frac{\alpha\rho\delta}{1-\alpha}.$$

Thus, by showing

$$(1-\theta)r(\rho) - \frac{\alpha\rho\delta}{1-\alpha} \ge \eta\delta > 0, \tag{68}$$

we can prove $(1-\theta)r(\rho) > \frac{\alpha\eta\delta}{1-\alpha}$, and $L_{\min}(\rho) \geq \eta\delta$. To do this, we show that the LHS of Eq (68) is a decreasing function of ρ, and then show Eq (68) for $\rho = \eta$. We can easily see that $-\theta r(\rho)$ is a decreasing function of ρ. For the rest of the LHS,

$$\begin{aligned}
&\frac{d}{d\rho}\left(r(\rho) - \frac{\alpha\rho\delta}{1-\alpha}\right) = \frac{\frac{2\alpha\delta^2\rho}{(1-\alpha)^2}}{2\sqrt{\frac{(1+\eta)^2\delta^2}{4} - \frac{\alpha}{(1-\alpha)^2}(\eta^2-\rho^2)\delta^2}} - \frac{\alpha\delta}{1-\alpha} \leq 0, \forall \rho \in [0,\eta] \\
\iff &\rho^2 \leq \frac{(1+\eta)^2(1-\alpha)^2}{4} - \alpha(\eta^2-\rho^2), \forall \rho \in [0,\eta] \\
\iff &(1-\alpha)\rho^2 \leq \frac{(1+\eta)^2(1-\alpha)^2}{4} - \alpha\eta^2, \forall \rho \in [0,\eta] \\
\iff &(1-\alpha)\eta^2 \leq \frac{(1+\eta)^2(1-\alpha)^2}{4} - \alpha\eta^2 \\
\iff &2\eta \leq (1+\eta)(1-\alpha) = 1+\eta-\alpha-\alpha\eta \\
\iff &\eta + \alpha + \alpha\eta \leq 1.
\end{aligned}$$

Since we know from Eq (37) that the last statement is true, we proved that the LHS of Eq (68) is decreasing. Finally, we examine Eq (68) for $\rho = \eta$.

$$(1-\theta)r(\eta) - \frac{\alpha\eta\delta}{1-\alpha} = \frac{2\eta}{(1+\eta)(1-\alpha)}\frac{(1+\eta)\delta}{2} - \frac{\alpha\eta\delta}{1-\alpha} = \eta\delta.$$

Thus far, we showed that with our choice of θ in Eq (39), the interpolation set $\mathbf{Itp}$ satisfies $\mathbf{Itp} \in \mathbf{B}^2_{x_1}(\eta\delta)^c \cap \mathbf{B}^2_{x_1}(\delta)$ for any $x_2 \in \mathbf{B}^2_{x_1}(\eta\delta)$.

E.2. Development of Algorithm 2

We define

$$\begin{aligned}
\dot{g}_-(w) &:= \nabla f_2(c + (1-\theta)rw) = \frac{2\alpha}{1-\alpha}(x_1 - x_2) + 2\alpha(1-\theta)rw, \\
\dot{g}_0(w) &:= \frac{\nabla f_2(c+rw) + \nabla f_1(c+rw)}{2} = \frac{2\alpha}{1-\alpha}(x_1 - x_2) + (1+\alpha)rw, \\
\dot{g}_+(w) &:= \nabla f_1(c + (1+\theta)rw) = \frac{2\alpha}{1-\alpha}(x_1 - x_2) + 2(1+\theta)rw
\end{aligned}$$

and then interpolate the gradients along each direction w:

$$\begin{aligned}
\dot{h}_-(\rho, w) &:= \dot{g}_-(w) + \frac{\dot{g}_0(w) - \dot{g}_-(w)}{\theta r}(\rho - (1-\theta)r) \\
&= \frac{2\alpha}{1-\alpha}(x_1 - x_2) - \frac{(1-\alpha)(1-\theta)r}{\theta}w + \left(\frac{1-\alpha}{\theta} + 2\alpha\right)\rho w \quad \text{for } \rho \in [(1-\theta)r, r],
\end{aligned} \tag{69}$$

$$\begin{aligned}
\dot{h}_+(\rho, w) &:= \dot{g}_+(w) - \frac{\dot{g}_0(w) - \dot{g}_+(w)}{\theta r}(\rho - (1+\theta)r) \\
&= \frac{2\alpha}{1-\alpha}(x_1 - x_2) - \frac{(1-\alpha)(1+\theta)r}{\theta}w + \left(\frac{1-\alpha}{\theta} + 2\right)\rho w \quad \text{for } \rho \in [r, (1+\theta)r].
\end{aligned} \tag{70}$$

We can see that $\dot{h}_-(r,w) = \dot{h}_+(r,w) = \dot{g}_0(w)$.

Interpolation of gradients also changes the function values on $\mathbf{Itp}$. The original function values at $x = c + (1-\theta)rw$ and $x = c + (1+\theta)rw$, which are the points we start the interpolation from, are

$$f_2(c+(1-\theta)rw) = \frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{2\alpha(1-\theta)r}{1-\alpha}\langle x_1 - x_2, w\rangle + \alpha(1-\theta)^2r^2 + \beta\delta^2,$$
$$f_1(c+(1+\theta)rw) = \frac{\alpha^2}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{2\alpha(1+\theta)r}{1-\alpha}\langle x_1 - x_2, w\rangle + (1+\theta)^2r^2.$$

The function values after interpolation is calculated by integrating the directional derivatives,

$$h_-(\rho, w) := f_2(c+(1-\theta)rw) + \int_{(1-\theta)r}^{\rho} \langle \dot{h}_-(t,w), w\rangle dt \qquad \text{for } \rho \in [(1-\theta)r, r], \tag{71}$$
$$h_+(\rho, w) := f_1(c+(1+\theta)rw) - \int_{\rho}^{(1+\theta)r} \langle \dot{h}_+(t,w), w\rangle dt \qquad \text{for } \rho \in [r, (1+\theta)r],$$

and the integrals are evaluated as

$$\begin{aligned}\int_{(1-\theta)r}^{\rho} \langle \dot{h}_-(t,w), w\rangle dt &= \left(\frac{2\alpha}{1-\alpha}\langle x_1 - x_2, w\rangle - \frac{(1-\alpha)(1-\theta)r}{\theta}\right)(\rho - (1-\theta)r) \\ &\quad + \frac{1}{2}\left(\frac{1-\alpha}{\theta} + 2\alpha\right)(\rho^2 - (1-\theta)^2r^2), \\ \int_{\rho}^{(1+\theta)r} \langle \dot{h}_+(t,w), w\rangle dt &= \left(\frac{2\alpha}{1-\alpha}\langle x_1 - x_2, w\rangle - \frac{(1-\alpha)(1+\theta)r}{\theta}\right)((1+\theta)r - \rho) \\ &\quad + \frac{1}{2}\left(\frac{1-\alpha}{\theta} + 2\right)((1+\theta)^2r^2 - \rho^2).\end{aligned}$$

Substituting the integrals and arranging the terms, we get the following:

$$\begin{aligned}h_-(\rho, w) :=& \frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \beta\delta^2 + \frac{(1-\alpha)(1-\theta)^2r^2}{2\theta} \\ &+ \left(\frac{2\alpha}{1-\alpha}\langle x_1 - x_2, w\rangle - \frac{(1-\alpha)(1-\theta)r}{\theta}\right)\rho + \left(\frac{1-\alpha}{2\theta} + \alpha\right)\rho^2 \quad \text{for } \rho \in [(1-\theta)r, r], \\ h_+(\rho, w) :=& \frac{\alpha^2}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{(1-\alpha)(1+\theta)^2r^2}{2\theta} \\ &+ \left(\frac{2\alpha}{1-\alpha}\langle x_1 - x_2, w\rangle - \frac{(1-\alpha)(1+\theta)r}{\theta}\right)\rho + \left(\frac{1-\alpha}{2\theta} + 1\right)\rho^2 \quad \text{for } \rho \in [r, (1+\theta)r].\end{aligned}$$

We can double-check that $h_-(r,w) = h_+(r,w)$ by substituting $\rho = r$ to both functions, arranging terms, and noting that

$$\frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \alpha r^2 + \beta\delta^2 = \frac{\alpha^2}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + r^2, \tag{72}$$

by definition of β in Eq (38).

Using $\dot{h}_-$, $\dot{h}_+$, h_-, and h_+ defined as above, we can define infinite number of hyperplanes corresponding to each point $c + \rho w$ in $\mathbf{Itp}$,

$$f_-^{\rho,w}(x) := \langle \dot{h}_-(\rho, w), x - (c + \rho w)\rangle + h_-(\rho, w) \qquad \text{for } \rho \in [(1-\theta)r, r], \|w\|_2 = 1,$$
$$f_+^{\rho,w}(x) := \langle \dot{h}_+(\rho, w), x - (c + \rho w)\rangle + h_+(\rho, w) \qquad \text{for } \rho \in [r, (1+\theta)r], \|w\|_2 = 1.$$

After substituting $\dot{h}_-(\rho, w)$, $\dot{h}_+(\rho, w)$, $h_-(\rho, w)$, and $h_+(\rho, w)$, and then arranging terms, we get

$$\begin{aligned} f_-^{\rho,w}(x) =& \frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \beta\delta^2 + \frac{(1-\alpha)(1-\theta)^2 r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta} + \alpha\right)\rho^2 \\ &+ \left\langle \frac{2\alpha}{1-\alpha}(x_1 - x_2) - \frac{(1-\alpha)(1-\theta)r}{\theta} w + \left(\frac{1-\alpha}{\theta} + 2\alpha\right)\rho w, x - c \right\rangle \qquad (73) \\ f_+^{\rho,w}(x) =& \frac{\alpha^2}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{(1-\alpha)(1+\theta)^2 r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta} + 1\right)\rho^2 \\ &+ \left\langle \frac{2\alpha}{1-\alpha}(x_1 - x_2) - \frac{(1-\alpha)(1+\theta)r}{\theta} w + \left(\frac{1-\alpha}{\theta} + 2\right)\rho w, x - c \right\rangle. \end{aligned}$$

We can finally state the definition of the interpolated function, which is

$$f(x) = \max\left\{ f_1(x), f_2(x), \sup_{\rho\in[(1-t)r,r],\|w\|_2=1} f_-^{\rho,w}(x), \sup_{\rho\in[r,(1+t)r],\|w\|_2=1} f_+^{\rho,w}(x) \right\}. \qquad (74)$$

E.3. Proof of Lemma 14

Before the proof of Lemma 14, we state and prove a "helper" lemma.

Lemma 23 *The following holds:*

1. *For any $0 < \rho \le \rho_1 \le \rho_2$ where $\rho_1, \rho_2 \in [(1-\theta)r, r]$, and any unit vectors w, w', we have $f_-^{\rho_1,w}(c + \rho w) \ge f_-^{\rho_2,w'}(c + \rho w)$, where equality holds if and only if $\rho_1 = \rho_2$ and $w = w'$.*
2. *For any $\rho \ge \rho_1 \ge \rho_2$ where $\rho_1, \rho_2 \in [(1-\theta)r, r]$, and any unit vectors w, w', we have $f_-^{\rho_1,w}(c + \rho w) \ge f_-^{\rho_2,w'}(c + \rho w)$, where equality holds if and only if $\rho_1 = \rho_2$ and $w = w'$.*
3. *For any $0 < \rho \le \rho_1 \le \rho_2$ where $\rho_1, \rho_2 \in [r, (1+\theta)r]$, and any unit vectors w, w', we have $f_+^{\rho_1,w}(c + \rho w) \ge f_+^{\rho_2,w'}(c + \rho w)$, where equality holds if and only if $\rho_1 = \rho_2$ and $w = w'$.*
4. *For any $\rho \ge \rho_1 \ge \rho_2$ where $\rho_1, \rho_2 \in [r, (1+\theta)r]$, and any unit vectors w, w', we have $f_+^{\rho_1,w}(c + \rho w) \ge f_+^{\rho_2,w'}(c + \rho w)$, where equality holds if and only if $\rho_1 = \rho_2$ and $w = w'$.*
5. *For any $\rho \in [(1-\theta)r, r]$, $h_-(\rho, w) \ge f_2(c + \rho w)$, where equality holds if and only if $\rho = (1-\theta)r$.*
6. *For any $\rho \in [r, (1+\theta)r]$, $h_+(\rho, w) \ge f_1(c + \rho w)$, where equality holds if and only if $\rho = (1+\theta)r$.*

Proof For Part 1, we first show that $f_-^{\rho_1,w}(c+\rho w) \geq f_-^{\rho_2,w}(c+\rho w)$, and then show $f_-^{\rho_2,w}(c+\rho w) \geq f_-^{\rho_2,w'}(c+\rho w)$. From Eq (73), we can see that

$$\begin{aligned}
&f_-^{\rho_1,w}(c+\rho w) \geq f_-^{\rho_2,w}(c+\rho w) \\
\iff &-\left(\frac{1-\alpha}{2\theta}+\alpha\right)\rho_1^2+\left(\frac{1-\alpha}{\theta}+2\alpha\right)\rho_1\rho \geq -\left(\frac{1-\alpha}{2\theta}+\alpha\right)\rho_2^2+\left(\frac{1-\alpha}{\theta}+2\alpha\right)\rho_2\rho \\
\iff &(\rho_2-\rho_1)(\rho_2+\rho_1) \geq 2(\rho_2-\rho_1)\rho,
\end{aligned}$$

which is true because $\rho \leq \rho_1 \leq \rho_2$. Also, we can check that equality holds if and only if $\rho_1 = \rho_2$. For the next step,

$$\begin{aligned}
&f_-^{\rho_2,w}(c+\rho w) \geq f_-^{\rho_2,w'}(c+\rho w) \\
\iff &\langle -\frac{(1-\alpha)(1-\theta)r}{\theta}w+\left(\frac{1-\alpha}{\theta}+2\alpha\right)\rho_2 w, \rho w\rangle \geq \langle -\frac{(1-\alpha)(1-\theta)r}{\theta}w'+\left(\frac{1-\alpha}{\theta}+2\alpha\right)\rho_2 w', \rho w\rangle \\
\iff &\left(\frac{1-\alpha}{\theta}(\rho_2-(1-\theta)r)+2\alpha\rho_2\right)\rho(1-\langle w', w\rangle) \geq 0,
\end{aligned}$$

which is true and equality holds if and only if $w = w'$. Parts 2–4 can be proved in a very similar way.

Part 5 holds because of the definition of $h_-(\rho, w)$ in Eq (71) and we can check that

$$\langle \dot{h}_-(\rho, w), w\rangle \geq \langle \nabla f_2(c+\rho w), w\rangle \text{ for } \rho \in [(1-\theta)r, r],$$

where equality holds if and only if $\rho = (1-\theta)r$. Part 6 can be proved similarly. ■

Given the helper lemma, we prove Lemma 14 by partitioning ρ into 8 intervals and prove each case separately. Specifically, for each case we show that for any given ρ and w, the supremum $f(c+\rho w)$ is achieved by exactly one or two functions among all the functions. If there are two functions that achieve the supremum, we show that they have the same gradients. If this is true, the statement about $\nabla f(c+\rho w)$ will naturally follow.

Case 1: $\rho = 0$. When $\rho = 0$, f_2 is strictly bigger than all other functions. We can show this by directly comparing function values at $x = c$. For this case, recall from Eq (72) and definition of f_1 and f_2 that

$$f_1(c) = \frac{\alpha^2}{(1-\alpha)^2}\|x_1-x_2\|_2^2 = \frac{\alpha}{(1-\alpha)^2}\|x_1-x_2\|_2^2+\beta\delta^2-(1-\alpha)r^2 = f_2(c)-(1-\alpha)r^2. \quad (75)$$

There are three things that we need to check:

1. $f_2(c) > f_-^{\rho',w'}(c), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$
   $\iff 0 > \frac{(1-\alpha)(1-\theta)^2r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta}+\alpha\right)\rho'^2, \forall \rho' \in [(1-\theta)r, r]$
   $\iff 0 > \frac{(1-\alpha)(1-\theta)^2r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta}+\alpha\right)(1-\theta)^2r^2 = -\alpha(1-\theta)^2r^2.$
2. $f_2(c) > f_+^{\rho',w'}(c), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$. Note that

$$f_+^{\rho',w'}(c) = \frac{\alpha^2}{(1-\alpha)^2}\|x_1-x_2\|_2^2 + \frac{(1-\alpha)(1+\theta)^2r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta}+1\right)\rho'^2$$

$$= f_2(c) - (1-\alpha)r^2 + \frac{(1-\alpha)(1+\theta)^2 r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta} + 1\right)\rho'^2,$$

so we need to show

$$-(1-\alpha)r^2 + \frac{(1-\alpha)(1+\theta)^2 r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta} + 1\right)\rho'^2 < 0, \forall \rho' \in [r, (1+\theta)r],$$
$$\iff -(1-\alpha)r^2 + \frac{(1-\alpha)(1+\theta)^2 r^2}{2\theta} - \left(\frac{1-\alpha}{2\theta} + 1\right) r^2 < 0 \iff (1-\alpha)\theta < 2,$$

which is true because $\alpha, \theta \in (0,1)$.

3. $f_2(c) > f_1(c)$. This is already shown by Eq (75).

Case 2: $\rho \in (0, (1-\theta)r)$**.** In this region,

$$f_2(c+\rho w) = \frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \beta\delta^2 + \frac{2\alpha\rho}{1-\alpha}\langle x_1 - x_2, w\rangle + \alpha\rho^2$$

dominates all other functions.

1. $f_2(c+\rho w) > f_-^{\rho', w'}(c+\rho w), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$. To show this, it suffices to show $f_2(c+\rho w) > f_-^{(1-\theta)r, w}(c+\rho w)$, and the rest follows because of Lemma 23.1. We can calculate and arrange $f_-^{(1-\theta)r,w}(c+\rho w)$ to get

   $$f_-^{(1-\theta)r,w}(c+\rho w) = \frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \beta\delta^2 + \frac{2\alpha\rho}{1-\alpha}\langle x_1 - x_2, w\rangle - \alpha(1-\theta)^2 r^2 + 2\alpha(1-\theta)r\rho,$$

   so

   $$f_2(c+\rho w) > f_-^{(1-\theta)r,w}(c+\rho w) \iff \alpha(1-\theta)^2 r^2 - 2\alpha(1-\theta)\rho + \alpha\rho^2 > 0 \iff \alpha((1-\theta)r - \rho)^2 > 0,$$

   which is true.

2. $f_2(c+\rho w) > f_+^{\rho', w'}(c+\rho w), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$. Notice that we just showed that $f_2(c+\rho w) > f_-^{r,w}(c+\rho w) = f_+^{r,w}(c+\rho w)$. The rest follows by Lemma 23.3.

3. $f_2(c+\rho w) > f_1(c+\rho w)$. This can be shown by direct comparison.

Case 3: $\rho = (1-\theta)r$**.** In this case, $f_2(c+(1-\theta)rw) = f_-^{(1-\theta)r,w}(c+(1-\theta)rw) = h_-((1-\theta)r, w)$ is strictly greater than all other functions. The gradient still exists because the gradients of these two functions are the same: $\nabla f_2(c+(1-\theta)rw) = \nabla f_-^{(1-\theta)r,w}(c+(1-\theta)rw) = \dot{h}_-((1-\theta)r, w)$, as we can see from Eq (69). We need to show:

1. $f_2(c+(1-\theta)rw) > f_-^{\rho', w'}(c+(1-\theta)rw), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$, as long as $(\rho', w') \neq ((1-\theta)r, w)$. This is true because of Lemma 23.1.

2. $f_2(c+(1-\theta)rw) > f_+^{\rho', w'}(c+(1-\theta)rw), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$. We just showed that $f_2(c+(1-\theta)rw) > f_-^{r,w}(c+(1-\theta)rw) = f_+^{r,w}(c+(1-\theta)rw)$, so the rest follows by Lemma 23.3.

3. $f_2(c+(1-\theta)rw) > f_1(c+(1-\theta)rw)$. This can be shown by direct comparison.

Case 4: $\rho \in ((1-\theta)r, r)$**.** In this region, the function $f_-^{\rho,w}$ dominates all other functions. Note that $f_-^{\rho,w}(c+\rho w) = h_-(\rho, w)$.

1. $f_-^{\rho,w}(c+\rho w) > f_2(c+\rho w)$. This is true by Lemma 23.5.
2. $f_-^{\rho,w}(c+\rho w) > f_-^{\rho',w'}(c+\rho w), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$, as long as $(\rho', w') \neq (\rho, w)$. This is true because of Lemma 23.1 and 23.2.
3. $f_-^{\rho,w}(c+\rho w) > f_+^{\rho',w'}(c+\rho w), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$. We just showed that $f_-^{\rho,w}(c+\rho w) > f_-^{r,w}(c+\rho w) = f_+^{r,w}(c+\rho w)$, so the rest follows by Lemma 23.3.
4. $f_-^{\rho,w}(c+\rho w) > f_1(c+(1-\theta)rw)$. This can be shown by noting that $f_2(c+\rho w) > f_1(c+\rho w)$ by direct comparison, and then Lemma 23.5.

Case 5: $\rho = r$**.** In this case, $f_-^{r,w}(c+rw) = h_-(r,w) = f_+^{r,w}(c+rw) = h_+(r,w)$ is greater than all other functions, but their gradients are also the same.

1. $f_-^{r,w}(c+rw) > f_2(c+rw)$. This is true by Lemma 23.5.
2. $f_-^{r,w}(c+rw) > f_-^{\rho',w'}(c+rw), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$, as long as $(\rho', w') \neq (r, w)$. This is because of Lemma 23.2.
3. $f_+^{r,w}(c+rw) > f_+^{\rho',w'}(c+rw), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$, as long as $(\rho', w') \neq (r, w)$, which is implied by Lemma 23.3.
4. $f_+^{r,w}(c+rw) > f_1(c+rw)$, by Lemma 23.6.

Case 6: $\rho \in (r, (1+\theta)r)$**.** We have to show that $f_+^{\rho,w}(c+\rho w) = h_+(\rho, w)$ is greater than values from all other functions.

1. $f_+^{\rho,w}(c+\rho w) > f_1(c+\rho w)$. This is true by Lemma 23.6.
2. $f_+^{\rho,w}(c+\rho w) > f_+^{\rho',w'}(c+\rho w), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$, as long as $(\rho', w') \neq (\rho, w)$. This is true because of Lemma 23.3 and 23.4.
3. $f_+^{\rho,w}(c+\rho w) > f_-^{\rho',w'}(c+\rho w), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$. We just showed that $f_+^{\rho,w}(c+\rho w) > f_+^{r,w}(c+\rho w) = f_-^{r,w}(c+\rho w)$, so the rest follows by Lemma 23.2.
4. $f_+^{\rho,w}(c+\rho w) > f_2(c+(1-\theta)rw)$. This can be shown by noting that $f_1(c+\rho w) > f_2(c+\rho w)$ by direct comparison, and then Lemma 23.6.

Case 7: $\rho = (1+\theta)r$**.** Here, $f_1(c+(1+\theta)rw) = f_+^{(1+\theta)r,w}(c+(1+\theta)rw) = h_+((1+\theta)r, w)$ dominate, but the gradients at the point are the same.

1. $f_1(c+(1+\theta)rw) > f_+^{\rho',w'}(c+(1+\theta)rw), \forall \rho' \in [r, (1+\theta)r], \|w'\|_2 = 1$, as long as $(\rho', w') \neq ((1+\theta)r, w)$. This is true because of Lemma 23.4.
2. $f_1(c+(1+\theta)rw) > f_-^{\rho',w'}(c+(1+\theta)rw), \forall \rho' \in [(1-\theta)r, r], \|w'\|_2 = 1$. We just showed that $f_1(c+(1+\theta)rw) > f_+^{r,w}(c+(1+\theta)rw) = f_-^{r,w}(c+(1+\theta)rw)$, so the rest follows by Lemma 23.2.
3. $f_1(c+(1+\theta)rw) > f_2(c+(1+\theta)rw)$. This can be shown by direct comparison.

Case 8: $\rho \in ((1+\theta)r, \infty)$**.** In the final case,

$$f_1(c+\rho w) = \frac{\alpha^2}{(1-\alpha)^2}\left\|x_1 - x_2\right\|_2^2 + \frac{2\alpha\rho}{1-\alpha}\langle x_1 - x_2, w\rangle + \rho^2.$$

dominates all other functions.

1. $f_1(c+\rho w) > f_+^{\rho', w'}(c+\rho w), \forall \rho' \in [r, (1+\theta)r], \left\|w'\right\|_2 = 1$. To show this, it suffices to show $f_1(c+\rho w) > f_+^{(1+\theta)r, w}(c+\rho w)$, and the rest follows because of Lemma 23.4. We can calculate and arrange $f_+^{(1+\theta)r,w}(c+\rho w)$ to get

   $$f_+^{(1+\theta)r,w}(c+\rho w) = \frac{\alpha^2}{(1-\alpha)^2}\left\|x_1 - x_2\right\|_2^2 + \frac{2\alpha\rho}{1-\alpha}\langle x_1 - x_2, w\rangle - (1+\theta)^2 r^2 + 2(1+\theta)r\rho,$$

   so

   $$f_1(c+\rho w) > f_+^{(1+\theta)r,w}(c+\rho w) \iff (1+\theta)^2 r^2 - 2(1+\theta)\rho + \rho^2 > 0 \iff ((1+\theta)r - \rho)^2 > 0.$$

2. $f_1(c+\rho w) > f_-^{\rho', w'}(c+\rho w), \forall \rho' \in [(1-\theta)r, r], \left\|w'\right\|_2 = 1$. Notice that we just showed that $f_1(c+\rho w) > f_+^{r,w}(c+\rho w) = f_-^{r,w}(c+\rho w)$. The rest follows by Lemma 23.2.

3. $f_1(c+\rho w) > f_2(c+\rho w)$. This can be shown by direct comparison.

E.4. Proof of Lemma 15

From Lemma 14, it became clear that

$$\nabla f(c+\rho w) = \begin{cases} \nabla f_2(c+\rho w) & \text{if } \rho \in [0, (1-\theta)r] \\ \dot{h}_-(\rho, w) & \text{if } \rho \in [(1-\theta)r, r] \\ \dot{h}_+(\rho, w) & \text{if } \rho \in [r, (1+\theta)r] \\ \nabla f_1(c+\rho w) & \text{if } \rho \in [(1+\theta)r, \infty). \end{cases}$$

In order to show that the function is $\left(2 + \frac{1-\alpha}{\theta}\right)$-smooth, it suffices to show that each piece of the function has $\left(2 + \frac{1-\alpha}{\theta}\right)$-Lipschitz gradient. Since f_1 and f_2 are quadratic functions, it is easy to see that they are 2-smooth and 2α-smooth, respectively. In case of $\dot{h}_-(\rho, w)$, for $\rho_1, \rho_2 \in [(1-\theta)r, r]$ and any arbitrary unit vectors w_1 and w_2,

$$\begin{aligned}
&\left\|\dot{h}_-(\rho_1, w_1) - \dot{h}_-(\rho_2, w_2)\right\|_2 \\
=&\left\|-\frac{(1-\alpha)(1-\theta)r}{\theta}(w_1 - w_2) + \left(\frac{1-\alpha}{\theta} + 2\alpha\right)(\rho_1 w_1 - \rho_2 w_2)\right\|_2 \\
=&\left\|\frac{1-\alpha}{\theta}[(\rho_1 - (1-\theta)r)w_1 - (\rho_2 - (1-\theta)r)w_2] + 2\alpha(\rho_1 w_1 - \rho_2 w_2)\right\|_2 \\
\leq&\frac{(1-\alpha)}{\theta}\left\|(\rho_1 - (1-\theta)r)w_1 - (\rho_2 - (1-\theta)r)w_2\right\|_2 + 2\alpha\left\|\rho_1 w_1 - \rho_2 w_2\right\|_2 \\
\leq&\left(2\alpha + \frac{1-\alpha}{\theta}\right)\left\|\rho_1 w_1 - \rho_2 w_2\right\|_2.
\end{aligned}$$

The last inequality sign used that $\|(\rho_1-(1-\theta)r)w_1-(\rho_2-(1-\theta)r)w_2\|_2 \le \|\rho_1 w_1-\rho_2 w_2\|_2$. To see this, whenever $0\le z_3\le z_1$ and $0\le z_3\le z_2$,

$$\begin{aligned}
&\|(z_1-z_3)w_1-(z_2-z_3)w_2\|_2 \le \|z_1w_1-z_2w_2\|_2\\
\iff &(z_1-z_3)^2+(z_2-z_3)^2-2(z_1-z_3)(z_2-z_3)\langle w_1,w_2\rangle \le z_1^2+z_2^2-2z_1z_2\langle w_1,w_2\rangle\\
\iff &((z_1-z_3)-(z_2-z_3))^2+2(z_1-z_3)(z_2-z_3)(1-\langle w_1,w_2\rangle)\le (z_1-z_2)^2+2z_1z_2(1-\langle w_1,w_2\rangle).
\end{aligned}$$

Similarly, for $\rho_1,\rho_2\in[r,(1+\theta)r]$ and any arbitrary unit vectors w_1 and w_2,

$$\begin{aligned}
&\|h_+(\rho_1,w_1)-h_+(\rho_2,w_2)\|_2\\
=&\left\|-\frac{(1-\alpha)(1+\theta)r}{\theta}(w_1-w_2)+\left(\frac{1-\alpha}{\theta}+2\right)(\rho_1w_1-\rho_2w_2)\right\|_2\\
=&\left\|\frac{1-\alpha}{\theta}[((1+\theta)r-\rho_2)w_2-((1+\theta)r-\rho_1)w_1]+2(\rho_1w_1-\rho_2w_2)\right\|_2\\
\le&\frac{1-\alpha}{\theta}\|((1+\theta)r-\rho_2)w_2-((1+\theta)r-\rho_1)w_1\|_2+2\|\rho_1w_1-\rho_2w_2\|_2\\
\le&\left(2+\frac{1-\alpha}{\theta}\right)\|\rho_1w_1-\rho_2w_2\|_2.
\end{aligned}$$

The last inequality sign used that $\|((1+\theta)r-\rho_2)w_2-((1+\theta)r-\rho_1)w_1\|_2\le\|\rho_1w_1-\rho_2w_2\|_2$. To see this, note that whenever $\frac{z_3}{2}\le z_1\le z_3$ and $\frac{z_3}{2}\le z_2\le z_3$,

$$\begin{aligned}
&\|(z_3-z_2)w_2-(z_3-z_1)w_1\|_2\le\|z_1w_1-z_2w_2\|_2\\
\iff&((z_3-z_2)-(z_3-z_1))^2+2(z_3-z_2)(z_3-z_1)(1-\langle w_1,w_2\rangle)\le(z_1-z_2)^2+2z_1z_2(1-\langle w_1,w_2\rangle)\\
\iff&(z_3-z_2)(z_3-z_1)\le z_2z_1.
\end{aligned}$$

The last statement holds because $0\le z_3-z_1\le z_1$ and $0\le z_3-z_2\le z_2$. Recalling $\theta\in(0,1)$, ρ_1, ρ_2, and $(1+\theta)r$ corresponds to z_1, z_2, and z_3, respectively. This concludes that the whole function has $\left(2+\frac{1-\alpha}{\theta}\right)$-Lipschitz gradient.

Now, since we know that $f_2(x)$ is 2α-strongly convex, the proof of 2α-strong convexity can be done by showing that $\tilde f(x):=f(x)-f_2(x)$ is convex, or equivalently, that $\nabla\tilde f(x)$ is monotone. The value of $\nabla\tilde f(c+\rho w)$ depending on different values of ρ is as the following:

$$\nabla\tilde f(c+\rho w)=\begin{cases}0 & \text{if } 0\le\rho\le(1-\theta)r\\ \frac{(1-\alpha)(\rho-(1-\theta)r)}{\theta}w & \text{if } (1-\theta)r\le\rho\le r\\ \frac{(1-\alpha)((1+2\theta)\rho-(1+\theta)r)}{\theta}w & \text{if } r\le\rho\le(1+\theta)r\\ 2(1-\alpha)\rho w & \text{if } \rho\ge(1+\theta)r.\end{cases}$$

Now, we need to show that $\nabla\tilde f(c+\rho w)$ is monotone, meaning that

$$\langle\nabla\tilde f(c+\rho'w')-\nabla\tilde f(c+\rho w),\rho'w'-\rho w\rangle\ge0,\ \forall\rho\ge0,\rho'\ge0,\|w\|_2=\|w'\|_2=1.$$

For notational simplicity in the proof, define $(\star)=\langle\nabla\tilde f(c+\rho'w')-\nabla\tilde f(c+\rho w),\rho'w'-\rho w\rangle$.

1. If $\rho,\rho'\in[0,(1-\theta)r]$, $(\star)=0$.

2. If $\rho \in [0, (1-\theta)r], \rho' \in ((1-\theta)r, r]$,
$$(\star) = \frac{(1-\alpha)(\rho' - (1-\theta)r)}{\theta}(\rho' - \rho\langle w, w'\rangle) \geq 0.$$
3. If $\rho \in [0, (1-\theta)r], \rho' \in (r, (1+\theta)r]$, the proof is similar to Case 2.
4. If $\rho \in [0, (1-\theta)r], \rho' \in ((1+\theta)r, \infty)$, the proof is similar to Case 2.
5. If $\rho, \rho' \in ((1-\theta)r, r]$,
$$\begin{aligned}(\star) &= \frac{(1-\alpha)}{\theta}\left((\rho' - (1-\theta)r)(\rho' - \rho\langle w, w'\rangle) + (\rho - (1-\theta)r)(\rho - \rho'\langle w, w'\rangle)\right) \\ &\geq \frac{(1-\alpha)}{\theta}\left((\rho' - (1-\theta)r)(\rho' - \rho) + (\rho - (1-\theta)r)(\rho - \rho')\right) \\ &= \frac{(1-\alpha)(\rho'-\rho)^2}{\theta} \geq 0.\end{aligned}$$
6. if $\rho \in ((1-\theta)r, r], \rho' \in (r, (1+\theta)r]$,
$$\begin{aligned}(\star) &\geq \frac{(1-\alpha)(\rho'-\rho)}{\theta}\left(((1+2\theta)\rho' - (1+\theta)r) - (\rho - (1-\theta)r)\right) \\ &= \frac{(1-\alpha)(\rho'-\rho)}{\theta}\left(2\theta(\rho' - r) + (\rho' - \rho)\right) \geq 0.\end{aligned}$$
7. if $\rho \in ((1-\theta)r, r], \rho' \in ((1+\theta)r, \infty)$,
$$\begin{aligned}(\star) &\geq \frac{(1-\alpha)(\rho'-\rho)}{\theta}\left(2\theta\rho' - (\rho - (1-\theta)r))\right) \\ &= \frac{(1-\alpha)(\rho'-\rho)}{\theta}\left(\theta(2\rho' - r) + r - \rho\right) \geq 0.\end{aligned}$$
8. If $\rho, \rho' \in (r, (1+\theta)r]$, the proof is similar to Case 5.
9. If $\rho \in (r, (1+\theta)r], \rho' \in ((1+\theta)r, \infty)$,
$$\begin{aligned}(\star) &\geq \frac{(1-\alpha)(\rho'-\rho)}{\theta}\left(2\theta\rho' - ((1+2\theta)\rho - (1+\theta)r))\right) \\ &= \frac{(1-\alpha)(\rho'-\rho)}{\theta}\left(2\theta(\rho' - \rho) + (1+\theta)r - \rho\right) \geq 0.\end{aligned}$$
10. If $\rho, \rho' \in ((1+\theta)r, \infty)$, the proof is similar to Case 5.

E.5. Proof of Lemma 16

By Lemma 13,
$$\mathbf{Itp} = \{c + \rho w \mid (1-\theta)r \leq \rho \leq (1+\theta)r, \|w\|_2 = 1\} \subset \mathrm{cl}(\mathbf{B}^2_{x_1}(\eta\delta)^c \cap \mathbf{B}^2_{x_1}(\delta)).$$
So, Lemma 16.1 and 16.2 are implied by Lemma 14.

For Lemma 16.3, By observing $f_2(x_2) = s\beta\delta^2 + t \le f_2(x) \le f(x)$ for all x, it is easy to see that the global minimum value of $f(x)$ is $s\beta\delta^2 + t$ and is attained at $x_2 \in \mathbf{B}^2_{x_1}(\eta\delta)$. Also, at any x such that $\|x - x_1\|_2 = \delta$, we can check $f(x) = f_1(x) = s\delta^2 + t$. By convexity, any point in between x and x_2 cannot be larger than $s\delta^2 + t$, which means that $s\beta\delta^2 + t \le f(x) \le s\delta^2 + t$ for all $x \in \mathbf{B}^2_{x_1}(\delta)$.

For the last statement Lemma 16.4, we will assume that the scaling factor $s = 1$ and prove that $\|\nabla f(x)\|_2 \le 2\delta$. The norm of gradient $\|\nabla f(x)\|_2$ naturally scales with s, so Lemma 16.4 follows for any $s > 0$. Note that $\|\nabla f_1(x)\|_2 = \|2(x - x_1)\|_2 \le 2\delta$ for any $x \in \mathbf{B}^2_{x_1}(\delta)$. Also, for $x \in \mathbf{B}^2_{x_1}(\delta)$,

$$\|\nabla f_2(x)\|_2 = 2\alpha \|x - x_2\|_2 = 2\alpha \|x - x_1\|_2 + 2\alpha \|x_1 - x_2\|_2 \le 2\alpha\delta + 2\alpha\eta\delta \le 2\delta,$$

where the last inequality is by Eq (37): $\eta + \alpha + \alpha\eta < 1$, which implies $\alpha + \alpha\eta < 1$. For $x = c + \rho w$ where $\rho \in [(1-\theta)r, r]$ and w is any unit vector, $\nabla f(x) = \dot{h}_-(x)$.

$$\begin{aligned}\left\|\dot{h}_-(c + \rho w)\right\|_2 &= \left\|\frac{2\alpha}{1-\alpha}(x_1 - x_2) + \left[\frac{(1-\alpha)(\rho - (1-\theta)r)}{\theta} + 2\alpha\rho\right] w\right\|_2 \\ \le &\frac{2\alpha}{1-\alpha}\|x_1 - x_2\|_2 + \left[\frac{(1-\alpha)(\rho - (1-\theta)r)}{\theta} + 2\alpha\rho\right] \le \frac{2\alpha\eta\delta}{1-\alpha} + (1+\alpha)r,\end{aligned}$$

where the last inequality is obtained by substituting $\rho = r$, the maximum possible ρ in the range.

$$\begin{aligned}\left\|\dot{h}_+(c + \rho w)\right\|_2 &= \left\|\frac{2\alpha}{1-\alpha}(x_1 - x_2) + \left[\frac{(1-\alpha)(\rho - (1+\theta)r)}{\theta} + 2\rho\right] w\right\|_2 \\ \le &\frac{2\alpha}{1-\alpha}\|x_1 - x_2\|_2 + \left[\frac{(1-\alpha)(\rho - (1+\theta)r)}{\theta} + 2\rho\right] \le \frac{2\alpha\eta\delta}{1-\alpha} + 2(1+\theta)r,\end{aligned}$$

also where the last inequality is obtained by substituting $\rho = (1+\theta)r$. We can check that $1 + \alpha < 2 < 2(1+\theta)$, so it suffices to show

$$\frac{2\alpha\eta\delta}{1-\alpha} + 2(1+\theta)r \le 2\delta. \tag{76}$$

First, recall from Eqs (36) and (38) that

$$r = \sqrt{\frac{\alpha}{(1-\alpha)^2}\|x_1 - x_2\|_2^2 + \frac{\beta\delta^2}{1-\alpha}}, \quad \text{and} \quad \beta = \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha}.$$

Substituting β,

$$r = \sqrt{\frac{(1+\eta)^2\delta^2}{4} - \frac{\alpha}{(1-\alpha)^2}(\eta^2\delta^2 - \|x_1 - x_2\|_2^2)} \le \frac{(1+\eta)\delta}{2}.$$

substituting this to LHS of Eq (76) and also $\theta = \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta}$ as defined in Eq (39),

$$\begin{aligned}\frac{2\alpha\eta\delta}{1-\alpha} + 2(1+\theta)r &\le \frac{2\alpha\eta\delta}{1-\alpha} + 2\left(1 + \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta}\right)\frac{(1+\eta)\delta}{2} \\ &= \frac{2\alpha\eta\delta}{1-\alpha} + 2\left(\frac{1-\alpha-\alpha\eta}{(1+\eta)(1-\alpha)}\right)(1+\eta)\delta = 2\delta.\end{aligned}$$

Thus, we have shown that for any $x \in \mathbf{B}^2_{x_1}(\delta)$, $\|\nabla f(x)\|_2 \le 2\delta$, which is our desired Lemma 16.4 with $s = 1$.

E.6. Proof of Lemma 17

We start by showing the following technical lemma, which illustrates how the functions after the smooth max operations look like. Its proof is deferred to Appendix E.8.

Lemma 24 *For any set of parameters $u_1, u_2, \ldots, u_M, v$ chosen by $u_1 \in \mathcal{U}^{(1)}$, $u_t \in \mathcal{U}^{(t)}_{u_{t-1}}$ for $t \in 2:M$, and $v \in \mathcal{V}$, run Algorithm 3 and get $f^v_{u_{1:M}}(x)$. Then, for any $t \in 2:M$, we have:*

1. $f_{u_{1:t}}(x) = \begin{cases} f_{u_{1:t-1}}(x) & \forall x \notin \mathbf{B}^2_{u_{t-1}}(\delta_{t-1}), \\ h_{u_{1:t}}(x) = g_{u_{1:t}}(x) & \forall x \in \mathbf{B}^2_{u_{t-1}}(\eta\delta_{t-1}), \end{cases}$
2. $C\beta \sum_{m=1}^{t-1} \alpha^{m-1}\delta_m^2 \le f_{u_{1:t}}(x) \le C\beta \sum_{m=1}^{t-2} \alpha^{m-1}\delta_m^2 + C\alpha^{t-2}\delta_{t-1}^2$ *for all* $x \in \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$.
3. $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), g_{u_{1:t}}(x)\}$ *is smooth.*

Also, at the final step,

4. $f^v_{u_{1:M}}(x) = \begin{cases} f_{u_{1:M}}(x) & \forall x \notin \mathbf{B}^2_{u_M}(\delta_M), \\ h^v_{u_{1:M}}(x) = g^v_{u_{1:M}}(x) & \forall x \in \mathbf{B}^2_{u_M}(\eta\delta_M), \end{cases}$
5. $C\beta \sum_{m=1}^{M} \alpha^{m-1}\delta_m^2 \le f^v_{u_{1:M}}(x) \le C\beta \sum_{m=1}^{M-1} \alpha^{m-1}\delta_m^2 + C\alpha^{M-1}\delta_M^2$ *for all* $x \in \mathbf{B}^2_{u_M}(\delta_M)$.
6. $f^v_{u_{1:M}}(x) := \max\{f_{u_{1:M}}(x), g^v_{u_{1:M}}(x)\}$ *is smooth.*

As done for Lemma 6 in Section D.2, we prove Lemma 17.1 and 17.4 using simple and intuitive argument that $\max$ operations done in Algorithm 3 only changes limited parts of the domain. From Lemma 24.1, note that whenever we have $f_{u_{1:t-1}}(x)$ and take $\max$ operation with $g_{u_{1:t}}(x)$ to construct $f_{u_{1:t}}(x)$, any point $\forall x \notin \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$ does not change its value; i.e. $f_{u_{1:t}}(x) = f_{u_{1:t-1}}(x)$. This means that the Line 6: $f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), g_{u_{1:t}}(x)\}$ in Algorithm 3 can only possibly change function values in $\mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$. Also, later iterations of the algorithm do not change that the function values at $x \notin \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$, because $\mathbf{B}^2_{u_{t-1}}(\delta_{t-1}) \supset \mathbf{B}^2_{u_t}(\delta_t) \supset \cdots \supset \mathbf{B}^2_{u_M}(\delta_M)$. From this argument, we can see that $f^v_{u_{1:M}}(x) = f^{\tilde{v}}_{u_{1:t-1},\tilde{u}_{t:M}}(x) = f_{u_{1:t-1}}(x)$ for all $x \notin \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$, therefore proving Lemma 17.1. Similarly, Line 10: $f^v_{u_{1:M}}(x) := \max\{f_{u_{1:M}}(x), g^v_{u_{1:M}}(x)\}$ in Algorithm 3 can only change function values in $\mathbf{B}^2_{u_M}(\delta_M)$, so $f^{-1}_{u_{1:M}}(x) = f^{+1}_{u_{1:M}}(x) = f_{u_{1:M}}(x)$ for all $x \notin \mathbf{B}^2_{u_M}(\delta_M)$, proving Lemma 17.4.

Lemma 17.5 can be implied directly by Lemma 24.5. In order to prove Lemma 17.2, note the following facts from Lemma 24.5 and 24.2:

$$C\beta \sum_{m=1}^{M} \alpha^{m-1}\delta_m^2 \le f^v_{u_{1:M}}(x) \le C\beta \sum_{m=1}^{M-1} \alpha^{m-1}\delta_m^2 + C\alpha^{M-1}\delta_M^2 \text{ for all } x \in \mathbf{B}^2_{u_M}(\delta_M),$$
$$C\beta \sum_{m=1}^{M-1} \alpha^{m-1}\delta_m^2 \le f_{u_{1:M}}(x) \le C\beta \sum_{m=1}^{M-2} \alpha^{m-1}\delta_m^2 + C\alpha^{M-2}\delta_{M-1}^2 \text{ for all } x \in \mathbf{B}^2_{u_{M-1}}(\delta_{M-1}).$$

Note from Lemma 24.4 that $f^v_{u_{1:M}}(x) = f_{u_{1:M}}(x)$ for all $x \notin \mathbf{B}^2_{u_M}(\delta_M)$, and that, for all $x \in \mathbf{B}^2_{u_M}(\delta_M)$,

$$f^v_{u_{1:M}}(x) \le C\beta \sum_{m=1}^{M-1} \alpha^{m-1}\delta_m^2 + C\alpha^{M-1}\delta_M^2 \le C\beta \sum_{m=1}^{M-2} \alpha^{m-1}\delta_m^2 + C\alpha^{M-2}\delta_{M-1}^2.$$

The last inequality is because $\frac{(1-\beta)\delta_{M-1}^2}{\alpha} \geq \delta_M^2$ holds for large enough n by assumption that $\delta_M = o(\delta_{M-1})$. From these observations, we have

$$C\beta \sum_{m=1}^{M-1} \alpha^{m-1}\delta_m^2 \leq f_{u_{1:M}}^v(x) \leq C\beta \sum_{m=1}^{M-2} \alpha^{m-1}\delta_m^2 + C\alpha^{M-2}\delta_{M-1}^2 \text{ for all } x \in \mathbf{B}_{u_{M-1}}^2(\delta_{M-1}).$$

Again note that, for any $x \notin \mathbf{B}_{u_{M-1}}^2(\delta_{M-1})$ we also have $x \notin \mathbf{B}_{u_M}^2(\delta_M)$, so $f_{u_{1:M}}^v(x) = f_{u_{1:M}}(x) = f_{u_{1:M-1}}(x)$. We can repeat a similar argument and obtain

$$C\beta \sum_{m=1}^{M-2} \alpha^{m-1}\delta_m^2 \leq f_{u_{1:M}}^v(x) \leq C\beta \sum_{m=1}^{M-3} \alpha^{m-1}\delta_m^2 + C\alpha^{M-3}\delta_{M-2}^2 \text{ for all } x \in \mathbf{B}_{u_{M-2}}^2(\delta_{M-2}).$$

For any $t \in 2:M$, we can repeat this argument until $\mathbf{B}_{u_{t-1}}^2(\delta_{t-1})$, so that we get

$$C\beta \sum_{m=1}^{t-1} \alpha^{m-1}\delta_m^2 \leq f_{u_{1:M}}^v(x) \leq C\beta \sum_{m=1}^{t-2} \alpha^{m-1}\delta_m^2 + C\alpha^{t-2}\delta_{t-1}^2 \text{ for all } x \in \mathbf{B}_{u_{t-1}}^2(\delta_{t-1}),$$

which directly implies Lemma 17.2 that we are after.

For Lemma 17.3 and 17.6, we will show that the function value $f_{u_{1:M}}^v(x)$ in $\mathbf{B}_{u_{t-1}}^2(\delta_{t-1})$ can be expressed as

$$f_{u_{1:M}}^v(x) = \max\left\{h_{u_{1:t-1}}(x), \max_{k\in t:M}\left\{g_{u_{1:k}}(x)\right\}, g_{u_{1:M}}^v(x)\right\}, \text{ for all } x \in \mathbf{B}_{u_{t-1}}^2(\delta_{t-1}). \tag{77}$$

Notice from Lemma 24.1 that $f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ for all $x \in \mathbf{B}_{u_{t-2}}^2(\eta\delta_{t-2})$. Recall that $\mathbf{B}_{u_{t-1}}^2(\delta_{t-1}) \subset \mathbf{B}_{u_{t-2}}^2(\eta\delta_{t-2})$, so $f_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ in $\mathbf{B}_{u_{t-1}}^2(\delta_{t-1})$. After this point, $f_{u_{1:M}}^v(x)$ is obtained from max operations with $g_{u_{1:t}}, \ldots, g_{u_{1:M}}, g_{u_{1:M}}^v$. This proves Eq (77). Given Eq (77), we prove Lemma 17.3 by showing that for any $x \in \mathbf{B}_{u_{t-1}}^2(\delta_{t-1})$, all the operands of the max operation in Eq (77) satisfy that ℓ_2 norm of the gradient is bounded above by $2C\alpha^{t-2}\delta_{t-1}$. First, the gradient of $h_{u_{1:t-1}}(x) := C\alpha^{t-2}\left\|x-u_{t-1}\right\|_2^2 + C\beta\sum_{m=1}^{t-2}\alpha^{m-1}\delta_m^2$ is $\nabla h_{u_{1:t-1}}(x) = 2C\alpha^{t-2}(x-u_{t-1})$, so

$$\left\|\nabla h_{u_{1:t-1}}(x)\right\|_2 \leq 2C\alpha^{t-2}\delta_{t-1}, \text{ for any } x \in \mathbf{B}_{u_{t-1}}^2(\delta_{t-1}).$$

Now, for $k \in t:M$, apply Lemma 16 for $g_{u_{1:k}}(x) = \textsc{sMax}(h_{u_{1:k-1}}, h_{u_{1:k}}, \alpha, \eta, \delta_{k-1})$. Recall the definition

$$h_{u_{1:k-1}}(x) := C\alpha^{k-2}\left\|x-u_{k-1}\right\|_2^2 + C\beta\sum_{m=1}^{k-2}\alpha^{m-1}\delta_m^2$$
$$h_{u_{1:k}}(x) := C\alpha^{k-1}\left\|x-u_{k}\right\|_2^2 + C\beta\sum_{m=1}^{k-1}\alpha^{m-1}\delta_m^2,$$

then we can note that in terms of the formulation in Lemma 16, $s = C\alpha^{k-2}$, $t = C\beta\sum_{m=1}^{k-2}\alpha^{m-1}\delta_m^2$, and $\delta = \delta_{k-1}$. From Lemma 16.1 and 16.4, we have

$$g_{u_{1:k}}(x) = h_{u_{1:k-1}}(x) \quad \text{for all } x \notin \mathbf{B}_{u_{k-1}}^2(\delta_{k-1}).$$
$$\left\|\nabla g_{u_{1:k}}(x)\right\|_2 \leq 2C\alpha^{k-2}\delta_{k-1} \leq 2C\alpha^{t-2}\delta_{t-1} \quad \text{for all } x \in \mathbf{B}_{u_{k-1}}^2(\delta_{k-1}).$$

For $x \in \mathbf{B}^2_{u_{k-1}}(\delta_{k-1})$ we already proved that $\|\nabla g_{u_{1:k}}(x)\|_2 \leq 2C\alpha^{t-2}\delta_{t-1}$. We now have to consider points in $\mathbf{B}^2_{u_{k-1}}(\delta_{k-1})^c \cap \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$. Note that for $k = t$, this is $\emptyset$. For $k \in (t+1) : M$,

$$\begin{aligned}
&\|\nabla g_{u_{1:k}}(x)\|_2 = \left\|\nabla h_{u_{1:k-1}}(x)\right\|_2 = \left\|2C\alpha^{k-2}(x - u_{k-1})\right\|_2 \\
\leq& 2C\alpha^{k-2}\|x - u_{t-1}\|_2 + 2C\alpha^{k-2}\|u_{t-1} - u_{k-1}\|_2 \leq 2C\alpha^{k-2}\delta_{t-1} + 2C\alpha^{k-2}\eta\delta_{t-1} \\
\leq& 2C\alpha^{t-1}\delta_{t-1} + 2C\alpha^{t-1}\eta\delta_{t-1} = 2C\alpha^{t-2}\delta_{t-1}(\alpha + \eta\alpha) \leq 2C\alpha^{t-2}\delta_{t-1}.
\end{aligned}$$

Thus, for any $x \in \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$, $\|\nabla g_{u_{1:k}}(x)\|_2 \leq 2C\alpha^{t-2}\delta_{t-1}$ for $k \in t : M$. We can prove this inequality for $g^v_{u_{1:M}}(x)$ in the same way. Since all the function in the max operation satisfies upper bound on ℓ_2 norm of gradient, we have

$$\left\|\nabla f^v_{u_{1:M}}(x)\right\|_2 \leq 2C\alpha^{t-2}\delta_{t-1} \text{ for all } x \in \mathbf{B}^2_{u_{t-1}}(\delta_{t-1}),$$

which implies Lemma 17.3. From a similar argument as Eq (77), we have

$$f^v_{u_{1:M}}(x) = \max\left\{h_{u_{1:M}}(x), g^v_{u_{1:M}}(x)\right\}, \text{ for all } x \in \mathbf{B}^2_{u_M}(\delta_M),$$

whereby we can prove Lemma 17.6.

Finally, we have to show Lemma 17.7. To do so, we first show that, for any choice of $u_1, u_2, \ldots, u_M$ and v,

$$\inf_x f^v_{u_{1:M}}(x) = C\beta\sum_{m=1}^{M}\alpha^{m-1}\delta_m^2. \tag{78}$$

In fact, from Lemma 24.4, we have $f^v_{u_{1:M}}(x) = h^v_{u_{1:M}}(x)$ for all $x \in \mathbf{B}^2_{u_M}(\eta\delta_M)$. Also, $h^v_{u_{1:M}}(x)$ is minimized at $u_M + v\eta\delta_M\mathbf{e}_1 \in \mathbf{B}^2_{u_M}(\eta\delta_M)$, whose minimum value is the RHS of Eq (78). So, for any $x \in \mathbf{D}$,

$$f^v_{u_{1:M}}(x) \geq h^v_{u_{1:M}}(x) \geq h^v_{u_{1:M}}(u_M + v\eta\delta_M\mathbf{e}_1) = C\beta\sum_{m=1}^{M}\alpha^{m-1}\delta_m^2,$$

proving Eq (78).

Next, we show that

$$\inf_x(f^{+1}_{u_{1:M}}(x) + f^{-1}_{u_{1:M}}(x)) = 2C\beta\sum_{m=1}^{M}\alpha^{m-1}\delta_m^2 + 2C\alpha^M\eta^2\delta_M^2. \tag{79}$$

Again note that $f^v_{u_{1:M}}(x) = h^v_{u_{1:M}}(x)$ for all $x \in \mathbf{B}^2_{u_M}(\eta\delta_M)$. That is, for $x \in \mathbf{B}^2_{u_M}(\eta\delta_M)$, we have $f^{+1}_{u_{1:M}}(x) = h^{+1}_{u_{1:M}}(x)$ and $f^{-1}_{u_{1:M}}(x) = h^{-1}_{u_{1:M}}(x)$. Therefore, for any $x \in \mathbf{B}^2_{u_M}(\eta\delta_M)$,

$$\begin{aligned}
&f^{+1}_{u_{1:M}}(x) + f^{-1}_{u_{1:M}}(x) = h^{+1}_{u_{1:M}}(x) + h^{-1}_{u_{1:M}}(x) \\
=&C\alpha^M\left(\|x - u_M - \eta\delta_M\mathbf{e}_1\|_2^2 + \|x - u_M + \eta\delta_M\mathbf{e}_1\|_2^2\right) + 2C\beta\sum_{m=1}^{M}\alpha^{m-1}\delta_m^2.
\end{aligned}$$

Note also that $x = u_M$ attains minimum, which evaluates to the RHS of Eq (79). So, for any $x \in \mathbf{D}$,

$$\begin{aligned}
f^{+1}_{u_{1:M}}(x) + f^{-1}_{u_{1:M}}(x) &\geq h^{+1}_{u_{1:M}}(x) + h^{-1}_{u_{1:M}}(x) \geq h^{+1}_{u_{1:M}}(u_M) + h^{-1}_{u_{1:M}}(u_M) \\
&= 2C\beta\sum_{m=1}^{M}\alpha^{m-1}\delta_m^2 + 2C\alpha^M\eta^2\delta_M^2,
\end{aligned}$$

thus proving Eq (79). Now, Lemma 17.7 follows from Eq (78) and Eq (79).

E.7. Proof of Lemma 18

First note that

$$f^v_{u_{1:M}}(x) = \max\left\{h_{u_1}(x), \max_{k\in 2:M}\left\{g_{u_{1:k}}(x)\right\}, g^v_{u_{1:M}}(x)\right\}. \tag{80}$$

As seen in Lemma 24.6, $f^v_{u_{1:M}}$ is smooth. Thus, the smoothness constant is determined by the "piece" of the function with the largest smoothness constant. This appears at $g_{u_{1:2}}(x)$, whose smoothness constant is $C\left(2+\frac{1-\alpha}{\theta}\right)$, as seen from Lemma 15. In order to prove that $f^v_{u_{1:M}}(x)$ is $2C\alpha^M$-strongly convex, it suffices to show that every operand in the $\max$ operation in Eq (80) is at least $2C\alpha^M$-strongly convex. This can be readily checked using Lemma 15.

Now, we are ready to pick parameters α, η, and C of Algorithm 3 so as to make sure output $f^v_{u_{1:M}}(x)$ is H-smooth and λ-strongly convex, for the case $H/5 \geq \lambda$. The other case ($H/5 < \lambda$) will be handled later. That is, we have to choose the right parameters to make

$$H \geq C\left(2+\frac{1-\alpha}{\theta}\right) \text{ and } \lambda \leq 2C\alpha^M. \tag{81}$$

We first choose

$$\eta = \frac{1-\alpha}{2},$$

so that $0 < \eta < \frac{1}{2}$ for $\alpha \in (0,1)$. Which this choice, we have

$$\eta + \alpha + \alpha\eta = \frac{1}{2} + \alpha - \frac{\alpha^2}{2},$$

which satisfies $\frac{1}{2} < \eta + \alpha + \alpha\eta < 1$ for $\alpha \in (0,1)$. Also, from Eq (38),

$$\beta := \frac{(1-\alpha)(1+\eta)^2}{4} - \frac{\alpha\eta^2}{1-\alpha} = \frac{(9-\alpha)(1-\alpha)^2}{16},$$

which satisfies $0 < \beta < \frac{9}{16}$ for $\alpha \in (0,1)$. From Eq (39),

$$\theta := \frac{1-\eta-\alpha-\alpha\eta}{1+\eta-\alpha-\alpha\eta} = \frac{1-\alpha}{3-\alpha},$$

which satisfies $0 < \theta < \frac{1}{3}$ for $\alpha \in (0,1)$. We finished checking that constraints on η, β, and θ are met, under this particular choice of η.

With this choice of η,

$$\frac{1-\alpha}{\theta} = 3-\alpha,$$

so $2 < \frac{1-\alpha}{\theta} < 3$ for $\alpha \in (0,1)$. This also means that

$$C\left(2+\frac{1-\alpha}{\theta}\right) \leq 5C.$$

Now choose

$$\alpha = \left(\frac{1}{2}\right)^{1/M}, \text{ and } C = \frac{H}{5}.$$

With $\eta = (1-\alpha)/2$, we can check that

$$C\left(2 + \frac{1-\alpha}{\theta}\right) \le 5C \le H, \text{ and } 2C\alpha^M = \frac{2H}{5} \cdot \frac{1}{2} \ge \lambda,$$

thus proving Eq (81).

For $H/5 < \lambda < H$, the choice of parameters is a bit more complicated; we choose $\eta = \frac{(1-\alpha)^2}{2}$, which satisfies $0 < \eta < \frac{1}{2}$ for $\alpha \in (0,1)$. With this choice, $\frac{1}{2} < \eta + \alpha + \eta\alpha < 1$. Also, $0 < \beta < \frac{9}{16}$ and $\frac{1}{3} < \theta < 1$ is satisfied, so all the constraints are met. With this choice of η,

$$\frac{H}{\lambda} = \frac{1}{2\alpha^M}\left(2 + \frac{1-\alpha}{\theta}\right) = \frac{1}{2\alpha^M}\left(2 + \frac{(1-\alpha)(\alpha^2 - 2\alpha + 3)}{1+\alpha^2}\right).$$

When expressed as a function of α, the RHS of the last equation has limit ∞ when $\alpha \to 0^+$, and limit 1 when $\alpha \to 1^-$. Therefore, for any ratio of $H/\lambda > 1$, there exists a α_0 that satisfies the above equation. Choose $\alpha = \alpha_0$, $\eta = \frac{(1-\alpha_0)^2}{2}$, and $C = \frac{\lambda}{2\alpha_0^M}$, then the output of Algorithm 3 is H-smooth and λ-strongly convex.

E.8. Proof of Lemma 24

We demonstrate in details the proof for Lemma 24 below, which is based on an induction argument.

Base case $t = 2$**.** In the base case, recall the definitions that

$$f_{u_1}(x) = h_{u_1}(x) := C\left\|x - u_1\right\|_2^2, \; h_{u_{1:2}}(x) := C\alpha\left\|x - u_2\right\|_2^2 + C\beta\delta_1^2,$$
$$g_{u_{1:2}}(x) := \textsc{sMax}(h_{u_1}, h_{u_{1:2}}, \alpha, \eta, \delta_1)$$

Apply Lemma 16 to h_{u_1} and $h_{u_{1:2}}$. Note that in this case $s = C$, $t = 0$, and $\delta = \delta_1$. Then, Lemma 24.1–2 is immediately proved by Lemma 16.1 and 16.3.

For Lemma 24.3, we want to prove that $f_{u_{1:2}}(x) = \max\{f_{u_1}(x), g_{u_{1:2}}(x)\}$ is smooth. By $f_{u_1}(x) = h_{u_1}(x)$, already $g_{u_{1:2}}(x) \ge h_{u_1}(x) = f_{u_1}(x)$. Thus, $f_{u_{1:2}}(x) = g_{u_{1:2}}(x)$, and it is proven to be smooth by Lemma 15.

Inductive case $2 < t \le M$**.** Recall the definitions that

$$\begin{aligned}
h_{u_{1:t-1}}(x) &:= C\alpha^{t-2}\left\|x - u_{t-1}\right\|_2^2 + C\beta\sum\nolimits_{m=1}^{t-2}\alpha^{m-1}\delta_m^2,\\
h_{u_{1:t}}(x) &:= C\alpha^{t-1}\left\|x - u_t\right\|_2^2 + C\beta\sum\nolimits_{m=1}^{t-1}\alpha^{m-1}\delta_m^2,\\
g_{u_{1:t-1}}(x) &:= \textsc{sMax}(h_{u_{1:t-2}}, h_{u_{1:t-1}}, \alpha, \eta, \delta_{t-2}),\\
g_{u_{1:t}}(x) &:= \textsc{sMax}(h_{u_{1:t-1}}, h_{u_{1:t}}, \alpha, \eta, \delta_{t-1}),\\
f_{u_{1:t-1}}(x) &:= \max\{f_{u_{1:t-2}}(x), g_{u_{1:t-1}}(x)\},\\
f_{u_{1:t}}(x) &:= \max\{f_{u_{1:t-1}}(x), g_{u_{1:t}}(x)\}.
\end{aligned}$$

Apply Lemma 16 to $h_{u_{1:t-1}}$ and $h_{u_{1:t}}$. Note that in this case $s = C\alpha^{t-2}$, $t = C\beta\sum_{m=1}^{t-2}\alpha^{m-1}\delta_m^2$, and $\delta = \delta_{t-1}$. By Lemma 16.1–3,

$$g_{u_{1:t}}(x) = h_{u_{1:t-1}}(x) \quad \text{for any } x \in \mathrm{cl}(\mathbf{B}^2_{u_{t-1}}(\delta_{t-1})^c), \tag{82}$$

$$g_{u_{1:t}}(x) = h_{u_{1:t}}(x) \quad \text{for any } x \in \mathbf{B}^2_{u_{t-1}}(\eta\delta_{t-1}), \tag{83}$$

$$C\beta \sum_{m=1}^{t-1} \alpha^{m-1}\delta_m^2 \le g_{u_{1:t}}(x) \le C\beta \sum_{m=1}^{t-2} \alpha^{m-1}\delta_m^2 + C\alpha^{t-2}\delta_{t-1}^2 \quad \text{for any } x \in \mathbf{B}^2_{u_{t-1}}(\delta_{t-1}), \tag{84}$$

$$\nabla g_{u_{1:t}}(x) = \nabla h_{u_{1:t-1}}(x) \quad \text{for any } x \in \mathrm{cl}(\mathbf{B}^2_{u_{t-1}}(\delta_{t-1})^c). \tag{85}$$

We will prove Lemma 24.1–3 using this list of facts. Note that $f_{u_{1:t-1}}(x) \ge g_{u_{1:t-1}}(x) \ge h_{u_{1:t-1}}(x)$ for all $x \in \mathbf{D}$. Together with Eq (82), this yields

$$f_{u_{1:t-1}}(x) \ge g_{u_{1:t}}(x) \text{ for all } x \in \mathrm{cl}(\mathbf{B}^2_{u_{t-1}}(\delta_{t-1})^c). \tag{86}$$

For the other case, by using Lemma 24.1 for case $t-1$ (induction hypothesis), we have $f_{u_{1:t-2}}(x) \le g_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ for all $x \in \mathbf{B}^2_{u_{t-2}}(\eta\delta_{t-2})$. From the definition $f_{u_{1:t-1}}(x) := \max\{f_{u_{1:t-2}}(x), g_{u_{1:t-1}}(x)\}$, we have $f_{u_{1:t-1}}(x) = g_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x)$ for $x \in \mathbf{B}^2_{u_{t-2}}(\eta\delta_{t-2})$. Note that $\mathbf{B}^2_{u_{t-2}}(\eta\delta_{t-2}) \supset \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$, so

$$f_{u_{1:t-1}}(x) = g_{u_{1:t-1}}(x) = h_{u_{1:t-1}}(x) \text{ for all } x \in \mathbf{B}^2_{u_{t-1}}(\delta_{t-1}). \tag{87}$$

By Eq (83), $g_{u_{1:t}}(x) = h_{u_{1:t}}(x) \ge h_{u_{1:t-1}}(x)$ for any $x \in \mathbf{B}^2_{u_{t-1}}(\eta\delta_{t-1})$. With Eq (87), this proves

$$f_{u_{1:t-1}}(x) \le g_{u_{1:t}}(x) = h_{u_{1:t}}(x) \text{ for } x \in \mathbf{B}^2_{u_{t-1}}(\eta\delta_{t-1}),$$

hence finishing the proof of Lemma 24.1.

By Eq (87), and that $h_{u_{1:t-1}}(x) \le g_{u_{1:t}}(x)$, we have $f_{u_{1:t-1}}(x) \le g_{u_{1:t}}(x)$ for all $x \in \mathbf{B}^2_{u_{t-1}}(\delta_{t-1})$. Together with Eq (86), this means

$$f_{u_{1:t}}(x) := \max\{f_{u_{1:t-1}}(x), g_{u_{1:t}}(x)\} = \begin{cases} g_{u_{1:t}}(x) & \text{if } \|x - u_{t-1}\|_2 \le \delta_{t-1}, \\ f_{u_{1:t-1}}(x) & \text{if } \|x - u_{t-1}\|_2 \ge \delta_{t-1}. \end{cases} \tag{88}$$

Combining Eqs (84) and (88), this proves Lemma 24.2.

We now have Lemma 24.3 to prove. If we look into Eq (88) more closely, we can see that $f_{u_{1:t-1}}(x) = g_{u_{1:t}}(x)$ if $\|x - u_{t-1}\|_2 = \delta_{t-1}$. We know by induction hypothesis that $f_{u_{1:t-1}}(x)$ is smooth, and by Lemma 15 that $g_{u_{1:t}}(x)$ is also smooth. So, the proof of smoothness of $f_{u_{1:t}}(x)$ suffices to check if $\nabla f_{u_{1:t-1}}(x) = \nabla g_{u_{1:t}}(x)$ for all x such that $\|x - u_{t-1}\|_2 = \delta_{t-1}$. From Eq (85),

$$\nabla g_{u_{1:t}}(x) = \nabla h_{u_{1:t-1}}(x) \text{ if } \|x - u_{t-1}\|_2 = \delta_{t-1}.$$

Also, $f_{u_{1:t-1}}(x)$ is a smooth function, meaning that $\nabla f_{u_{1:t-1}}(x) = \nabla g_{u_{1:t-1}}(x)$ whenever $f_{u_{1:t-1}}(x) = g_{u_{1:t-1}}(x)$. Together with Eq (87), we have

$$\nabla f_{u_{1:t-1}}(x) = \nabla g_{u_{1:t-1}}(x) = \nabla h_{u_{1:t-1}}(x) \text{ if } \|x - u_{t-1}\|_2 = \delta_{t-1}.$$

This shows $\nabla f_{u_{1:t-1}}(x) = \nabla g_{u_{1:t}}(x)$ whenever x satisfies $\|x - u_{t-1}\|_2 = \delta_{t-1}$. So $f_{u_{1:t}}$ is smooth, hence Lemma 24.3 is shown.

Final Case. It is left to prove Lemma 24.4–6. Their proof can be done in a similar way as Lemma 24.1–3 for the inductive cases, hence omitted.

Appendix F. Proof of Theorem 3: Roadmap

The proof of theorem 3 is fairly involved. Indeed, we divide the proof of Thoerem 3 into four parts: in Section G, we present the proof for the case where the function class $\mathcal{F} = \mathcal{F}_{H,\lambda}$ and the oracle is first order oracle, in Section H, we present the proof for the case where the function class $\mathcal{F} = \mathcal{F}_{H,\lambda}$ and the oracle is zeroth-order oracle, in Section I, we present the proof for the case where the function class $\mathcal{F} = \mathcal{F}_L$ and the oracle is the first order oracle, and lastly, in Section J, we present the proof for the case where the function class $\mathcal{F} = \mathcal{F}_L$ and the oracle is zeroth order oracle. Broadly speaking, for each of those four cases, we prove the corresponding upper bound via first introducing a concrete algorithm, and then showing that the algorithm achieves the upper bound in Theorem 3 through careful theoretical justifications.

Common Notations from Section G to Section J. The following notations are useful throughout the proofs of the upper bounds from Section G to Section J. For any $c \in \mathbb{R}^d$, $r \in \mathbb{R}_+$ and $k \in \mathbb{N}$, we use $G(c, r, k)$ to denote the following grids in $\mathbb{R}^d$:

$$G(c,r,k) := \left\{ c + \left(\frac{r}{k} i_1, \frac{r}{k} i_2, \ldots, \frac{r}{k} i_d\right)' \mid i_j \in \{-k, -(k-1), \ldots, k-1, k\} \text{ for all } 1 \le j \le d \right\}.$$

Denote $\gamma : \mathbb{R}_+ \to \mathbb{R}$ to be the function $\gamma(x) := x/\log(x)$, which we heavily use in the proof.

Appendix G. Smooth Function with First Order Oracle

G.1. Description of Algorithms

In this section, we propose two generic algorithms: algorithm 4 parameterized by parameters $(c, r, k, T) \in \mathbf{D} \times \mathbb{R}_+ \times \mathbb{N} \times \mathbb{N}$ and algorithm 5 that builds from algorithm 4. We note here that, algorithm 5 in essence, builds from M times of repeated calls of algorithm 4. As will be shown immediately in the later subsections, it turns out that the two algorithms with careful choice of parameters return the minimax estimator in single and multi rounds respectively.

G.2. Analysis of Algorithm 4: Single-Stage Analysis

In this section, we analyze the single-stage algorithm 4. For purpose of convenience in later discussion for multi-stage algorithm 5, we slightly generalize the domain of interest. In fact, we consider the domain to be $\mathbf{D}_{c,r} = \{x : \|x - c\|_\infty \le r\}$ parameterized by $c \in \mathbb{R}^d$ and $r \in \mathbb{R}_+$, and the goal of the algorithm is to find $x^\star_{f,c,r}$, the minimum of f in $\mathbf{D}_{c,r}$ so that the performance of the algorithm is measured by $\mathbb{E}[f(\hat{x}) - f(x^\star_{f,c,r}))]$. Note that, if we substitute $c = \frac{1}{2} \cdot \mathbf{1}$ and $r = \frac{1}{2}$ into the results below, it leads to corresponding results to the original domain $\mathbf{D} = [0,1]^d$.

Proposition 25 *Given any fix $c \in \mathbb{R}^d$ and $r \in (0, 1]$, suppose there exists $k \in \mathbb{N}$ satisfying*

$$(2k+1)^d \left\lceil 2k^2 \log(2k+1) \right\rceil \le nr^2, \text{ and } k > \frac{10}{\lambda} \left(\sigma \left(\log \frac{1}{\delta} + 3d \right)^{\frac{1}{2}} + Hd^{\frac{1}{2}} \right). \tag{89}$$

Then, pick any k satisfying Eq (89) and set $T = \left\lfloor \frac{n}{(2k+1)^d} \right\rfloor$. Denote $\hat{c}$, $\hat{r}$ and $\hat{x}$ to be the output from algorithm 4 when we input (c, k, T, r) as the input parameters. Then, we have,

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,c,r}) \le 2Hd\hat{r}^2 \text{ and } \left\| x^\star_{f,c,r} - \hat{c} \right\|_\infty \le \hat{r} \right) \ge 1 - \delta.$$

Algorithm 4 Generic Routine for One Stage Smooth Functions $\mathcal{F}_{H,\lambda}$ (First-order Oracle)

Input: Prior knowledge on $\lambda, H \in \mathbb{R}_+$ satisfying $\lambda \leq H$ and the noise level $\sigma \in \mathbb{R}_+$. User specify the sampling center $c \in \mathbf{D}$, radius $r \in \mathbb{R}_+$, grid size parameter $k \in \mathbb{N}$, the sampling times $T \in \mathbb{N}$ and the confidence level $\delta \in (0,1)$.

1: Compute the grid points $G = G(c, r, k)$.

2: At each point $x \in G$, query the first oracle T times and denote each sample gradient value via $\{\widehat{\nabla} f(x)^{(1)}, \widehat{\nabla} f(x)^{(2)}, \ldots, \widehat{\nabla} f(x)^{(T)}\}$.

3: Compute the gradient estimate at each point $x \in G$ via $\widehat{\nabla} f(x) := \frac{1}{T}\sum_{i=1}^{T} \widehat{\nabla} f(x)^{(i)}$.

4: Define r^S as $r^S := \frac{4r}{k\lambda} \cdot \left(2\sigma\left(\log\frac{1}{\delta} + 3d\right)^{\frac{1}{2}} + Hd^{\frac{1}{2}}\right)$. Compute the 'candidate' set $S \in G$:

$$S = \left\{ x \in G : \left\langle \widehat{\nabla} f(x), y - x \right\rangle + \frac{\lambda}{2}\|y - x\|_2^2 > 0 \text{ for all } y \in G \text{ satisfying } \|y - x\|_2 \geq r^S \right\}.$$

5: Find the center $x^S \in S$, defined by $x^S := \operatorname{argmin}_{x \in S} \max_{y \in S} \|x - y\|_2$.

6: Define $\hat{r}$ as $\hat{r} := \frac{10r}{\lambda k}\left(\sigma\left(\log\frac{1}{\delta} + 3d\right)^{\frac{1}{2}} + Hd^{\frac{1}{2}}\right)$. Define the following rectangular W by:

$$W := \left\{ x \in \mathbf{D} : \left\| x - x^S \right\|_\infty \leq \hat{r} \right\}$$

7: Return the center $\hat{c} = x^S$, the radius $\hat{r}$ and the estimate $\hat{x} \in W$, defined by

$$\hat{x} := \operatorname*{argmax}_{x \in W} |\{i \in \{1, 2, \ldots, d\} : |x_i - c_i| = r\}|,$$

where for each $1 \leq i \leq d$, x_i, c_i denotes the ith coordinate of $x, c \in \mathbb{R}^d$.

Algorithm 5 Generic Routine for Multi-stage Smooth Functions $\mathcal{F}_{H,\lambda}$ (First-order Oracle)

Input: Prior knowledge on $\lambda, H \in \mathbb{R}_+$ satisfying $\lambda \leq H$, the noise level $\sigma \in \mathbb{R}_+$ and number of rounds $R \in \mathbb{N}_+$. Initialization of parameters $(c_1, r_1, k_1, T_1) \in \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{N} \times \mathbb{N}$. User specifies the confidence level $\delta \in (0, 1]$ and the updating rule used in line (3) of the algorithm.

1: **for** $i = 1$ to M **do**

2: Run algorithm 4 with input parameters (c_i, r_i, k_i, T_i). Denote the output to be $\hat{c}_i$, radius $\hat{r}_i$ and estimate $\hat{x}_i$.

3: Update $(c_{i+1}, r_{i+1}, k_{i+1}, T_{i+1})$. The updating rule may take $\hat{c}_i$, $\hat{r}_i$ and $\hat{x}_i$ as input.

4: **end for**

5: **return** $\hat{x}_M$ as estimate of $x_f^\star$ and the radius r_{M+1}.

Remark 26 *Before we give the proof of proposition 25, we give some high-level intuitions why the algorithm 4* should *work. First of all, since the noise are all light-tailed random variables, it is expected that all gradient estimate $\hat{\nabla} f(x)$ concentrates at the true gradient $\nabla f(x)$, so that they give useful information of the gradient value at all grid points in G. So the question becomes, how can we utiliize the noisy gradient information in G to find the minimum $x^\star_{f,c,r}$, which can be characterized as the following:*

$$\langle \nabla f(x^\star_{f,c,r}), x - x^\star_{f,c,r} \rangle \geq 0 \text{ for all } x \in \mathbf{D}_{c,r}. \tag{90}$$

Due to the construction of the grids, it is immediate that there always exists some point $\bar{x}$ belonging to the grid set G such that $\bar{x}$ is close to $x^\star_{f,c,r}$ up to $\left\|\bar{x} - x^\star_{f,c,r}\right\|_\infty \leq \frac{r}{k}$. A more careful analysis from lemma 28 gives us a stronger but useful result: we can always find some $\bar{x} \in G$ such that it satisfies below two equations simultaneously,

$$\left\|\bar{x} - x^\star_{f,c,r}\right\|_\infty \leq \frac{r}{k} \text{ and } \langle \nabla f(x^\star_{f,c,r}), x - \bar{x} \rangle \geq 0 \text{ for all } x \in \mathbf{D}_{c,r}. \tag{91}$$

Basically, Eq (91) and Eq (90) tell us that, there always exists some grid point $\bar{x} \in G$ close enough to $x^\star_{f,c,r}$ so that, it has similar property as $x^\star_{f,c,r}$. This motivates us to search for points as $\bar{x} \in G$. One difficulty in searching for $\bar{x}$ is that its characterization from the second inequality of Eq (91), as it requires knowledge of the value $\nabla f(x^\star_{f,c,r})$. Our strategy is to approximate the unknown gradient value $\nabla f(x^\star_{f,c,r})$ by $\hat{f}(\bar{x})$. However, due to such approximation, we need additional regularization to make the term to be non-negative. This motivates us to define the set S in line 4 of the algorithm 4. As will be made more precise in lemma 29, we know that, with high probability, $\bar{x} \in S$ and all points in S is close to $\bar{x}$ up to $O(k^{-1})$. This crucial observation also motivates the construction of our algorithm from line 5 to line 7. In line 5, we first find the center of the set S, i.e, x^S and then in line 6 construct the set W centered at x^S with an appropriate radius so that we can make sure both $\bar{x}$ and thus $x^\star_{f,c,r}$ are within the box W with high probability. Finally, we carefully select $\hat{x}$ from the box W in line 7 to make sure the function value $f(\hat{x})$ is also close to the minimum $f(x^\star_{f,c,r})$.

Proof We start proving the proposition by considering the following high probability event. Denote Γ to be the following event:

$$\Gamma = \left\{ \left\|\hat{\nabla} f(x) - \nabla f(x)\right\|_2 \leq r^a := 2\sigma \sqrt{\frac{2}{T}\left(\log\frac{1}{\delta} + 2d + d\log(2k+1)\right)} \text{ for all } x \in G \right\}.$$

The lemma below shows that Γ happens with probability at least $1 - \delta$.

Lemma 27 *We have $\mathbb{P}(\Gamma) \geq 1 - \delta$.*

Proof First, let us for each $x \in G$, denote $\epsilon(x) := \hat{\nabla} f(x) - \nabla f(x)$. Then, since by our assumption the noise vectors $\{\hat{\nabla} f(x)^{(i)} - \nabla f(x)\}_{i=1}^{T_1}$ are mean $\mathbf{0}$, independent and is subgaussian with parameter σ^2, we have that $\epsilon(x)$ is mean $\mathbf{0}$ and is subgaussian with parameter σ^2/T. Therefore, we have, for any fix $x \in G$,

$$\mathbb{P}\left(\|\epsilon(x)\|_2 \geq r^a\right) \leq \exp(2d)\exp\left(-\frac{(r^a)^2 T}{8\sigma^2}\right) \leq \delta(2k+1)^{-d},$$

where the last inequality above uses the definition of r^a. Now, the desired claim of the lemma follows from the fact that $|G| = (2k+1)^d$ and the union bound of the above events. ■

Note that, since our condition on k in Eq (89), we get, $T \geq 2k^2 \log(2k+1) r^{-2}$, and hence,

$$r^a = 2\sigma \sqrt{\frac{2}{T}\left(\log\frac{1}{\delta} + 2d + d\log(2k+1)\right)} \leq \frac{2\sigma r}{k}\left(\log\frac{1}{\delta} + 3d\right)^{\frac{1}{2}}. \tag{92}$$

Lemma 28 *There exists some $\bar{x} \in G$ satisfying the below conditions:*

$$\left\|\bar{x} - x^\star_{f,c,r}\right\|_\infty \leq \frac{r}{k} \text{ and } \left\langle \nabla f(x^\star_{f,c,r}), \bar{x} - x^\star_{f,c,r}\right\rangle = 0. \tag{93}$$

Any $\bar{x}$ satisfying Eq (93) satisfies the crucial property below:

$$\left\langle \nabla f(x^\star_{f,c,r}), x - \bar{x}\right\rangle \geq 0 \text{ for all } x \in \mathbf{D}_{c,r}. \tag{94}$$

Proof To start with, let us define the set $K^\star$ to be:

$$K^\star = \{i \in \{1,2,\ldots,d\} : |(x^\star_{f,c,r})_i - c_i| = r\},$$

where in above, $(x^\star_{f,c,r})_i$ denotes the ith coordinate of $x^\star_{f,c,r}$. Now, using optimality condition of $x^\star_{f,c,r}$ (actually complementary slackness condition from the KKT characterization of $x^\star_{f,c,r}$), we know that, $(\nabla f(x^\star_{f,c,r}))_i = 0$ for all $i \notin K^\star$. Now, denote the following sets:

$$R^{K^\star} = \{x \in \mathbb{R}^d : x_i = (x^\star_{f,c,r})_i \text{ for all } i \in K^\star\},\ G^{K^\star} = G \cap R^{K^\star} \text{ and } \mathbf{D}^{K^\star} = \mathbf{D} \cap R^{K^\star}.$$

Then $x^\star_{f,c,r} \in \mathbf{D}^{K^\star}$. Now, denote $\bar{x} = \operatorname{argmin}_{x \in G^{K^\star}} \left\|x - x^\star_{f,c,r}\right\|_\infty$. Then, it is easy to see that, $\left\|\bar{x} - x^\star_{f,c,r}\right\|_\infty \leq \frac{r}{k}$ since $G^{K^\star}$ forms a set of grid points of $\mathbf{D}^{K^\star}$. Thus, we have,

$$\left\langle \nabla f(x^\star_{f,c,r}), \bar{x} - x^\star_{f,c,r}\right\rangle = \sum_{i \in K^\star} (\nabla f(x^\star_{f,c,r}))_i \underbrace{(\bar{x}_i - x^\star_{f,i})}_{0} + \sum_{i \notin K^\star} \underbrace{(\nabla f(x^\star_{f,c,r}))_i}_{0} (\bar{x}_i - x^\star_{f,i}) = 0.$$

Now, the above identity and the optimality condition of $x^\star_{f,c,r}$ in Eq (90) together imply that

$$\left\langle \nabla f(x^\star_{f,c,r}), x - \bar{x}\right\rangle = \left\langle \nabla f(x^\star_{f,c,r}), x - x^\star_{f,c,r}\right\rangle \geq 0 \text{ for all } x \in \mathbf{D}_{c,r}.$$

■

Lemma 29 *Let $\bar{x}$ be any point in $\mathbf{D}_{c,r}$ satisfying Eq (93). Denote S to be the 'candidate' set $S \subseteq G$ in the line 4 of algorithm 4. Then, on event Γ, we have,*

$$\bar{x} \in S \text{ and } S \subseteq \left\{x \in \mathbf{D}_{c,r} : \|x - \bar{x}\|_2 \leq \frac{4r}{\lambda k}\left(2\sigma\left(\log\frac{1}{\delta} + 3d\right)^{\frac{1}{2}} + Hd^{\frac{1}{2}}\right)\right\}. \tag{95}$$

Proof Throughout the proof, we assume event Γ happens. Note the upper bound on r^a in Eq (92), we know that, on event Γ,

$$\sup_{x\in G}\left\|\widehat{\nabla}f(x)-\nabla f(x)\right\|_2 \le \frac{2\sigma r}{k}\left(\log\frac{1}{\delta}+3d\right)^{\frac{1}{2}}. \tag{96}$$

To show the desired result, let us define $\mathbf{D}^{\text{ex}}_{c,r}$ to be the following subset of $\mathbf{D}$,

$$\mathbf{D}^{\text{ex}}_{c,r}=\left\{x:\|x-\bar{x}\|_2>\frac{4r}{\lambda k}\left(2\sigma\left(\log\frac{1}{\delta}+3d\right)^{\frac{1}{2}}+Hd^{\frac{1}{2}}\right)\right\}.$$

By definition of S, it suffices to show that, for all $x\in\mathbf{D}^{\text{ex}}_{c,r}$,

$$\left\langle\widehat{\nabla}f(\bar{x}),x-\bar{x}\right\rangle+\frac{\lambda}{2}\|x-\bar{x}\|_2^2>0 \text{ and } \left\langle\widehat{\nabla}f(x),\bar{x}-x\right\rangle+\frac{\lambda}{2}\|x-\bar{x}\|_2^2<0. \tag{97}$$

To do so, let $\bar{x}\in\mathbf{D}_{c,r}$ satisfying Eq (93). Then for any $x\in\mathbf{D}_{c,r}$, using Eq (94), we get,

$$\langle\nabla f(\bar{x}),x-\bar{x}\rangle\ge\left\langle\nabla f(\bar{x})-\nabla f(x^\star_{f,c,r}),x-\bar{x}\right\rangle. \tag{98}$$

Now, note that $\bar{x}$ satisfies $\left\|\bar{x}-x^\star_{f,c,r}\right\|_\infty\le\frac{r}{k}$. Since the function f is smooth, we get that,

$$\left\|\nabla f(\bar{x})-\nabla f(x^\star_{f,c,r})\right\|_2\|x-\bar{x}\|_2\le H\left\|\bar{x}-x^\star_{f,c,r}\right\|_2\le H\sqrt{d}\left\|\bar{x}-x^\star_{f,c,r}\right\|_\infty\le\frac{Hr\sqrt{d}}{k}.$$

Hence, by Cauchy Schwartz inequality, we further get,

$$\left\langle\nabla f(\bar{x})-\nabla f(x^\star_{f,c,r}),x-\bar{x}\right\rangle\ge-\left\|\nabla f(\bar{x})-\nabla f(x^\star_{f,c,r})\right\|_2\|x-\bar{x}\|_2\ge-\frac{Hr\sqrt{d}}{k}\|x-\bar{x}\|_2. \tag{99}$$

Noticing the definition of $\mathbf{D}^{\text{ex}}_{c,r}$, Eq (98) and Eq (99) together imply that, for all $x\in\mathbf{D}^{\text{ex}}_{c,r}$,

$$\langle\nabla f(\bar{x}),x-\bar{x}\rangle+\frac{\lambda}{4}\|x-\bar{x}\|_2^2\ge-\frac{Hr\sqrt{d}}{k}\|x-\bar{x}\|_2+\frac{\lambda}{4}\|x-\bar{x}\|_2^2\ge0. \tag{100}$$

Now, using Cauchy Schwartz inequality, we get that, for all $x\in\mathbf{D}^{\text{ex}}_{c,r}$,

$$\left\langle\widehat{\nabla f}(\bar{x})-\nabla f(\bar{x}),x-\bar{x}\right\rangle+\frac{\lambda}{4}\|x-\bar{x}\|_2^2\ge-\left\|\widehat{\nabla f}(\bar{x})-\nabla f(\bar{x})\right\|_2\|x-\bar{x}\|_2+\frac{\lambda}{4}\|x-\bar{x}\|_2^2>0 \tag{101}$$

Now, inequality (100) and (101) gives that for all $x\in\mathbf{D}^{\text{ex}}_{c,r}$,

$$\begin{aligned}&\left\langle\widehat{\nabla}f(\bar{x}),x-\bar{x}\right\rangle+\frac{\lambda}{2}\|x-\bar{x}\|_2^2\\&=\left(\langle\nabla f(\bar{x}),x-\bar{x}\rangle+\frac{\lambda}{4}\|x-\bar{x}\|_2^2\right)+\left(\left\langle\widehat{\nabla}f(\bar{x})-\nabla f(\bar{x}),x-\bar{x}\right\rangle+\frac{\lambda}{4}\|x-\bar{x}\|_2^2\right)>0\end{aligned} \tag{102}$$

This gives the first part of Eq (97). Now, since the function $f(\cdot)$ is λ strongly convex, for all $x \in \mathbf{D}_{c,r}$, we have,

$$\langle \nabla f(\bar{x}) - \nabla f(x), \bar{x} - x \rangle \geq \lambda \left\| x - \bar{x} \right\|_2^2 .$$

Together with Eq (100), it shows that, for all $x \in \mathbf{D}_{c,r}^{\text{ex}}$,

$$\langle \nabla f(x), \bar{x} - x \rangle + \frac{3\lambda}{4} \left\| x - \bar{x} \right\|_2^2 \leq \lambda \left\| x - \bar{x} \right\|_2^2 - \langle \nabla f(\bar{x}) - \nabla f(x), \bar{x} - x \rangle \leq 0. \tag{103}$$

Now, using again by Cauchy Schwartz inequality, we get for all $x \in \mathbf{D}_{c,r}^{\text{ex}}$,

$$\left\langle \widehat{\nabla} f(\bar{x}) - \nabla f(\bar{x}), x - \bar{x} \right\rangle - \frac{\lambda}{4} \left\| x - \bar{x} \right\|_2^2 \leq \left\| \widehat{\nabla} f(\bar{x}) - \nabla f(\bar{x}) \right\|_2 \left\| x - \bar{x} \right\|_2 - \frac{\lambda}{4} \left\| x - \bar{x} \right\|_2^2 < 0. \tag{104}$$

Thus, for all $x \in \mathbf{D}_{c,r}^{\text{ex}}$, the inequality below holds

$$\begin{aligned} &\left\langle \widehat{\nabla} f(x), \bar{x} - x \right\rangle + \frac{\lambda}{2} \left\| x - \bar{x} \right\|_2^2 \\ &= \left(\langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{3\lambda}{4} \left\| x - \bar{x} \right\|_2^2 \right) + \left(\left\langle \widehat{\nabla} f(\bar{x}) - \nabla f(\bar{x}), x - \bar{x} \right\rangle - \frac{\lambda}{4} \left\| x - \bar{x} \right\|_2^2 \right) < 0. \end{aligned}$$

This gives the second part of Eq (97). As discussed previously, this shows that $\bar{x} \in S$. ■

The lemma above has a lot of nice implications. Indeed, denote W to be the set that we construct in the 6th line of the algorithm. In fact, using Eq (93) and Eq (95), on event Γ, the distance between x^S and $x^\star_{f,c,r}$ can be upper bounded by,

$$\left\| x^S - x^\star_{f,c,r} \right\|_2 \leq \left\| x^S - \bar{x} \right\|_2 + \left\| \bar{x} - x^\star_{f,c,r} \right\|_2 \leq \frac{10r}{\lambda k} \left(\sigma \left(\log \frac{1}{\delta} + 3d \right)^{\frac{1}{2}} + Hd^{\frac{1}{2}} \right) = \hat{r}.$$

By definition, this means that $x^\star_{f,c,r} \in W$ on event Γ. Finally, we show that, our careful choice of $\hat{x} \in W$ makes $f(\hat{x})$ is close to $f(x^\star_{f,r,c})$ close up to $O((\hat{r})^2)$.

Lemma 30 *Assume that k satisfies Eq (89). Then, $\hat{r} < r$, and on event Γ, $\hat{x}$ satisfies,*

$$f(\hat{x}) - f(x^\star_{f,c,r}) \leq 2Hd\hat{r}^2.$$

Proof We prove the desired inequality by showing the following crucial property of $\hat{x}$:

$$\left\langle \nabla f(x^\star_{f,c,r}), \hat{x} - x^\star_{f,c,r} \right\rangle = 0 \ \text{ and } \ \left\| \hat{x} - x^\star_{f,c,r} \right\|_2 \leq 2\sqrt{d}\hat{r}. \tag{105}$$

Given above equation, the desired inequality follows, since by smoothness of f,

$$f(\hat{x}) - f(x^\star_{f,c,r}) \leq \left\langle \nabla f(x^\star_{f,c,r}), \hat{x} - x^\star_{f,c,r} \right\rangle + \frac{H}{2} \left\| \hat{x} - x^\star_{f,c,r} \right\|_2^2 \leq 2Hd\hat{r}^2,$$

The rest of the proof is thus devoted to proving Eq (105). Note that, the second inequality of Eq (105) follows easily since on Γ, $x^\star_{f,c,r} \in W$ and the diameter of W is exactly $2\sqrt{d}\hat{r}$. Now, we prove the first equality in Eq (105). To do so, denote $K^\star$ and $\hat{K}$ respectively as follows:

$$K^\star = \{ i \in \{1, 2, \ldots, d\} : |(x^\star_{f,c,r})_i - c_i| = r \} \ \text{ and } \ \hat{K} = \{ i \in \{1, 2, \ldots, d\} : |\hat{x}_i - c_i| = r \}.$$

Now, we show $K^\star \subseteq \hat{K}$. Suppose on the contrary, then, consider some x such that $x_i = \hat{x}_i$ for $i \in K_1 \cup (K^\star)^c$ and $x_i = (x^\star_{f,c,r})_i$ for $i \in K^\star \setminus K_1$. Then $x \in W$ and the set K_x defined below,

$$K_x := \{i \in \{1, 2, \ldots, d\} : |x_i - c_i| = r\} \supseteq K^\star \cup \hat{K},$$

and hence K_x contains K. This contradicts the definition of $\hat{x}$. Thus, we have, $K^\star \subseteq \hat{K}$. Now, the optimality condition of $x^\star$ gives that $(\nabla f(x^\star_{f,c,r}))_i = 0$ for all $i \notin K^\star$. In addition, if k satisfies Eq (89), then $\hat{r} < r$. Since both $x^\star_{f,c,r}$ and $\hat{x}$ belong to W, we know that, $(x^\star_{f,c,r})_i = \hat{x}_i$ for all $i \in K^\star$. Together, it gives the first part of Eq (105), as

$$\left\langle \nabla f(x^\star_{f,c,r}), \hat{x} - x^\star_{f,c,r} \right\rangle = \sum_{i \in K^\star} (\nabla f(x^\star_{f,c,r}))_i \underbrace{(\hat{x}_i - (x^\star_{f,c,r})_i)}_{0} + \sum_{i \notin K^\star} \underbrace{(\nabla f(x^\star_{f,c,r}))_i}_{0} (\hat{x}_i - (x^\star_{f,c,r})_i) = 0.$$

■

The desired claim of the proposition now follows easily from Lemma 27 and Lemma 30. ■

Motivated by Proposition 25, it becomes important to understand when such k exists in Eq (89) and how large it is.

Lemma 31 *Assume n is large enough satisfying*

$$nr^2 \geq (6B)^{2(d+2)}, \quad \textit{where } B = \frac{10}{\lambda}\left(\sigma\left(\log\frac{1}{\delta} + 3d\right)^{\frac{1}{2}} + Hd^{\frac{1}{2}}\right). \tag{106}$$

Denote $k(r) = \left(\gamma\left(nr^2\right)\right)^{\frac{1}{d+2}}$ and $k^\star = \left\lfloor \frac{1}{3}k(r) \right\rfloor$. Then $k^\star \in \mathbb{N}$, and $k^\star$ satisfies Eq (89).

Proof Note that, $\gamma(x) \geq \sqrt{x}$ whenever $x \geq 3$. Thus, by our assumption on n, we get that,

$$\left(\gamma(nr^2)\right)^{\frac{1}{d+2}} \geq 6B \geq 6.$$

This immediately gives us that $k^\star \geq 1$ and $k^\star$ satisfies the second inequality of Eq (89). Now, we show that $k^\star$ satisfies the first inequality of Eq (89). In fact, when $k = k^\star$, we have,

$$(2k+1)^d \left\lceil 2k^2 \log(2k+1) \right\rceil \leq (3k)^{d+2} \log(3k)^{d+2} \leq (k(r))^{d+2} \log(k(r))^{d+2} \leq nr^2,$$

where the last inequality follows from the fact that, for any $x > 0$, $\gamma(x)\log\gamma(x) \leq x$. ■

Proposition 25 and Lemma 31 together immediately give the corollary below.

Corollary 32 *Given any fix $c \in \mathbb{R}^d$ and $r \in [0,1]$, set $k = \left\lfloor \frac{1}{3}\left(\gamma\left(nr^2\right)\right)^{\frac{1}{d+2}} \right\rfloor$, and $T = \left\lfloor \frac{n}{(2k+1)^d} \right\rfloor$. Assume n is large enough satisfying Eq (106). Then if we denote $\hat{c}, \hat{r}$ and $\hat{x}$ to be the output of Algorithm 4 when we input (c, k, T, r) as the input parameters, we have,*

$$\hat{r} \leq \min\{r, 6Br^{\frac{d}{d+2}} n^{-\frac{1}{d+2}} \log(nr^2)^{\frac{1}{d+2}}\}, \quad \textit{where } B = \frac{10}{\lambda}\left(\sigma\left(\log\frac{1}{\delta} + 3d\right)^{\frac{1}{2}} + Hd^{\frac{1}{2}}\right).$$

In addition, we get that,

$$\mathbb{P}\left(f(\hat{x}) - f(x - f, c, r^\star) \leq \gamma^\star\right) \geq 1 - \delta,$$

where

$$\gamma^\star = 2Hd\hat{r}^2 \leq 2Hd(6B)^2 r^{\frac{2d}{d+2}} n^{-\frac{2}{d+2}} \log(nr^2)^{\frac{2}{d+2}}.$$

G.3. Analysis of Algorithm 5: Multi-Stage Analysis

In this section, we show that, with careful choice of input parameters (c_1, r_1, k_1, T_1) and updating rule, algorithm 5 returns some minimax estimator $\hat{x}$. In essence, algorithm 5 recursively uses algorithm 4 to build smaller confidence region of the optimum $x_f^\star$ through iterations. Indeed, an important message from proposition 25 shows that, given any $\delta > 0$, with appropriate choice of parameters, one can find some rectangular W such that $x_f^\star$ lies inside the rectangular with probability at least $1 - \delta$. This means that, after one round, one can 'localize' the search of the optimum $x_f^\star$ by searching the optimum of f inside W. Now, treating W as the original $\mathbf{D}$, one can thus get an improved rate of convergence in the second round. Finally, we note that such 'localized' search can be recursively applied in all rounds from the first to the last round.

To be clear about how we specify the updating rule in line 3 of the algorithm 5, we summarize it as follows: given the output parameters $(\hat{c}_i, \hat{r}_i)$, we update $(c_{i+1}, r_{i+1}, k_{i+1}, T_{i+1})$ via algorithm 6. The next proposition shows that with appropriate choice of the initial parameter, we have nice convergence guarantees for the output of algorithm 5.

Algorithm 6 Updating Rule in Algorithm 5

Input: $\hat{c}_i \in \mathbb{R}^d$ and $\hat{r}_i \in \mathbb{R}_+$.
(i) Update c_{i+1} coordinate-wisely via:

$$c_{i+1,j} = \min\{1 - \hat{r}_i, \max\{r_i, \hat{c}_{i,j}\}\},$$

where $c_{i,j}$ and $c_{i+1,j}$ denote the jth coordinate of c_i and c_{i+1}.
(ii) Update r_{i+1} as $r_{i+1} = \hat{r}_i$.
(iii) Update k_{i+1} to be the largest $k \in \mathbb{N}$ such that

$$(2k+1)^d \left\lceil 2k^2 \log(2k+1) \right\rceil \leq n r_i^2. \tag{107}$$

If no such k exists, return FAIL.
(iv)Update T_{i+1} to be $T_{i+1} = \left\lfloor \frac{n}{(2k_i+1)^d} \right\rfloor$.

Proposition 33 *Let $\mathbf{D} = [0,1]^d$, and we are given the confidence level $\delta > 0$. We initialize the initial parameters as follows: set $c_1 = \frac{1}{2} \cdot \mathbf{1}$, $r_1 = \frac{1}{2}$, and set k_1 to be the largest $k \in \mathbb{N}$ (if exists) such that, $(2k+1)^d \left\lceil 2k^2 \log(2k+1) \right\rceil \leq n r_1^2$ and $T_1 = \left\lfloor \frac{n}{(2k_1+1)^d} \right\rfloor$. Consider algorithm 5 that uses algorithm 6 to be the updating rule. Denote $\hat{x}_M$ and $r_{M+1} = \hat{r}_M$ to be the output of estimate and radius from algorithm 5. Now, assuming that the $\{k_i\}_{i=1}^M$ exist for Eq (107) and satisfy the following lower bounds:*

$$k_i > \frac{10}{\lambda} \left(\sigma \left(\log \frac{M}{\delta} + 3d \right)^{\frac{1}{2}} + H d^{\frac{1}{2}} \right) \text{ for all } 1 \leq i \leq M. \tag{108}$$

Then, we have,

$$\mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq 2Hd r_{M+1}^2 \right) \geq 1 - \delta.$$

Proof Note that, when $M = 1$, proposition 33 reduces to proposition 25. When $M > 1$, note that, the definition of k_1 and T_1 takes the same form as that in Eq (89). In addition, the condition of k_1 in

Eq (108) changes by substituting δ by δ/M. Hence, denoting $W_1' = \{x \in \mathbf{D} : \|x - \hat{c}_1\|_\infty \leq \hat{r}_1\}$, Proposition 25 gives $\mathbb{P}(x_f^\star \in W_1') \geq 1 - \delta/M$. Now, denote similarly $W_1 = \{x : \|x - c_2\|_\infty \leq r_2\}$. Then, by definition of c_2 and r_2, $W_1 \subseteq W_1' \subseteq \mathbf{D}$, and hence $\mathbb{P}(x_f^\star \in W_1) \geq \mathbb{P}(x_f^\star \in W_1') \geq 1 - \delta/M$.

Now, in the second round of sampling and estimation, the algorithm essentially view W_1 as the entire domain $\mathbf{D}$ and sample the points using the same strategy as that in the first round. As the noise vectors in the second round is independent of the first one, Proposition 25 gives that $\mathbb{P}\left(x_f^\star \in W_2 \mid x_f^\star \in W_1\right) \geq 1 - \delta/M$, where $W_2 = \{x : \|x - c_3\|_\infty \leq r_3\}$. Indeed, the same conclusion holds for the ith round: denoting W_i analogously to be $W_i = \{x : \|x - c_{i+1}\| \leq r_{i+1}\}$, we get, $\mathbb{P}\left(x_f^\star \in W_{i+1} \mid x_f^\star \in W_i\right) \geq 1 - \delta/M$ for all $1 \leq i \leq M-1$. Thus, denoting via convection that $W_0 = \mathbf{D}$, we get,

$$\mathbb{P}(x_f^\star \in W_M) \geq \Pi_{i=1}^M \left[\mathbb{P}\left(x_f^\star \in W_i \mid x_f^\star \in W_{i-1}\right)\right] \geq (1 - \delta/M)^M \geq 1 - \delta.$$

Now, we introduce the notation $\Gamma = \{x_f^\star \in W_M\}$. Then, $\mathbb{P}(\Gamma) \geq 1 - \delta$. Now the requirements on k in Eq (108) guarantees that the size of W_i is strictly decreasing when $1 \leq i \leq M$. Hence, we can prove similarly to lemma 30 to get that, on event Γ, $f(\hat{x}_M) - f(x_f^\star) \leq 2Hd\hat{r}_m^2$ for our careful choice of the estimator $\hat{x}_M \in W_M'$. ■

Motivated by Proposition 33, it becomes important to understand when $k_i \in \mathbb{N}$ exists to satisfy both Eq (107) and Eq (108).

Lemma 34 *Assume n is large enough satisfying the bounds below:*

$$\log\log n > M \log\left(1 + \frac{2}{d}\right) + \log\left(2M \log(6B \log n) + (2d+5)\log(6B)\right), \tag{109}$$

where we denote B to be $B = \frac{10}{\lambda}\left(\sigma\left(\log\frac{M}{\delta} + 3d\right)^{\frac{1}{2}} + Hd^{\frac{1}{2}}\right)$. Then, the sequence $\{r_i\}_{i=1}^M$, $\{k_i\}_{i=1}^M$ and $\{T_i\}_{i=1}^M$ are well defined via algorithm 6. In addition, the sequence $\{k_i\}_{i=1}^M$ satisfy Eq (108). Finally, the output $\{r_i\}_{i=1}^M$ satisfy the bound below:

$$r_i \leq (6B)^{\frac{d+2}{2}\left(1-\left(\frac{d}{d+2}\right)^M\right)} D_n^{\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^M\right)} n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^M\right)} \text{ for all } 1 \leq i \leq M.$$

Proof We prove the desired claim of the lemma via induction. Our strategy is to show via induction that the below hypothesis hold for all $1 \leq i \leq M$:

$$(i) nr_i^2 > (6B)^{2(d+2)} \; (ii) k_i \text{ is well defined and } k_i > B \tag{110}$$

$$(iii)\; r_i \leq \min\left\{1, (6B)^{\frac{d+2}{2}\left(1-\left(\frac{d}{d+2}\right)^{i-1}\right)} (\log n)^{\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^{i-1}\right)} n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^{i-1}\right)}\right\}. \tag{111}$$

We first show the base case $i = 1$. Note that, the first part of Eq (110) is implied by the assumption on n, the second part of Eq (110) follows from the first part and corollary 32, and Eq (111) is trivial when $i = 1$. Now, for some $i < M$, assuming that the induction hypothesis holds for all $j \in \{1, 2, \ldots, i\}$, we show that the hypothesis holds for $i + 1$. We first show the second part of Eq (110) for $i + 1$. Indeed, by induction hypothesis, we know that, the first inequality of Eq (110)

is true for i, and thus the second part of Eq (110) follows from a direct application of corollary 32 (by substituting δ by δ/M and r by r_i there). Next, we show Eq (111) holds for $i+1$. Again, we know from corollary 32 that $r_{i+1} \leq r_i \leq 1$ and the bound that $r_{i+1} \leq (6B) r_i^{\frac{d}{d+2}} n^{-\frac{1}{d+2}} (\log n)^{\frac{1}{d+2}}$. Now, using the induction hypothesis of the upper bound on r_i, we get,

$$\begin{aligned} r_{i+1} &\leq 6B \left((6B)^{\frac{d+2}{2}\left(1-\left(\frac{d}{d+2}\right)^{i-1}\right)} (\log n)^{\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^{i-1}\right)} n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^{i-1}\right)} \right)^{\frac{d}{d+2}} (\log n)^{\frac{1}{d+2}} n^{-\frac{1}{d+2}} \\ &= (6B)^{\frac{d+2}{2}\left(1-\left(\frac{d}{d+2}\right)^{i}\right)} (\log n)^{\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^{i}\right)} n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+2}\right)^{i}\right)}, \end{aligned}$$

so we have shown Eq (111) for $i+1$. Finally, we show the first part of Eq (110) for $i+1$. Note that, by definition of $\{r_i\}_{i=1}^M$, we know that,

$$r_{i+1} = r_1 \cdot \Pi_{j=1}^{i} \frac{r_{j+1}}{r_j} = r_1 \cdot B^i \cdot \frac{1}{\Pi_{j=1}^{i} k_j} = \frac{B^i}{2\Pi_{j=1}^{i} k_j}.$$

Hence, the first part of Eq (119) for $i+1$ is equivalent to

$$nB^{2i} > 4(6B)^{2(d+2)} \cdot \Pi_{j=1}^{i} k_j^2. \tag{112}$$

To establish Eq (121), note first that, by definition of k_j, we have, for all $j \leq i$,

$$k_j \leq \frac{1}{3}(\gamma(nr_j^2))^{\frac{1}{d+2}} \leq \frac{1}{3}(nr_j^2)^{\frac{1}{d+2}}.$$

Now, using Eq (121) from the induction hypothesis for $j \leq i$, we get that,

$$k_j \leq (nr_j^2)^{\frac{1}{d+2}} \leq (6B)^{\left(1-\left(\frac{d}{d+2}\right)^{i}\right)} n^{\frac{1}{d+2}\left(\frac{d}{d+2}\right)^{j-1}} (\log n)^{\frac{1}{d+2}\left(1-\left(\frac{d}{d+2}\right)^{i}\right)} \leq 6B n^{\frac{1}{d+2}\left(\frac{d}{d+2}\right)^{j-1}} (\log n).$$

Hence, to prove Eq (121), it suffices to show that, for all $1 \leq i \leq M$,

$$nB^{2i} > 4(6B)^{2(d+2)} (6B \log n)^{2i} n^{\frac{2}{d+2}\sum_{j=0}^{i-1}\left(\frac{d}{d+2}\right)^{j}} = 4(6B)^{2(d+2)} (6B \log n)^{2i} n^{1-\left(\frac{d}{d+2}\right)^{i}}.$$

Note that, it suffices if n is large enough satisfying the bound below,

$$n^{\left(\frac{d}{d+2}\right)^{M}} > 4(6B)^{2(d+2)} (6B \log n)^{2M},$$

which would suffice is n satisfies

$$\log \log n > M \log\left(1 + \frac{2}{d}\right) + \log\left(2M \log(6B \log n) + (2d+5)\log(6B)\right).$$

■

Now, Proposition 33 and Lemma 34 together immediately give the corollary below.

Corollary 35 *Let $\mathbf{D} = [0,1]^d$. Consider algorithm 5. Suppose we use the same initialization rule as that in Proposition 33 and use Algorithm 6 to be the updating rules for algorithm 5, then, when n is large enough so that it satisfies Eq (109), then the output $\hat{x}_M$ from algorithm 5 satisfies*

$$\mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq \gamma_M^\star\right) \geq 1 - \delta$$

with

$$\gamma_M^\star := 2Hd(6B)^{(d+2)\left(1-\left(\frac{d}{d+2}\right)^{M}\right)} (\log n)^{\left(1-\left(\frac{d}{d+2}\right)^{M}\right)} n^{-\left(1-\left(\frac{d}{d+2}\right)^{M}\right)}.$$

Appendix H. Smooth Function with Zeroth Order Oracle

H.1. Description of Algorithms

In this section, we propose two generic algorithms: algorithm 7 parameterized by parameters $(c, r, k, T) \in \mathbf{D} \times \mathbb{R}_+ \times \mathbb{N} \times \mathbb{N}$ and algorithm 8 that builds from algorithm 7. We note here that, algorithm 8 in essence, builds from M times of repeated calls of algorithm 7. As will be shown immediately in the later subsections, it turns out that the two algorithms with careful choice of parameters return the minimax estimator in single and multi rounds respectively.

Algorithm 7 Generic Routine for One Stage Smooth Functions $\mathcal{F}_{H,\lambda}$ (Zeroth-order Oracle)

Input: Prior knowledge on $\lambda, H \in \mathbb{R}_+$ satisfying $\lambda \leq H$ and the noise level $\sigma \in \mathbb{R}_+$. User specifies the sampling center $c \in \mathbf{D}$, radius $r \in \mathbb{R}_+$, grid size parameter $k \in \mathbb{N}$, the sampling times $T \in \mathbb{N}$ and the confidence level $\delta \in (0, 1]$.

1: Compute the grid points $G = G(c, r, k)$.
2: At each point $x \in G$, query the first oracle T times and denote each sample function value via $\{\widehat{f}(x)^{(1)}, \widehat{f}(x)^{(2)}, \ldots, \widehat{f}(x)^{(T)}\}$.
3: Compute the function value estimate at each point $x \in G$ via $\widehat{f}(x) := \frac{1}{T}\sum_{i=1}^{T} \widehat{f}(x)^{(i)}$.
4: Compute the estimate $\hat{x} \in G$, defined as, $\hat{x} := \operatorname{argmin}_{x \in G} \widehat{f}(x)$.
5: Return the estimator $\hat{x}$ and the confidence radius $\hat{r} = \frac{r}{k} \cdot \left(\frac{2\sigma}{\lambda}\left(\log\frac{2}{\delta} + d\right)^{\frac{1}{2}} + \frac{H}{\lambda}d\right)^{\frac{1}{2}}$.

Algorithm 8 Generic Routine for Multi-stage Smooth Functions $\mathcal{F}_{H,\lambda}$ (Zeroth-order Oracle)

Input: Prior knowledge on $\lambda, H \in \mathbb{R}_+$ satisfying $\lambda \leq H$ and the noise level $\sigma \in \mathbb{R}_+$ and number of rounds $R \in \mathbb{N}_+$. Initialization of parameters $(c_1, r_1, k_1, T_1) \in \mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{N} \times \mathbb{N}$. User specifies the confidence level $\delta \in (0, 1]$ and the updating rule that used in line (3) of the algorithm.

1: **for** $i = 1$ to M **do**
2: Run algorithm 4 with input parameters (c_i, r_i, k_i, T_i). Denote the output to be $\hat{x}_i$, confidence radius $\hat{r}_i$.
3: Update $(c_{i+1}, r_{i+1}, k_{i+1}, T_{i+1})$. The updating rule may take $\hat{x}_i$ and $\hat{r}_i$ as input.
4: **end for**
5: **return** $\hat{x}_M$ as estimate of $x_f^\star$ and the radius r_{M+1}.

H.2. Analysis of Algorithm 7: Single-Stage Analysis

In this section, we show that a single call of algorithm 7 with careful choice of input parameters (c, r, k, T) returns some estimator $\hat{x}$ that is minimax optimal. To serve for the purpose for latter discussion on multi-stage algorithm, in this section, we slightly generalize the domain of interest $\mathbf{D}_{c,r} = \{x \in \mathbb{R}^d : \|x - c\|_\infty \leq r\}$ and denote $x^\star_{f,c,r}$ the unique minimum of f on domain $\mathbf{D}_{c,r}$. We consider finding the minimax estimator $\hat{x} \in \mathbf{D}_{c,r}$ for $x^\star_{f,c,r}$ evaluated by $f(\hat{x}) - f(x^\star_{f,c,r})$. Substituting $c = \frac{1}{2} \cdot 1$ and $r = \frac{1}{2}$ gives the result for the single-stage algorithm for the original domain $\mathbf{D} = [0, 1]^d$.

Proposition 36 *Given any fix $c \in \mathbb{R}^d$ and $r \in (0,1]$, suppose there exists some $k \in \mathbb{N}$ satisfying*

$$(2k+1)^d \left\lfloor 2k^4 \log(2k+1) \right\rfloor \leq nr^4 \text{ and } k \geq \left(\frac{2\sigma}{\lambda} \left(\log \frac{2}{\delta} + d \right)^{\frac{1}{2}} + \frac{H}{\lambda} d \right)^{\frac{1}{2}}. \tag{113}$$

Set $T = \left\lfloor \frac{n}{(2k+1)^d} \right\rfloor$. Then, if we denote $\hat{r}$ and $\hat{x}$ to be the output from algorithm 7 when inputting (c, k, T, r) as above, then, we have $\hat{r} < r$ and

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,c,r}) \leq \frac{1}{2}\lambda \hat{r}^2 \text{ and } \left\| x^\star_{f,c,r} - \hat{x} \right\|_\infty \leq \hat{r} \right) \geq 1 - \delta.$$

Remark 37 *Before we give the proof of proposition 36, we give some high-level intuitions why algorithm 7* should *work. As shown previously in lemma 28, we know that there always exists some point $\bar{x} \in G$ satisfying*

$$\left\| \bar{x} - x^\star_{f,c,r} \right\|_\infty \leq \frac{r}{k} \text{ and } \left\langle \nabla f(x^\star_{f,c,r}), \bar{x} - x^\star_{f,c,r} \right\rangle = 0.$$

It turns out that such $\bar{x}$ also has function value $f(\bar{x})$ close to $f(x^\star_{f,c,r})$ up to $O(\hat{r}^2)$. Since $\hat{x}$ is defined to be the smallest point in grid that minimizes $\hat{f}$, intuitively it makes sense that $f(\hat{x})$ should also be as good as $f(x^\star_{f,c,r})$ when the sample size n is large enough.

Proof We first recall lemma 28 in the proof of proposition 25.

Lemma 38 *There exists some $\bar{x} \in G$ satisfying the below conditions:*

$$\left\| \bar{x} - x^\star_{f,c,r} \right\|_\infty \leq \frac{r}{k} \text{ and } \left\langle \nabla f(x^\star_{f,c,r}), \bar{x} - x^\star_{f,c,r} \right\rangle = 0. \tag{114}$$

Now, take any point $\bar{x} \in G$ satisfying Eq (114). Note first that, by smoothness assumption of the objective function f, we have,

$$f(\bar{x}) - f(x^\star_{f,c,r}) \leq \left\langle \nabla f(x^\star_{f,c,r})), \bar{x} - x^\star_{f,c,r} \right\rangle + \frac{H}{2} \left\| \bar{x} - x^\star_{f,c,r} \right\|_2^2 \leq \frac{Hr^2 d}{2k^2}.$$

Now, let us consider the following event:

$$\Gamma = \left\{ \left| \widehat{f}(x) - f(x) \right| \leq r^a := \sigma \sqrt{\frac{2}{T} \log \frac{2(2k+1)^d}{\delta}} \text{ for all } x \in G \right\},$$

The next lemma shows that Γ happens with probability at least $1 - \delta$.

Lemma 39 *We have $\mathbb{P}(\Gamma) \geq 1 - \delta$.*

Proof First, for each $x \in G$, denote $\epsilon(x) := \widehat{f}(x) - f(x)$. Then, since by our assumption, the noise $\{\widehat{f}(x) - f(x)\}_{x=1}^T$ is mean 0, independent and subgaussian with parameter σ^2, we have that $\epsilon(x)$ is mean 0 and subgaussian with parameter σ^2/T. Therefore, for any fix $x \in G$,

$$\mathbb{P}\left(|\epsilon(x)| \geq r^a \right) \leq 2 \exp\left(-\frac{(r^a)^2 T}{2\sigma^2} \right) \leq \delta (2k+1)^{-d}.,$$

where the first inequality above uses the subgaussianity of $\epsilon(x)$, and the second inequality uses the definition of k and T. Now, the desired claim of the lemma follows from the fact that $|G| = (2k+1)^d$ and the union bound of the above events. ■

Since the assumption on k, T shows that $T \geq 2k^4 \log(2k+1) r^{-4}$, and therefore, we have,

$$r^a = \sigma \left(\frac{2 \log \frac{2}{\delta} + 2d \log(2k+1)}{T} \right)^{\frac{1}{2}} \leq \frac{\sigma r^2}{k^2} \left(\log \frac{2}{\delta} + d \right)^{\frac{1}{2}},$$

this shows that, on Γ, for all $x \in \mathbf{D}_{c,r}$ such that $\left\| x - x^\star_{f,c,r} \right\|_2 \geq \hat{r}$, we have,

$$f(x) - f(x^\star_{f,c,r}) \geq \left\langle \nabla f(x^\star_{f,c,r})), x - x^\star_{f,c,r} \right\rangle + \frac{\lambda}{2} \left\| x - x^\star_{f,c,r} \right\|_2^2 \geq \frac{\lambda}{2} \hat{r}^2 \geq \frac{Hr^2 d}{2k^2} + 2r^a.$$

Hence, on event Γ, we have, for all $x \in \mathbf{D}$ satisfying $\left\| x - x^\star_{f,c,r} \right\|_2 \geq \hat{r}$, we have,

$$\hat{f}(x) - \hat{f}(\bar{x}) = \underbrace{\left(\hat{f}(x) - f(x) \right)}_{\geq -r^a} + \underbrace{\left(f(x) - f(x^\star_{f,c,r}) \right)}_{\geq \frac{Hr^2 d}{2k^2} + 2r^a} + \underbrace{\left(f(x^\star_{f,c,r}) - f(\bar{x}) \right)}_{\geq -\frac{Hr^2 d}{2k^2}} + \underbrace{\left(f(\bar{x}) - \hat{f}(\bar{x}) \right)}_{\geq -r^a} > 0,$$

which gives us that on event Γ, we must have

$$\left\| \hat{x} - x^\star_{f,c,r} \right\|_2 \leq \hat{r} \quad \Leftrightarrow \quad x^\star_{f,c,r} \in W.$$

Finally, since always $\hat{f}(\hat{x}) \leq \hat{f}(\bar{x})$, on Γ, we have below upper bound on $f(\hat{x})$ on event Γ:

$$f(\hat{x}) - f(x^\star_{f,c,r}) = \underbrace{f(\hat{x}) - \hat{f}(\hat{x})}_{\leq r^a} + \underbrace{\hat{f}(\hat{x}) - \hat{f}(\bar{x})}_{\leq 0} + \underbrace{\hat{f}(\bar{x}) - f(\bar{x})}_{\leq r^a} + \underbrace{f(\bar{x}) - f(x^\star_{f,c,r}))}_{\leq \frac{Hr^2 d}{2k^2}} \leq \frac{Hr^2 d}{2k^2} + 2r^a.$$

The desired claim of the proposition follows from $\frac{Hr^2 d}{2k^2} + 2r^a \leq \frac{\lambda}{2} \hat{r}^2$. ■

Motivated by Proposition 36, it becomes important to understand when such k exists in Eq (113) and how large it is.

Lemma 40 *Assume n is large enough satisfying*

$$nr^4 \geq (6B)^{2(d+4)}, \quad \textit{where } B = 12 \left(\frac{2\sigma}{\lambda} \left(\log \frac{2}{\delta} + d \right)^{\frac{1}{2}} + \frac{H}{\lambda} d \right)^{\frac{1}{2}}. \tag{115}$$

Denote $k(r) = \left(\gamma \left(nr^4 \right) \right)^{\frac{1}{d+4}}$ and $k^\star = \left\lfloor \frac{1}{3} k(r) \right\rfloor$. Then $k^\star \in \mathbb{N}$, and $k^\star$ satisfies Eq (113).

Proof Note that, $\gamma(x) \geq \sqrt{x}$ whenever $x \geq 3$. Thus, by our assumption on n, we get that,

$$\left(\gamma(nr^4) \right)^{\frac{1}{d+4}} \geq 6B \geq 6.$$

This immediately gives us that $k^\star \geq 1$ and $k^\star$ satisfies the second inequality of Eq (113). Now, we show that $k^\star$ satisfies the first inequality of Eq (113). In fact, when $k = k^\star$, we have,

$$(2k+1)^d \left\lceil 2k^4 \log(2k+1) \right\rceil \leq (3k)^{d+4} \log(3k)^{d+4} \leq (k(r))^{d+4} \log(k(r))^{d+4} \leq nr^4,$$

where the last inequality follows from the fact that, for any $x > 0$, $\gamma(x) \log \gamma(x) \leq x$. ■

Proposition 36 and Lemma 40 together immediately give us the corollary below.

Corollary 41 *Given any fix $c \in \mathbb{R}^d$ and $r \in [0,1]$, set $k = \left\lfloor \frac{1}{3}\gamma(nr^4)^{\frac{1}{d+4}} \right\rfloor$ and $T = \left\lfloor \frac{n}{2k+1^d} \right\rfloor$. Assume n is large enough satisfying Eq (115). Then, if we denote $\hat{r}$ and $\hat{x}$ to be the output of of Algorithm 7 when we input (c, k, T, r) as the input parameters, we have,*

$$\hat{r} \leq \min\{r, 6Br^{\frac{d}{d+4}} n^{-\frac{1}{d+4}} \log(nr^4)^{\frac{1}{d+4}}\}, \text{ where } B = \left(\frac{2\sigma}{\lambda} \left(\log \frac{2}{\delta} + d \right)^{\frac{1}{2}} + \frac{H}{\lambda} d \right)^{\frac{1}{2}}.$$

In addition, we get that,

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,c,r}) \leq \gamma^\star \right) \geq 1 - \delta,$$

where

$$\gamma^\star = \frac{1}{2}\lambda \hat{r}^2 \leq (6B)^2 r^{\frac{2d}{d+4}} n^{-\frac{2}{d+4}} \log(nr^4)^{\frac{2}{d+4}}.$$

H.3. Analysis of Algorithm 8: Multi-Stage Analysis

In this section, we show that with careful choice of input parameters (c_1, r_1, k_1, T_1) and the updating rule, algorithm 8 returns some estimator $\hat{x}$ that is minimax optimal. To do so, our strategy is similar to the strategy used in algorithm 5. In fact, an important message from proposition 36 shows that, given any $\delta > 0$, with appropriate choice of parameters, one can find some rectangular W such that $x^\star_f$ lies inside the rectangular with probability at least $1 - \delta$. This means that, after one round, one can 'localize' the search of the optimum $x^\star_f$ by searching the optimum of f inside W. Now, treating W as the original $\mathbf{D}$, one can thus get an improved rate of convergence in the second round.

To be clear about how we specify the updating rule in line 3 of the algorithm 5, we summarize it as follows: given the output parameters $(\hat{c}_i, \hat{r}_i)$, we update $(c_{i+1}, r_{i+1}, k_{i+1}, T_{i+1})$ via algorithm 6. The next proposition shows that with appropriate choice of the initial parameter, we have nice convergence guarantees for the output of algorithm 5.

Below proposition quantifies such 'localization' idea and shows us that an appropriate choice of the initial parameter (c_1, r_1, k_1, T_1) and appropriate choice of the updating rule gives back some estimator that optimal minimax rate.

Proposition 42 *Let $\mathbf{D} = [0,1]^d$, and we are given the confidence level $\delta > 0$. We initialize the initial parameters as follows: set $c_1 = \frac{1}{2} \cdot \mathbf{1}$, $r_1 = \frac{1}{2}$, and set k_1 to be the largest $k \in \mathbb{N}$ (if exists) such that, $(2k+1)^d \left\lceil 2k^4 \log(2k+1) \right\rceil \leq nr_1^4$ and $T_1 = \left\lfloor \frac{n}{(2k_1+1)^d} \right\rfloor$. Consider algorithm 8 that uses algorithm 9 to be the updating rule. Denote $\hat{x}_M$ and $r_{M+1} = \hat{r}_M$ to be the output of estimate and radius from algorithm 5. Now, assuming that the $\{k_i\}_{i=1}^M$ exist for Eq (107) and satisfy the following lower bounds:*

$$k_i \geq \left(\frac{2\sigma}{\lambda} \left(\log \frac{2M}{\delta} + d \right)^{\frac{1}{2}} + \frac{H}{\lambda} d \right)^{\frac{1}{2}} \text{. for all } 1 \leq i \leq M. \tag{117}$$

Algorithm 9 Updating Rule in Algorithm 8

Input: $\hat{x}_i \in \mathbb{R}^d$ and $\hat{r}_i \in \mathbb{R}_+$.
(i) Update c_{i+1} coordinate-wisely via:

$$c_{i+1,j} = \min\{1 - \hat{r}_i, \max\{r_i, \hat{x}_{i,j}\}\},$$

where $\hat{x}_{i,j}$ and $c_{i+1,j}$ denote the jth coordinate of $\hat{x}_i$ and c_{i+1}.
(ii) Update r_{i+1} as $r_{i+1} = \hat{r}_i$.
(iii) Update k_{i+1} to be the largest $k \in \mathbb{N}$ such that

$$(2k+1)^d \left\lceil 2k^4 \log(2k+1) \right\rceil \leq n r_i^4. \tag{116}$$

If no such k exists, return FAIL.
(iv)Update T_{i+1} to be $T_{i+1} = \left\lfloor \frac{n}{(2k_i+1)^d} \right\rfloor$.

Then, we have,

$$\mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq \frac{1}{2}\lambda r_{M+1}^2 \right) \geq 1 - \delta.$$

Proof Note that, when $M = 1$, proposition 42 reduces to proposition 36. When $M > 1$, note that, the definition of k_1 and T_1 take the same form as that in Eq (113) However, the condition of k_1 in Eq (117) changes by substituting δ by δ/M. Denote $W_1' = \{x \in \mathbf{D} : \|x - \hat{x}_1\|_\infty \leq \hat{r}_1\}$, then proposition 36 asserts that $\mathbb{P}(x_f^\star \in W_1) \geq 1 - \delta/M$. Now, denote $W_1 = \{x : \|x - x_2\|_\infty \leq r_2\}$. Then $W_1 \subseteq W_1' \subseteq \mathbf{D}$ by definition of $\hat{x}_1'$ and the fact that $\hat{r}_1 \leq r$. Thus, we have $\mathbb{P}(x_f^\star \in W_1) \geq \mathbb{P}(x_f^\star \in W_1') \geq 1 - \delta/M$. Now, in the second round of sampling and estimation, the algorithm essentially view W_1 as the entire domain $\mathbf{D}$ and sample the points using the same strategy as that in the first round. Since the noise in the second round is independent of the noise in the first round, by proposition 36 that $\mathbb{P}\left(x_f^\star \in W_2 \mid x_f^\star \in W_1\right) \geq 1 - \delta/M$, where we define analogously $W_2 = \{x : \|x - x_3\|_\infty \leq r_3\}$. The same reasoning applies to any round, which shows that, we have $\mathbb{P}\left(x_f^\star \in W_{i+1} \mid x_f^\star \in W_i\right) \geq 1 - \delta/M$, where we define W_i analogously by $W_i = \{x : \|x - x_{i+1}\|_\infty \leq r_{i+1}\}$. In addition to that, in the Mth round, with the same reasoning as proposition 36, one can easily show that $\mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq \frac{\lambda}{2} r_{M+1}^2 \mid x_f^\star \in W_{M-1}\right) \geq 1 - \delta/M$. Thus, if we denote via convection that $W_0 = \mathbf{D}$, then we have,

$$\begin{aligned}
&\mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq \frac{\lambda}{2} r_{M+1}^2 \right) \\
&= \mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq \frac{\lambda}{2} r_{M+1}^2 \mid x_f^\star \in W_{M-1} \right) \cdot \Pi_{i=1}^{M-1} \left[\mathbb{P}\left(x_f^\star \in W_i \mid x_f^\star \in W_{i-1}\right) \right] \\
&\geq (1 - \delta/M)^M \geq 1 - \delta,
\end{aligned}$$

which gives the desired claim of the proposition. ■

Motivated by Proposition 42, it becomes important to understand when $k_i \in \mathbb{N}$ exists to satisfy both Eq (116) and Eq (117).

Lemma 43 *Assume n is large enough satisfying the bounds below:*

$$\log\log n \geq M \log\left(1+\frac{4}{d}\right) + \log\left(2M\log(12B\log n) + (2d+6)\log(6B)\right). \tag{118}$$

where we denote B to be $B = \left(\frac{2\sigma}{\lambda}\left(\log\frac{2}{\delta}+d\right)^{\frac{1}{2}} + \frac{H}{\lambda}d\right)^{\frac{1}{2}}$. Then, the sequence $\{r_i\}_{i=1}^M$, $\{k_i\}_{i=1}^M$ and $\{T_i\}_{i=1}^M$ are well defined via algorithm 9. In addition, the sequence $\{k_i\}_{i=1}^M$ satisfy Eq (117). Finally, the output $\{r_i\}_{i=1}^M$ satisfy the bound below:

$$r_i \leq (6B)^{\frac{d+4}{4}\left(1-\left(\frac{d}{d+4}\right)^M\right)} D_n^{\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^M\right)} n^{-\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^M\right)} \text{ for all } 1 \leq i \leq M.$$

Proof We prove the desired claim of the corollary via induction. Our strategy is to show via induction that the below hypothesis hold for all $1 \leq i \leq M$.

$$(i) nr_i^4 \geq (6B)^{2(d+4)} \; (ii) k_i \in \mathbb{N} \text{ and } k_i \geq B \tag{119}$$

$$(iii)\; r_i \leq \min\left\{1, (6B)^{\frac{d+4}{4}\left(1-\left(\frac{d}{d+4}\right)^{i-1}\right)} (\log n)^{\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^{i-1}\right)} n^{-\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^{i-1}\right)}\right\}. \tag{120}$$

We first show the base case $i = 1$. Note that, the first part of Eq (119) is implied by the assumption on n, the second part of Eq (119) follows from the first part and corollary 41, and Eq (120) is trivial when $i = 1$. Now, for some $i < M$, assuming that the induction hypothesis holds for all $j \in \{1, 2, \ldots, i\}$, we show that the hypothesis holds for $i+1$. We first show the second part of Eq (119) for $i+1$. Indeed, by induction hypothesis, we know that, the first inequality of Eq (119) is true for i, and thus the second part of Eq (119) follows from a direct application of corollary 41 (by substituting δ by δ/M and r by r_i there). Next, we show Eq (120) holds for $i+1$. Again, we know from corollary 41 that $r_{i+1} \leq r_i \leq 1$, and the bound that, $r_{i+1} \leq 6Br_i^{\frac{d}{d+4}} n^{-\frac{1}{d+4}} (\log n)^{\frac{1}{d+4}}$. Now, using the induction hypothesis of the upper bound on r_i, we get,

$$\begin{aligned} r_{i+1} &\leq 6B\left((6B)^{\frac{d+4}{4}\left(1-\left(\frac{d}{d+4}\right)^{i-1}\right)} (\log n)^{\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^{i-1}\right)} n^{-\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^{i-1}\right)}\right)^{\frac{d}{d+4}} (\log n)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \\ &= (6B)^{\frac{d+4}{4}\left(1-\left(\frac{d}{d+4}\right)^{i}\right)} (\log n)^{\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^{i}\right)} n^{-\frac{1}{4}\left(1-\left(\frac{d}{d+4}\right)^{i}\right)}, \end{aligned}$$

so we have shown the Eq (120) for $i+1$. Finally, we show the first part of Eq (119) for $i+1$. Note that, by definition of $\{r_i\}_{i=1}^M$,

$$r_{i+1} = r_1 \cdot \Pi_{j=1}^i \frac{r_{j+1}}{r_j} = r_1 \cdot B^i \cdot \frac{1}{\Pi_{j=1}^i k_j} = \frac{B^i}{2\Pi_{j=1}^i k_j}.$$

Hence, the first part of Eq (119) for $i+1$ is equivalent to

$$nB^i \geq 16 \cdot (6B)^{2(d+4)} \cdot \Pi_{j=1}^i k_j^4. \tag{121}$$

To establish Eq (121), note first that, by definition of k_j, we have, for all $j \leq i$,

$$k_j \leq \frac{1}{3}(\gamma(nr_j^4))^{\frac{1}{d+4}} \leq \frac{1}{3}(nr_j^4)^{\frac{1}{d+4}}.$$

Now, using Eq (121) from the induction hypothesis for $j \leq i$, we get that,

$$k_j \leq (nr_j^4)^{\frac{1}{d+4}} \leq (6B)^{\left(1-\left(\frac{d}{d+4}\right)^i\right)} n^{\frac{1}{d+4}\left(\frac{d}{d+4}\right)^{j-1}} (\log n)^{\frac{1}{d+4}\left(1-\left(\frac{d}{d+4}\right)^i\right)} \leq 6Bn^{\frac{1}{d+4}\left(\frac{d}{d+4}\right)^{j-1}} (\log n).$$

Hence, to prove Eq (121), it suffices to show that, for all $1 \leq i \leq M$,

$$nB^i \geq 16(6B)^{2(d+4)} (6B \log n)^{2i} n^{\frac{4}{d+4}\sum_{j=0}^{i-1}\left(\frac{d}{d+4}\right)^j} = 16(6B)^{2(d+4)} (6B \log n)^{2i} n^{1-\left(\frac{d}{d+4}\right)^i}.$$

Note that, it suffices if n is large enough satisfying the bound below,

$$n^{\left(\frac{d}{d+4}\right)^M} \geq 16(6B)^{2(d+4)} (6B \log n)^{2M},$$

which would suffice if

$$\log \log n \geq M \log\left(1 + \frac{4}{d}\right) + \log\left(2M \log(6B \log n) + (2d+6)\log(6B)\right).$$

■

Now, Proposition 42 and Lemma 43 together immediately give the corollary below.

Corollary 44 *Let $\mathbf{D} = [0,1]^d$. Consider algorithm 8. Suppose we use the same initialization rule as that in Proposition 42 and use Algorithm 9 to be the updating rules for algorithm 8, then, when n is large enough so that it satisfies Eq (118), then the output $\hat{x}_M$ from algorithm 8 satisfies*

$$\mathbb{P}\left(f(\hat{x}_M) - f(x_f^\star) \leq \gamma_M^\star\right) \geq 1 - \delta$$

with

$$\gamma_M^\star := \frac{1}{2}\lambda(6B)^{\frac{d+4}{2}\left(1-\left(\frac{d}{d+4}\right)^M\right)} (\log n)^{\frac{1}{2}\left(1-\left(\frac{d}{d+4}\right)^M\right)} n^{-\frac{1}{2}\left(1-\left(\frac{d}{d+4}\right)^M\right)}.$$

Appendix I. Lipschitz Function with First-Order Oracle

I.1. Lipschitz Function with First-Order Oracle, $d = 1$

I.1.1. DESCRIPTION OF ALGORITHM

In this section, we propose a generic algorithm: Algorithm 10 to solve the problem when the dimension $d = 1$. The algorithm is a one-round algorithm. As will be seen immediately, this one-round algorithm achieves the best possible statistical minimax rate (up to logarithmic factors and constants). Thus, under the first oracle situtation, *adaptivity gives no advantage* for optimization in one dimension.

Algorithm 10 Routine for One Stage Lipschitz Function $\mathcal{F}_\lambda$ (First-Order Oracle)

Input: User's choice of the sampling center $c \in \mathbf{D}$, radius $r \in \mathbb{R}_+$, resolution M, grid size parameter $\{k_i\}_{i=1}^M$, sampling times $\{T_i\}_{i=1}^M$ and the precision $\{\Delta_i\}_{i=1}^M$.

Initialize the interval $I = [l_1, l_2] = [c - r, c + r]$.
for $i = 1$ to K **do**
  Compute the grid $G_i = G(c, r, k_i) \cap I$.
  At each point $x \in G_i$, query the first-order oracle T_i times. Denote each sample derivative value via $\{\hat{f}'(x)^{(1)}, \hat{f}'(x)^{(2)}, \ldots, \hat{f}'(x)^{(T_i)}\}$.
  Compute the derivative estimate at each point $x \in G_i$ via averaging:

$$\hat{f}'(x) = \frac{1}{T_i} \sum_{i=1}^{T_i} \hat{f}'(x)^{(i)}$$

  if there exists $x_i^\star \in G_i$ such that $|\hat{f}'(x_i^\star)| \leq \Delta_i$ **then**
    Return the estimator $\hat{x} = x_i^\star$.
  else if for all $x \in G_i$, $\hat{f}'(x_i^\star) < -\Delta_i$ **then**
    Return the estimator $\hat{x} = r$
  else if for all $x \in G_i$, $\hat{f}'(x_i^\star) > \Delta_i$ **then**
    Return the estimator $\hat{x} = l$
  else update the interval $I = [l_1, l_2]$ via:

$$l_1 = \max_{x \in G_i} \{x \in I : \hat{f}'(x) < -\Delta_i\} \text{ and } l_2 = \min_{x \in G_i} \{x \in I : \hat{f}'(x) > \Delta_i\}.$$

  end if
end for
Return the estimator $\hat{x} = (l_1 + l_2)/2$.

I.1.2. ANALYSIS OF ALGORITHM 10

In this section, we show that, with careful choice of input parameters, algorithm 10 returns some estimator $\hat{x}$ so that its risk is upper bounded by $\tilde{O}\left(n^{-1/2}\right)$. As before, we slightly generalize the domain of interest to be $\mathbf{D}_{c,r} = [c - r, c + r]$. The target of interest now becomes $x^\star_{f,c,r}$, the unique minimum of f in the domain $\mathbf{D}_{c,r}$ and the risk of interest would be $\mathbb{E}\left[f(\hat{x}) - f(x^\star_{f,c,r})\right]$.

Proposition 45 *Given any fix $c \in \mathbb{R}$ and $r \in \mathbb{R}_+$. Let us choose the parameters*

$$k_i = 2^{i-1},\ M = \left\lfloor \frac{1}{2} \log_2 n \right\rfloor,\ T_i = \left\lfloor \frac{n}{M(2k_i + 1)} \right\rfloor \text{ and } \Delta_i = \sqrt{\frac{2\sigma^2}{T_i} \log \frac{2M(2k_i + 1)}{\delta}}. \quad (122)$$

If we denote $\hat{x}$ to be the output of the algorithm 10 with the above input parameters, then,

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,c,r}) \leq \max\{4 \max_{1 \leq j \leq M} \left\{\Delta_j k_j^{-1}\right\}, L k_M^{-1}\} r\right) \geq 1 - \delta.$$

Proof Now, let us consider the following event:

$$\Gamma = \bigcap_{i=1}^{M} \left\{ |\hat{f}'(x) - f'(x)| \leq \Delta_i \text{ for all } x \in G_i \right\}$$

The next lemma shows that Γ happens with probability at least $1 - \delta$.

Lemma 46 *We have* $\mathbb{P}(\Gamma) \geq 1 - \delta$.

Proof First, for each $x \in G_i$, denote $\epsilon(x) := \widehat{f}'(x) - f'(x)$. Then, since by our assumption, the noise $\{\widehat{f}'(x) - f'(x)\}_{x=1}^{T_i}$ is mean 0, independent and subgaussian with parameter σ^2, we have $\epsilon(x)$ is mean 0 and subgaussian with parameter σ^2/T_i. Therefore, for any fix $x \in G_i$,

$$\mathbb{P}\left(|\epsilon(x)| \geq \Delta_i\right) \leq 2\exp\left(-\frac{\Delta_i^2 T_i}{2\sigma^2}\right) \leq \delta M^{-1}(2k_i + 1)^{-d},$$

where the first inequality above uses the subgaussianity of $\epsilon(x)$, and the second inequality uses the definition of k and T. Now, since $|G_i| = (2k+1)^d$, by union bound, we have,

$$\mathbb{P}(\Gamma^c) \leq \sum_{i=1}^{M} \mathbb{P}(\exists x \in G_i : |\epsilon(x)| \geq \Delta_i) \leq \delta,$$

which gives the desired claim of the proposition. ■

Now, assuming in the rest of the proof that event Γ happens. We now take a look into the *first* iteration from line 2 to line 14. Note that, basically, from line 6 to line 14, we are checking whether there exist $x_1 \in G_1$ such that $\hat{f}'(x_1) < -\Delta_1$ or $x_2 \in G_2$ such that $\hat{f}'(x_2) > \Delta_2$. There are four different circumstances, and here we discuss them one by one:

1. Suppose there exist some $x_1^\star \in G_1$ such that $|\hat{f}'(x_1)| < \Delta_1$, then the algorithm terminates at line 7. In this case, we know that $|f'(x_1)| \leq |\hat{f}'(x)| + |f(x_1) - \hat{f}'(x_1)| \leq 2\Delta_1$. By convexity, this gives the following upper bound on $f(x_1^\star) - f(x_{f,c,r}^\star)$:

   $$f(x_1^\star) - f(x_{f,c,r}^\star) \leq f'(x_1^\star)(x_1^\star - x_{f,c,r}^\star) \leq |f'(x_1^\star)||x_1^\star - x_{f,c,r}^\star| \leq 4\Delta_1 k_1^{-1} r.$$

   In this case, the algorithm returns $x_1^\star$.

2. Suppose for all $x \in G_1$, $\hat{f}'(x) < -\Delta_1$, then the algorithm terminates at line 9. Hence, for all $x \in G_1$, we indeed have $f'(x) < 0$, and the function is monotonically decreasing on the interval $[l_1, l_2]$. In this case, we return r, the minimum of f.

3. Suppose for all $x \in G_1$, $\hat{f}'(x) > \Delta_1$, then the algorithm terminates at line 11. Hence, for all $x \in G_1$, we indeed have $f'(x) > 0$, and the function is monotonically increasing on the interval $[l_1, l_2]$. In this case, we return l, the minimum of f.

4. Lastly, suppose there exists some $x_1, x_2 \in G_1$ such that $\hat{f}'(x_1) < -\Delta_1$ and $\hat{f}'(x_2) > \Delta_1$. Thus in this case, $f'(x_1) < 0$ and $f'(x_2) > 0$. As $f$is a convex function on the interval $I = [l_1, l_2]$, its derivative f' can only flip the sign at most once on I. Our way of updating the interval, will make sure that the minimum $x_{f,c,r}^\star$ lie in the updated interval I. Note that, the length of the interval I now decreases to $k_1^{-1}r = 2k_2^{-1}r$ in the next for loop.

Now, following exactly the same reasoning as above, one can prove via induction that, in the jth loop, where $1 \leq j \leq M$, we have, $|I| = 2k_j^{-1}r$.

1. Suppose there exist some $x_j^\star \in G_1$ such that $|\hat{f}'(x_j)| < \Delta_j$. Then, the algorithm returns $x_j^\star$, which satisfies,
$$f(x_j^\star) - f(x_{f,c,r}^\star) \leq |f'(x_j^\star)||x_j^\star - x_{f,c,r}^\star| \leq 4\Delta_j k_j^{-1} r.$$
2. Suppose for all $x \in G_j$, $\hat{f}'(x) < -\Delta_j$, then the algorithm returns the minimum of f.
3. Suppose for all $x \in G_j$, $\hat{f}'(x) > \Delta_j$, then the algorithm returns the minimum of f.
4. Suppose there exists some $x_1, x_2 \in G_1$ such that $\hat{f}'(x_1) < -\Delta_j$ and $\hat{f}'(x_2) > \Delta_j$. Then I is updated, and its length shrinks to $k_j^{-1}r$. I contains the minimum $x_{f,c,r}^\star$.

Now, when the algorithm does not terminate until we finish the M-th loop, the interval I now contains $x_{f,c,r}^\star$ with length at most $k_M^{-1}r$. In this case, we return the middle point of the interval, so that by Lipschitzness of the objective function f, we have,
$$f(\hat{x}) - f(x_{f,c,r}^\star) \leq L|\hat{x} - x_{f,c,r}^\star| \leq L|I| \leq Lk_M^{-1}r.$$
The desired claim of the proposition now thus follows. ■

Corollary 47 *Assume $n \geq 30$, then the parameters in Eq (122) satisfy*
$$\min_{1\leq j\leq M} T_j \geq 1, k_M \leq n^{-1/2} \text{ and } \max_{1\leq j\leq M}\left\{\Delta_j k_j^{-1}\right\} \leq 3\sigma n^{-\frac{1}{2}}(\log_2(n))^{\frac{1}{2}}\left(\log\frac{3}{\delta} + \log n\right)^{\frac{1}{2}}.$$
Moreover, the output $\hat{x}$ from algorithm 10 with the input defined in Eq (122) satisfies,
$$\mathbb{P}\left(f(\hat{x}) - f(x_{f,c,r}^\star) \leq \gamma^\star\right) \geq 1 - \delta,$$
where
$$\gamma^\star := \max\left\{12\sigma(\log_2(n))^{\frac{1}{2}}\left(\log\frac{3}{\delta} + \log n\right)^{\frac{1}{2}}, 2L\right\} n^{-\frac{1}{2}} r$$

Proof Note that, the second part of the corollary follows immediately from the first part and proposition 45. To prove the first part, note first that, $k_j \leq k_M \leq n^{\frac{1}{2}}$ for $1 \leq j \leq M$. Since $n \geq 30$, we get for $1 \leq j \leq M$,
$$\frac{n}{M(2k_j+1)} \geq \frac{n}{\frac{1}{2}\log_2 n(2n^{\frac{1}{2}}+1)} \geq 1 \;\Rightarrow\; T_j \geq 1.$$
Finally, note that, since we have, for all $1 \leq j \leq M$:
$$T_j \geq \frac{n}{2M(2k_j+1)} \geq \frac{n}{\log_2(n)\cdot(2k_j+1)} \geq \frac{n}{3k_j\log_2(n)} \text{ and } 2k_j + 1 \leq 3n^{\frac{1}{2}}.$$
this gives us that, when $n \geq 30$, for all $1 \leq j \leq M$:
$$\Delta_j k_j^{-1} \leq \sqrt{\frac{6\sigma^2\log_2 n}{nk_j}\log\frac{3n^{1/2}\log_2(n)}{\delta}} \leq 3\sigma\sqrt{\frac{\log_2(n)}{n}\left(\log\frac{3}{\delta} + \log n\right)}.$$
■

Appendix J. Lipschitz Function with Zeroth-Order Oracle

J.1. Lipschitz Function with Zeroth-Order Oracle, $d = 1$ and $M \geq 1$

J.1.1. DESCRIPTION OF ALGORITHMS

In this section, we propose two generic algorithms: algorithm 11 and algorithm 12 that builds from algorithm 11. We note here that, algorithm 12 in essence, builds from M times of repeated calls of algorithm 11. As will be shown immediately in the later subsections, it turns out that the two algorithms with careful choice of parameters return the minimax estimator in the single and multi rounds respectively.

Algorithm 11 Routine for One Stage $d = 1$ Lipschitz Functions $\mathcal{F}_\lambda$ (Zeroth-Order Oracle)

Input: User's choice of the left point and right point of interval l_1 and l_2, grid size parameter $k \in \mathbb{N}$ and sampling times $T \in \mathbb{N}$, value m_1 and m_2 and function gap $U \in \mathbb{R}_+$.

1: Set $c = \frac{1}{2}(l_1 + l_2)$ and $r = \frac{1}{2}(l_2 - l_1)$. Compute the grid points $G = G(c, r, k)$.

2: At each point $x \in G$, query the zeroth oracle T times and denote each sample function value via $\{f(x)^{(1)}, f(x)^{(2)}, \ldots, f(x)^{(T)}\}$.

3: Compute the function value estimate at each point $x \in G$ via $\hat{f}(x) = \frac{1}{T}\sum_{i=1}^{T} f(x)^{(i)}$.

4: Compute the estimate $x \in G$ via $\hat{x} = \operatorname{argmin}_{x \in G} \hat{f}(x)$.

5: Compute $\hat{m}_1$ and $\hat{m}_2$ as follows:

$$\hat{m}_1 = \min\{x \in G : \hat{f}(x) - \hat{f}(\hat{x}) \leq U\} \text{ and } \hat{m}_2 = \max\{x \in G : \hat{f}(x) - \hat{f}(\hat{x}) \leq U\}$$

6: Compute $\hat{l}_1$ and $\hat{l}_2$ as follows:

$$\hat{l}_1 = \max\left\{m_1 - \frac{r}{k}, l_1\right\} \text{ and } \hat{l}_2 = \min\left\{m_2 + \frac{r}{k}, l_1\right\}$$

7: Return the the estimator $\hat{x}$, the value of $\hat{l}_1$, $\hat{l}_2$, $\hat{m}_1$ and $\hat{m}_2$.

Algorithm 12 Routine for Multi Stage $d = 1$ Lipschitz Functions $\mathcal{F}_\lambda$ (Zeroth-Order Oracle)

Input: Initialization of th eparameters $(l_1^{(1)}, l_2^{(1)}, m_1^{(1)}, m_2^{(1)}, k^{(1)}, T^{(1)}, U^{(1)}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{N} \times \mathbb{N} \times \mathbb{R}_+$. User specifies the updating rule used in line (3) of the algorithm.

1: **for** $i = 1$ to M **do**

2: Run algorithm 11 with input parameter $(l_1^{(i)}, l_2^{(i)}, m_1^{(i)}, m_2^{(i)}, k^{(i)}, T^{(i)}, U^{(i)})$. Denote the output to be $\hat{l}_1^{(i)}$, $\hat{l}_2^{(i)}$, $\hat{m}_1^{(i)}$, $\hat{m}_2^{(i)}$ and $\hat{x}^{(i)}$.

3: Update $(l_1^{(i+1)}, l_2^{(i+1)}, m_1^{(i+1)}, m_2^{(i+1)}, k^{(i+1)}, T^{(i+1)}, U^{(i+1)})$.

4: **end for**

5: **return** $\hat{x}_M$ as the estimate of $x_f^\star$, and $\hat{U} = U^{(M+1)}$.

J.1.2. ANALYSIS OF ALGORITHM 11: SINGLE-STAGE ANALYSIS

In this section, we show that, with careful choice of input parameters, algorithm 11 returns some estimator $\hat{x}$ so that its associated risk $\mathcal{R}$ is upper bounded by $\tilde{O}\left(n^{-\frac{1}{3}}\right)$. For purpose of convenience

in later discussion on upper bounds for multi-stage algorithm, in this section, we slightly generalize the domain of interest by considering the domain $\mathbf{D} = [l_1, l_2]$ (while noting that $l_1 = 0$ and $l_2 = 1$ corresponds to the original domain $\mathbf{D} = [0, 1]$). In this sense, the algorithm discussed in this section seeks to estimate $x^\star_{f,l_1,l_2}$, the unique minimum of f in the domain $\mathbf{D}_{l_1,l_2}$, while the risk of interest would be $\mathbb{E}\left[f(\hat{x}) - f(x^\star_{f,l_1,l_2})\right]$.

Proposition 48 *Given any fix $c \in \mathbb{R}^d$ and r, suppose there exists some $k \in \mathbb{N}$ satisfying*

$$(2k+1)\left\lceil 2k^2 \log(2k+1)\right\rceil \le nr^2 \ \text{ and } \ k \ge 6. \tag{123}$$

Set $T = \left\lfloor \frac{n}{2k+1} \right\rfloor$ and $U = 2Lr$. Then, if we denote $\hat{l}_1$, $\hat{l}_2$, $\hat{m}_1$, $\hat{m}_2$ and $\hat{x}$ to be the output from algorithm 12 with input $l_1 = m_1 = c - r$, $l_2 = m_2 = c + r$, and k, T, U set up as above, we have, with probability at least $1 - \delta$:

1. *The output set $[\hat{m}_1, \hat{m}_2]$ is amenable to $[\hat{l}_1, \hat{l}_2]$ with parameter $(k, t) = \left(6, \frac{1}{6}\right)$.*
2. *The output estimator $\hat{x}$ satisfies $f(\hat{x}) - f(x^\star_{f,l_1,l_2}) \le k^{-1} r \left(6L + 2\sigma \left(\log \frac{2}{\delta} + 1\right)^{\frac{1}{2}}\right)$.*
3. *For any $x \in \{\hat{m}_1, \hat{m}_2\}$, we have, $f(x) - f(x^\star_{f,l_1,l_2}) \le k^{-1} r \left(18L + 2\sigma \left(\log \frac{2}{\delta} + 1\right)^{\frac{1}{2}}\right)$.*
4. *The output $[\hat{l}_1, \hat{l}_2]$ contains the minimum: any minimum $x^\star_{f,l_1,l_2}$ satisfies $x^\star_{f,l_1,l_2} \in [\hat{l}_1, \hat{l}_2]$.*

Proof The proof of the proposition relies on the following two critical lemma. The first lemma bounds the deviation of the function value $f(x)$ and its estimate $\hat{f}(x)$ on the grid G. The second lemma establishes a critical property of the algorithm, which turns out to be very useful in the analysis of multi-stage algorithm. We start with the first lemma. Consider the following event,

$$\Gamma := \left\{ \left| \hat{f}(x) - f(x) \right| \le r^a := \sigma \sqrt{\frac{2}{T} \log \frac{2(2k+1)}{\delta}} \ \text{ for all } x \in G \right\}.$$

The next lemma shows that Γ happens with probability at least $1 - \delta$.

Lemma 49 *We have $\mathbb{P}(\Gamma) \ge 1 - \delta$.*

Proof First, for each $x \in G$, denote $\epsilon(x) := \widehat{f}(x) - f(x)$. Then, since by our assumption, the noise $\{\widehat{f}(x) - f(x)\}_{x=1}^T$ is mean 0, independent and subgaussian with parameter σ^2, we have that $\epsilon(x)$ is mean 0 and subgaussian with parameter σ^2/T. Therefore, for any fix $x \in G$,

$$\mathbb{P}\left(|\epsilon(x)| \ge r^a\right) \le 2 \exp\left(-\frac{(r^a)^2 T}{2\sigma^2}\right) \le \delta (2k+1)^{-1},$$

where the first inequality above uses the subgaussianity of $\epsilon(x)$, and the second inequality uses the definition of k and T. Now, the desired claim of the lemma follows from the fact that $|G| = 2k + 1$ and the union bound of the above events. ■

Before we introduce the next lemma, we introduce the following concept.

Definition 50 *An interval $[m_1, m_2]$ is amenable to another interval $[l_1, l_2]$ with parameter $k^\star \in \mathbb{N}, t^\star \in \mathbb{R}_+$ if it satisfies the following two conditions:*

1. *The set $[m_1, m_2] \subseteq [l_1, l_2]$.*
2. *For any $k \in \mathbb{N}$, denote G_k to be the grid points $G_k = G(c, r, k)$ with $c = \frac{1}{2}(l_1 + l_2)$ and $r = \frac{1}{2}(l_2 - l_1)$. Then, for any $x \in [l_1, l_2]$ and $k \geq k^\star$, there exists three consecutive grids $x_0 \leq x_1 \leq x_2 \in G_k$ such that, either $m_1 \leq x_0 \leq x_1 \leq x \leq x_2, x - m_1 \geq t^\star(l_2 - l_1)$ or $x_0 \leq x \leq x_1 \leq x_2 \leq m_2, m_2 - x \geq t^\star(l_2 - l_1)$.*

The lemma below is helpful in understanding the above concept.

Lemma 51 *Let $[l_1, l_2]$ be some interval on $\mathbb{R}$. Suppose m_1, m_2 satisfy the following condition:*

$$m_1 \leq l_1 \leq l_2 \leq m_2 \text{ and } m_2 - m_1 \geq \frac{1}{3}(l_2 - l_1). \tag{124}$$

Then, the interval $[m_1, m_2]$ is amenable to $[l_1, l_2]$ with parameter $\left(6, \frac{1}{6}\right)$

Proof By definition, we need to show that, $[m_1, m_2]$ satisfies the following conditions:

1. The set $[m_1, m_2] \subseteq [l_1, l_2]$.
2. For any $k \geq 6$, denote G_k to be the grid points $G_k = G(c, r, k)$ with $c = \frac{1}{2}(l_1 + l_2)$ and $r = \frac{1}{2}(l_2 - l_1)$. Then, for any $x \in [l_1, l_2]$ and $k \geq 6$, there exists three consecutive grids $x_0 \leq x_1 \leq x_2 \in G_k$ such that, either $m_1 \leq x_0 \leq x_1 \leq x \leq x_2, x - m_1 \geq \frac{1}{6}(l_2 - l_1)$ or $x_0 \leq x \leq x_1 \leq x_2 \leq m_2, m_2 - x \geq \frac{1}{6}(l_2 - l_1)$.

The first condition is satisfied according to the first group of inequality of Eq (124). To show the second inequality, pick any $x \in [l_1, l_2]$, and we divide our discussion into two cases. (i) $x \geq \frac{1}{2}(m_1 + m_2)$. Then, since $k \geq 6$, we can take three consecutive grids $x_0, x_1, x_2 \in G_k$ such that $x \in [x_1, x_2]$. Then, by definition,we know that, $x_2 - x_1 = x_1 - x_0 = \frac{1}{2k}(l_2 - l_1)$. This gives $x_0 \geq x - \frac{1}{6}(l_2 - l_1)$. Since $m_2 - m_1 \geq \frac{1}{3}(l_2 - l_1)$ by Eq (124), we get that,

$$x - m_1 \geq \frac{1}{2}(m_2 - m_1) \geq \frac{1}{6}(l_2 - l_1) \text{ and } x_0 \geq x - \frac{1}{6}(l_2 - l_1) \geq m_1.$$

The desired claim of the lemma now thus follows. (ii)$x \leq \frac{1}{2}(m_1 + m_2)$. One can similarly show that, for $k \geq 6$. there exists consecutive grids $x_0, x_1, x_2 \in G_k$ such that $x_0 \leq x \leq x_1 \leq x_2 \leq m_2, m_2 - x \geq \frac{1}{6}(l_2 - l_1)$. The proof under this case is essentially the same as that under the case where $x \geq \frac{1}{2}(m_1 + m_2)$. ■

Now, we are ready to introduce the following lemma. It is deterministic in nature.

Lemma 52 *Assume that, the following four conditions hold:*

1. *The set $[m_1, m_2]$ is amenable to the set $[l_1, l_2]$ with parameter $(k, t) = \left(6, \frac{1}{6}\right)$.*
2. *For any $x \in [m_1, m_2]$, we have, $f(x) - f(x^\star_{f,l_1,l_2}) \leq U'$ for some $U' > 0$.*
3. *The grid $G = G(c, r, k)$ with $c = \frac{1}{2}(l_1 + l_2)$, $r = \frac{1}{2}(l_2 - l_1)$ and $k \geq 6$.*

4. *For any $x \in G$, we have, $|\hat{f}(x) - f(x)| \leq r^a$.*

Now, let us denote U to be $U = 9U'k^{-1} + 2r^a > 0$. Then, the output of algorithm 11 with input parameters l_1, l_2, m_1, m_2, k, T and U satisfies:

1. *The output set $[\hat{m}_1, \hat{m}_2]$ is amenable to $[\hat{l}_1, \hat{l}_2]$ with parameter $(k, t) = \left(6, \frac{1}{6}\right)$.*
2. *The output estimator $\hat{x}$ satisfies $f(\hat{x}) - f(x^\star_{f,l_1,l_2}) \leq 3U'k^{-1} + 2r^a$.*
3. *For any $x \in [\hat{m}_1, \hat{m}_2]$, we have, $f(x) - f(x^\star_{f,l_1,l_2}) \leq 9U'k^{-1} + 2r^a = U$.*
4. *The output $[\hat{l}_1, \hat{l}_2]$ contains the minimum: any minimum $x^\star_{f,l_1,l_2}$ satisfies $x^\star_{f,l_1,l_2} \in [\hat{l}_1, \hat{l}_2]$.*

Proof First note that, $[m_1, m_2]$ is amenable to $[l_1, l_2]$ with parameters $(6, \frac{1}{6})$. Since $x^\star_{f,l_1,l_2} \in [l_1, l_2]$, and the grid size $k \geq 6$, thus, by definition, we know that, there exist three consecutive grid points $x_0 \leq x_1 \leq x_2 \in G_k$ such that, either $m_1 \leq x_0 \leq x_1 \leq x^\star_{f,l_1,l_2} \leq x_2$, $x^\star_{f,l_1,l_2} - m_1 \geq \frac{1}{6}(l_2 - l_1)$ or $x_0 \leq x^\star_{f,l_1,l_2} \leq x_1 \leq x_2 \leq m_2$ and $m_2 - x^\star_{f,l_1,l_2} \geq \frac{1}{6}(l_2 - l_1)$. We only prove the desired claim of the lemma under the first situation, that is $m_1 \leq x_0 \leq x_1 \leq x^\star_{f,l_1,l_2} \leq x_2$, $x^\star_{f,l_1,l_2} - m_1 \geq \frac{1}{6}(l_2 - l_1)$, while noting that the desired result of the lemma can be proved in a totally similar way under the other situation.

We start by showing the first claim of the desired result. To do so, we show that, $\hat{m}_1 \leq x_0 \leq x_1 \leq \hat{m}_2$. Indeed whenever $z \in \{x_0, x_1\}$, we know that, $m_1 \leq z \leq x^\star_{f,l_1,l_2}$, and therefore, by convexity of f, we have,

$$f(z) \leq \frac{z_1 - m_1}{x^\star_{f,l_1,l_2} - m_1} f(x^\star_{f,l_1,l_2}) + \frac{x^\star_{f,l_1,l_2} - z_1}{x^\star_{f,l_1,l_2} - m_1} f(m_1).$$

Hence, the value $f(z) - f(x^\star_{f,l_1,l_2})$ for $z \in \{x_0, x_1\}$ can be upper bounded by:

$$f(z) - f(x^\star_{f,l_1,l_2}) \leq \frac{x^\star_{f,l_1,l_2} - z}{x^\star_{f,l_1,l_2} - m_1} \left(f(m_1) - f(x^\star_{f,l_1,l_2})\right).$$

Now, note that, since we have $0 \leq x^\star_{f,l_1,l_2} - x_0 \leq \frac{1}{k}(l_2 - l_1)$, $0 \leq x^\star_{f,l_1,l_2} - x_1 \leq \frac{1}{2k}(l_2 - l_1)$, $\frac{1}{6}(l_2 - l_1) \leq x^\star_{f,l_1,l_2} - m_1$ and $f(m_1) - f(x^\star_{f,l_1,l_2}) \leq U'$, we get that,

$$f(x_1) - f(x^\star_{f,l_1,l_2}) \leq 3U'k^{-1} \text{ and } f(x_0) - f(x^\star_{f,l_1,l_2}) \leq 6U'k^{-1}. \tag{125}$$

Thus, whenever $z \in \{x_0, x_1\}$, we get,

$$\hat{f}(z) - \hat{f}(x^\star_{f,l_1,l_2}) \leq \left|\hat{f}(z) - f(z)\right| + f(z) - f(x^\star_{f,l_1,l_2}) + \left|\hat{f}(x^\star_{f,l_1,l_2}) - f(x^\star_{f,l_1,l_2})\right| \leq 2r^a + 6U'k^{-1} = U.$$

This gives $\hat{m}_1 \leq x_0 \leq x_1 \leq \hat{m}_2$. As an immediate consequence, we get $\hat{m}_2 \geq \hat{m}_1 + k^{-1}r$. Since by definition, we know $\hat{l}_1 \leq \hat{m}_1 \leq \hat{l}_1 + k^{-1}r$ and $\hat{m}_2 \leq \hat{l}_2 \leq \hat{m}_2 + k^{-1}r$, we get that,

$$\hat{l}_1 \leq \hat{m}_1 \leq \frac{2}{3}\hat{l}_1 + \frac{1}{3}\hat{l}_2 \text{ and } \frac{1}{3}\hat{l}_1 + \frac{2}{3}\hat{l}_2 \leq \hat{m}_2 \leq \hat{l}_2.$$

Thus, by lemma 51, we get that, $[\hat{m}_1, \hat{m}_2]$ is amenable to $[\hat{l}_1, \hat{l}_2]$ with parameter $(6, \frac{1}{6})$.

Next, we show the second and third claim of the desired result. Indeed, note first that, by definition, $\hat{f}(\hat{x}) \leq \hat{f}(x_1)$. In addition, we know from the condition 4 of the lemma that, $f(\hat{x}) - \hat{f}(\hat{x}) \leq r^a$ and $\hat{f}(x_1) - f(x_1) \leq r^a$. Lastly, we know from Eq (125) that, $f(x_1) - f(x^\star_{f,l_1,l_2}) \leq 3U'k^{-1}$. Thus, combining all these bounds together, we have,

$$f(\hat{x}) - f(x^\star_{f,l_1,l_2}) \leq f(\hat{x}) - \hat{f}(\hat{x}) + \hat{f}(\hat{x}) - \hat{f}(x_1) + \hat{f}(x_1) - f(x_1) + f(x_1) - f(x^\star_{f,l_1,l_2}) \leq 3U'k^{-1} + 2r^a.$$

This gives us the second claim of the desired result. Similarly, notice that, for any $z \in \{\hat{m}_1, \hat{m}_2\}$, we have $f(z) - \hat{f}(z) \leq r^a$. and the definition of $\hat{m}_1$, $\hat{m}_2$ and $\hat{x}$ implies that, $\hat{f}(z) - \hat{f}(x_1) \leq \hat{f}(z) - \hat{f}(\hat{x}) \leq 6U'k^{-1} + 2r^a$. Thus, combining these bounds together, we get for $z \in \{\hat{m}_1, \hat{m}_2\}$,

$$f(z) - f(x^\star_{f,l_1,l_2}) \leq f(z) - \hat{f}(z) + \hat{f}(z) - \hat{f}(x_1) + \hat{f}(x_1) - f(x_1) + f(x_1) - f(x^\star_{f,l_1,l_2}) \leq 9U'k^{-1} + 2r^a.$$

This gives us the third claim of the desired result.

In the last step, we show the fourth claim of the desired result. To do so, we need to show that, $\hat{l}_1 \leq x^\star_{f,l_1,l_2} \leq \hat{l}_2$. First, we know that $x^\star_{f,l_1,l_2} \geq \hat{m}_1 \geq \hat{l}_1$. Next, we show that $x^\star_{f,l_1,l_2} \leq \hat{m}_2$. We divide our discussion into two cases: (i) $\hat{m}_2 = l_2$. In this case, the result is trivial. (ii) $\hat{m}_2 < l_2$. In this case, by definition of $\hat{l}_2$, we know that in this case $\hat{l}_2 = \hat{m}_2 + k^{-1}r$. Since we have already shown that $\hat{m}_2 \geq x_1$, the fact that, $x^\star_{f,l_1,l_2} \leq x_1 + k^{-1}r$ gives us that, $x^\star_{f,l_1,l_2} \leq \hat{l}_2$. Now, altogether, we have shown the fourth claim of the desired result. ■

Now, we are ready to show the desired claim of the proposition. Since the function $f(x)$ is known to be Lipschitz with parameter L on the interval $[l_1, l_2]$, thus, it is known that, for all $x \in [l_1, l_2]$, we have, $f(x) - f(x^\star_{f,l_1,l_2}) \leq L(l_2 - l_1) \leq 2Lr$. Now that, the condition on k shows that $T \geq 2k^2 \log(2k+1) r^{-2}$. As a consequence, this gives us that,

$$r^a = \sigma \sqrt{\frac{2}{T} \log \frac{2(2k+1)}{\delta}} \leq \frac{\sigma r}{k} \cdot \sqrt{\log \frac{2}{\delta} + 1}.$$

Now, the desired claim of the proposition follows from lemma 52 and above bound on r^a. ■

Motivated by proposition 48, it becomes important to understand when such k exists in Eq (123) and how large it is.

Lemma 53 *Assume n is large enough satisfying $nr^2 \geq 6^{12}$. Denote $k(r) = (\gamma(nr^2))^{\frac{1}{3}}$, and $k^\star = \lfloor \frac{1}{3} k(r) \rfloor$. Then $k^\star \in \mathbb{N}$ and it satisfies Eq (123).*

Proof Note that, the second part of the corollary follows easily from the first part, and proposition 48. To show the first part, note that, $\gamma(x) \geq \sqrt{x}$ when $x \geq 3$. Therefore, whenever $nr^2 \geq 6^{12}$, we have, $\gamma(nr^2) \geq 6^6$. As a consequence, $k(r) \geq 36$, and hence, $k^\star \geq 6$. Now, we show that $k^\star$ satisfies Eq (123). Indeed, when $k = k^\star$,

$$(2k+1) \left\lfloor 2k^2 \log(2k+1) \right\rfloor \leq (3k)^3 \log(3k)^3 \leq k(r)^3 \log(k(r))^3 \leq nr^2,$$

where the last inequality follows from the fact that, for any $x > 0$, $\gamma(x) \log \gamma(x) \leq x$. ■

Now, proposition 48 and Lemma 53 immediately give us the following corollary.

Corollary 54 *Given any fix $c \in \mathbb{R}^d$ and $r \in (0,1]$, set $k = \left\lfloor \frac{1}{3}\gamma(nr^2)^{\frac{1}{3}} \right\rfloor$ and $T = \left\lfloor \frac{n}{2k+1} \right\rfloor$. Assume n is large enough satisfying $nr^2 \geq 6^{12}$. Then, if we denote $\hat{x}$ to be the output from algorithm 12 with input $l_1 = m_1 = c - r$, $l_2 = m_2 = c + r$, and k, T, U set up as above, we get,*

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,l_1,l_2}) \leq 12Bn^{-\frac{1}{3}}r^{\frac{1}{3}}\log(nr^2)^{\frac{1}{3}}\right) \geq 1-\delta \text{ where } B = 3L + \sigma\left(\log\frac{2}{\delta}+1\right)^{\frac{1}{2}}.$$

J.1.3. ANALYSIS OF ALGORITHM 12: MULTI-STAGE ANALYSIS

In this section, we show that with careful choice of input parameters l_1, l_2, k_1, T_1 and U_1, and appropriate updating rule, then algorithm returns some estimator $\hat{x}$ that is minimax optimal. The idea is from the crucial lemma 52: lemma 52 helps locate an interval such that all point in this new interval has smaller function value gap to the minimum value $f(x^\star_{f,l_1,l_2})$. The idea is to apply this technique in multiple rounds, and after each round, get a better convergence rate of the function value.

To be clear about how we specify the updating rule in line 3 of the algorithm 12, we summarize it as follows: given the output parameters $\hat{l_1}^{(i)}$, $\hat{l_2}^{(i)}$, $\hat{m_1}^{(i)}$, $\hat{m_2}^{(i)}$, we update $l_1^{(i+1)}$, $l_2^{(i+1)}$, $m_1^{(i+1)}$, $m_2^{(i+1)}$, $k^{(i+1)}$, $T^{(i+1)}$ and $U^{(i+1)}$ as described in Algorithm 13. Our update also requires knowledge of $\{k^{(j)}\}_{j=1}^i$ and $\{U^{(j)}\}_{j=1}^i$.

Algorithm 13 Updating Rule in Algorithm 12

Input:$l_1^{(i)}$, $l_2^{(i)}$, $m_1^{(i)}$, $m_2^{(i)}$, $\{k^{(j)}\}_{j=1}^i$ and $\{U^{(j)}\}_{j=1}^i$.
(i) Update $l_1^{(i+1)} = \hat{l}_1^{(i)}$, $l_2^{(i+1)} = \hat{l}_2^{(i)}$, $m_1^{(i+1)} = \hat{m}_1^{(i)}$, $m_2^{(i+1)} = \hat{m}_2^{(i)}$.
(ii) Update $k^{(i+1)}$ to be the largest $k \in \mathbb{N}$ such that,

$$(2k+1)\Pi_{j=1}^i\left(k^{(j)}\right)^2\lceil 2\log(2k+1)\rceil \leq n \tag{126}$$

If no such k exists, return FAIL.
(iii) Update $T^{(i+1)}$ to be $T^{(i+1)} = \left\lfloor \frac{n}{2k^{(i)}+1} \right\rfloor$.
(iv) Update $U^{(i+1)}$ to be $U^{(i+1)} = 9U^{(i)}\left(k^{(i)}\right)^{-1} + 2\sigma\Pi_{j=1}^i(k^{(j)})^{-1}\left(\log\frac{2}{\delta}+1\right)^{\frac{1}{2}}$.

Proposition 55 *Let $\mathbf{D} = [0,1]$ and we are given the confidence level $\delta > 0$. Set the initial parameter $l_1^{(1)} = m_1^{(1)} = 0$ and $l_2^{(1)} = m_2^{(1)} = 1$. Set the parameters $k^{(1)}$ to be the largest $k \in \mathbb{N}$ such that, $(2k+1)\lceil 2\log(2k+1)\rceil \leq n$. Set the initial parameters $T^{(1)} = \left\lfloor \frac{n}{2k+1} \right\rfloor$ and $U^{(1)} = L$. Consider algorithm 12 that uses algorithm 13 to be the updating rule. Now, assuming that the $\{k^{(i)}\}_{i=1}^M$ exist and satisfy $k^{(i)} \geq 6$ for all $1 \leq i \leq M$. Then, if we denote $\hat{x}$ and $\hat{U}$ to be the output from algorithm 12, we get,*

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,l_1,l_2}) \leq U^{(M+1)}\right) \geq 1-\delta.$$

Proof The proof of the proposition relies on the following two components. The first component is the following probabilistic lemma that bounds the deviation of the function value $f(x)$ and its

estimate $\hat{f}(x)$ on the grid $\{G_i\}_{i=1}^M$. The second component is the crucial lemma 51 that we have shown before. We start with the first component. Consider the following event Γ,

$$\Gamma = \bigcap_{i=1}^{M} \left\{ |\hat{f}(x) - f(x)| \leq r_i^a := \sigma \sqrt{\frac{2}{T_i} \log \frac{2M(2k_i+1)}{\delta}} \text{ for all } x \in G_i \right\}.$$

The following lemma shows that Γ happens with probability at least $1-\delta$.

Lemma 56 *We have* $\mathbb{P}(\Gamma) \geq 1-\delta$.

Proof First, fix $1 \leq i \leq M$. Note that, for each $x \in G_i$, denote $\epsilon(x) := \widehat{f}(x) - f(x)$. Then, since by our assumption, the noise $\{\widehat{f}(x) - f(x)\}_{x=1}^{T_i}$ is mean 0, independent and subgaussian with parameter σ^2, we have that $\epsilon(x)$ is mean 0 and subgaussian with parameter σ^2/T_i. Therefore, for any fix $x \in G_i$,

$$\mathbb{P}\left(|\epsilon(x)| \geq r^a\right) \leq 2\exp\left(-\frac{(r_i^a)^2 T_i}{2\sigma^2}\right) \leq \delta M^{-1}(2k_i+1)^{-1},$$

where the first inequality above uses the subgaussianity of $\epsilon(x)$, and the second inequality uses the definition of r^a. Since $|G_i| = 2k_i+1$, after taking a union bound, we get for $1 \leq i \leq M$,

$$\mathbb{P}\left(\exists x \in G_i \text{ such that } |\hat{f}(x) - f(x)| \geq r_i^a\right) \leq \delta M^{-1}.$$

Now, the desired claim of the lemma follows from a union bound on $1 \leq i \leq M$. ■

We are now ready to prove the desired claim of the proposition. Indeed, we can prove the following via induction on the rounds $1 \leq i \leq M$:

1. The output $[\hat{m}_1^{(i+1)}, \hat{m}_2^{(i+1)}]$ is amenable to $[\hat{l}_1^{(i+1)}, \hat{l}_2^{(i+1)}]$ with parameter $(k,t) = \left(6, \frac{1}{6}\right)$.
2. The output estimator $\hat{x}^{(i+1)}$ satisfies $f(\hat{x}^{(i)}) - f(x^\star_{f,l_1,l_2}) \leq 3U^{(i)}(k^{(i)})^{-1} + 2r_i^a$.
3. For any $x \in [\hat{m}_1^{(i+1)}, \hat{m}_2^{(i+1)}]$, we have, $f(x) - f(x^\star_{f,l_1,l_2}) \leq 9U^{(i)}(k^{(i)})^{-1} + 2r_i^a \leq U^{(i+1)}$.
4. The output $[\hat{l}_1, \hat{l}_2]$ contains the minimum: any minimum $x^\star_{f,l_1,l_2}$ satisfies $x^\star_{f,l_1,l_2} \in [\hat{l}_1, \hat{l}_2]$.

The only trick is to apply lemma 52 repeatedly and notice the fact that $T^{(i)} \geq 2\log(2k^{(i)}+1)\Pi_{j=1}^{i}(k^{(j)})^2$ implies the following upper bound on r_i^a:

$$r_i^a := \sigma\sqrt{\frac{2}{T^{(i)}} \log \frac{2M(2k^{(i)}+1)}{\delta}} \leq \sigma \Pi_{j=1}^{i}(k^{(j)})^{-1} \left(\log \frac{2M}{\delta} + 1\right)^{\frac{1}{2}}.$$

■

Motivated by Proposition 55, it becomes important to understand how large $\{k^{(i)}\}_{i=1}^M$ and $\{U^{(i)}\}_{i=1}^M$ are.

Lemma 57 *Suppose that n is large enough satisfying*

$$\log\log n \geq M\log 3 + \log 12 + \log\log 6. \tag{127}$$

Then, the sequence $\{k^{(i)}\}_{i=1}^{M}$, $\{T^{(i)}\}_{i=1}^{M}$ and $\{U^{(i)}\}_{i=1}^{M}$ are well-defined via algorithm 13. In addition, $k^{(i)} \geq 6$ for all $1 \leq i \leq M$. Finally, if we denote the output $U^{(M+1)}$ satisfies,

$$U^{(M+1)} \leq 9^{M+2}B(\log n)^{\frac{1}{2}} \cdot n^{\frac{1}{2}\left(1-\frac{1}{3^M}\right)} \text{ where } B = L + \sigma\left(\log\frac{2}{\delta}+1\right)^{\frac{1}{2}}. \tag{128}$$

Proof First of all, we show that $\{k^{(i)}\}_{i=1}^{M}$ are well defined via algorithm 13. To do so, we show that the following hypothesis hold for all $1 \leq i \leq M$:

$$(i) n \geq 6^{12}\Pi_{j=1}^{i}(k^{(j)})^2 \quad (ii) k^{(i)} \in \mathbb{N} \text{ and } k^{(i)} \geq 6 \tag{129}$$

$$(iii)\Pi_{j=1}^{i}k^{(j)} \geq (12)^{-\frac{3}{2}\left(1-\frac{1}{3^i}\right)}(\log n)^{-\frac{1}{2}\left(1-\frac{1}{3^i}\right)}n^{\frac{1}{2}\left(1-\frac{1}{3^i}\right)} \tag{130}$$

$$(iv)\Pi_{j=1}^{i}k^{(j)} \leq 6^{-\frac{3}{2}\left(1-\frac{1}{3^i}\right)}n^{\frac{1}{2}\left(1-\frac{1}{3^i}\right)} \tag{131}$$

We first show the base case $i = 1$. Note that, the first part of Eq (129) is implied by the assumption on n, both the second part of Eq (129), Eq (130) and Eq (131) follow from the first part of Eq (129) and corollary 54. Now, for some $i < M$, assuming that the induction hypothesis holds for all $j \in \{1, 2, \ldots, i\}$, we show that the hypothesis holds for $i+1$. We first show the second part of Eq (129) for $i+1$. Indeed, by induction hypothesis, we know that, the first inequality of Eq (129) is true for i, and thus the second part of Eq (129) follows from a direct application of corollary 54 (by substituting δ by δ/M and r by $\Pi_{j=1}^{i}k_j^{-1}$ there). Next, we show Eq (130) holds for $i+1$. In fact, note that, by definition of $k^{(i)}$, we know that,

$$k^{(i+1)} \geq \frac{1}{12}\left[\gamma\left(n\Pi_{j=1}^{i}(k^{(j)})^{-2}\right)\right]^{\frac{1}{3}} \geq \frac{1}{12}\left(n\Pi_{j=1}^{i}(k^{(j)})^{-2}\right)^{\frac{1}{3}}(\log n)^{-\frac{1}{3}},$$

and therefore, we get that,

$$\Pi_{j=1}^{i+1}k^{(j)} = \Pi_{j=1}^{i}k^{(j)} \cdot k^{(i+1)} \geq 12^{-1}n^{\frac{1}{3}}(\log n)^{-\frac{1}{3}}\left(\Pi_{j=1}^{i}k^{(j)}\right)^{\frac{1}{3}}$$

Now, using the induction hypothesis Eq (130), we get,

$$\begin{aligned}\Pi_{j=1}^{i+1}k^{(j)} &\geq 12^{-1}n^{-\frac{1}{3}}(\log n)^{\frac{1}{3}}\left((12)^{-\frac{3}{2}\left(1-\frac{1}{3^i}\right)}(\log n)^{-\frac{1}{2}\left(1-\frac{1}{3^i}\right)}n^{\frac{1}{2}\left(1-\frac{1}{3^i}\right)}\right)^{\frac{1}{3}} \\ &= (12)^{-\frac{3}{2}\left(1-\frac{1}{3^{i+1}}\right)}(\log n)^{-\frac{1}{2}\left(1-\frac{1}{3^{i+1}}\right)}n^{\frac{1}{2}\left(1-\frac{1}{3^{i+1}}\right)}\end{aligned}$$

This gives Eq (130) for the case $i+1$. In the third step, we show Eq (131) for $i+1$. The proof idea is similar to that of Eq (130). In fact, by definition of $k^{(i)}$, we know that,

$$k^{(i+1)} \leq \frac{1}{6}\left[\gamma\left(n\Pi_{j=1}^{i}(k^{(j)})^{-2}\right)\right]^{\frac{1}{3}} \leq \frac{1}{6}\left(n\Pi_{j=1}^{i}(k^{(j)})^{-2}\right)^{\frac{1}{3}},$$

and hence immediately we can get that,

$$\Pi_{j=1}^{i+1} k^{(j)} = \Pi_{j=1}^{i} k^{(j)} \cdot k^{(i+1)} \leq 6^{-1} n^{\frac{1}{3}} \left(\Pi_{j=1}^{i} k^{(j)} \right)^{\frac{1}{3}}.$$

Now, using the induction hypothesis Eq (131), we get,

$$\Pi_{j=1}^{i+1} k^{(j)} \leq 6^{-1} n^{-\frac{1}{3}} \left(6^{-\frac{3}{2}\left(1-\frac{1}{3^i}\right)} n^{\frac{1}{2}\left(1-\frac{1}{3^i}\right)} \right)^{\frac{1}{3}} \leq 6^{-\frac{3}{2}\left(1-\frac{1}{3^{i+1}}\right)} n^{\frac{1}{2}\left(1-\frac{1}{3^{i+1}}\right)}$$

Finally, we show the first part of Eq (129). Note that, by the proven fact of Eq (131) for $i+1$, it suffices to show that, $n^{1/3^{i+1}} \geq 6^{12}$, which would suffice if n is large enough so that,

$$\log \log n \geq M \log 3 + \log 12 + \log \log 6.$$

Now, we are ready to show the rest of the corollary. We first show Eq (128). Indeed, it is easy to use induction argument to show the following result: for all $0 \leq i \leq M-1$:

$$U^{(M+1)} = 9^{i+1} U^{(M-i)} \Pi_{j=0}^{i} k_{M-j}^{-1} + \frac{1}{4}(9^{i+1}-1)\sigma \left(\log \frac{2}{\delta} + 1 \right)^{\frac{1}{2}} \Pi_{j=0}^{M} k_j^{-1}.$$

Therefore, if we plug in $i = M-1$ in the above argument, we get that,

$$U^{(M+1)} \leq 9^M \left(U_1 + \sigma \left(\log \frac{2}{\delta} + 1 \right)^{\frac{1}{2}} \right) \Pi_{j=0}^{M} k_j^{-1} = 9^M \Pi_{j=0}^{M} k_j^{-1} \left(2L + \sigma \left(\log \frac{2}{\delta} + 1 \right)^{\frac{1}{2}} \right)$$

■

Now, Proposition 55 and lemma 57 together immediately give us the corollary below.

Corollary 58 *Let* $\mathbf{D} = [0,1]$. *Consider algorithm 12. Suppose we use the same initialization rule and use algorithm 13 to be the updating rule as that in Proposition 55. Then, when* n *is large enough satisfying Eq* (127)*, then the output* $\hat{x}$ *from algorithm 12 satisfies,*

$$\mathbb{P}(f(\hat{x}) - f(x^\star) \leq \gamma_M^\star) \geq 1 - \delta.$$

where

$$\gamma_M^\star = 9^{M+2} (\log n)^{\frac{1}{2}} \cdot n^{\frac{1}{2}\left(1-\frac{1}{3^M}\right)} \cdot \left(L + \sigma \left(\log \frac{2}{\delta} + 1 \right)^{\frac{1}{2}} \right), \tag{132}$$

J.2. Lipschitz Function with Zeroth-Order Oracle, $d \geq 1$ and $M = 1$

J.2.1. DESCRIPTION OF ALGORITHMS

In this section, we introduce the one stage algorithm for minimization of Lipschitz function under zeroth-order oracle: algorithm 14 parameterized by $(c, r, k, T) \in \mathbf{D} \times \mathbb{R}_+ \times \mathbb{N} \times \mathbb{N}$. Note that, algorithm 14 is in essence the same as algorithm 7.

Algorithm 14 Routine for One Stage Lipschitz Function $\mathcal{F}_\lambda$ (Zeroth-Order Oracle)

Input: User's choice of the sampling center $c \in \mathbf{D}$, radius $r \in \mathbb{R}_+$, grid size parameter $k \in \mathbb{N}$ and the sampling times $T \in \mathbb{N}$.

1: Compute the grid points $G = G(c, r, k)$.
2: At each point $x \in G$, query the zeroth-order oracle T times and denote each sample function value via $\{\widehat{f}(x)^{(1)}, \widehat{f}(x)^{(2)}, \ldots, \widehat{f}(x)^{(T)}\}$.
3: Compute the function value estimate at each point $x \in G$ via $\hat{f}(x) = \frac{1}{T}\sum_{i=1}^{T} \widehat{f}(x)^{(i)}$.
4: Compute the estimate $\hat{x} \in G$, defined by $\hat{x} := \operatorname{argmin}_{x \in G} \hat{f}(x)$.
5: Return the estimator $\hat{x}$.

J.2.2. ANALYSIS OF ALGORITHM 14

In this section, we show that, with careful choice of input parameters (c, r, k, T), algorithm 14 returns some estimator $\hat{x}$ that achieves the minimax risk (up to constants and logarithmic factors). The analysis is pretty close to that of algorithm 7. For convenience, we slightly generalize the domain of interest by considering $\mathbf{D}_{c,r} = \{x : \|x - c\|_\infty \le r\}$ parameterized by $c \in \mathbb{R}^d$ and $r \in \mathbb{R}$. The target now becomes $x^\star_{f,c,r}$, the unique minimum of f in the domain $\mathbf{D}_{c,r}$, and the risk of interest would be $\mathbb{E}\left[f(\hat{x}) - f(x^\star_{f,c,r})\right]$.

Proposition 59 *Given any fix $c \in \mathbb{R}^d$ and $r \in (0, 1]$, suppose there exists $k \in \mathbb{N}$ satisfying*

$$(2k+1)^d \left\lceil 2k^2 \log(2k+1) \right\rceil \le nr^2. \tag{133}$$

Then, pick any $k \in \mathbb{N}$ satisfying Eq (133), and set $T = \left\lfloor \frac{n}{(2k+1)^d} \right\rfloor$. Denote $\hat{x}$ to be the output from algorithm 14 when we input (c, r, k, T) as the input parameters. Then, we have

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,c,r}) \le \frac{r}{k} \cdot \left(2\sigma\left(\sqrt{2\log\frac{2}{\delta} + d}\right) + L\sqrt{d}\right)\right) \ge 1 - \delta.$$

Proof Denote $\bar{x} = \operatorname{argmin}_{x \in G} \left\|\bar{x} - x^\star_{f,c,r}\right\|_2$. Then, by construction of the grids G and Lipschitzness of the function f, we know that,

$$\left\|\bar{x} - x^\star_{f,c,r}\right\|_2 \le \frac{r\sqrt{d}}{k} \quad \text{and} \quad f(\bar{x}) - f(x^\star_{f,c,r}) \le L\left\|\bar{x} - x^\star_{f,c,r}\right\|_2 \le \frac{Lr\sqrt{d}}{k}.$$

Now, let us consider the following event:

$$\Gamma = \left\{\left|\widehat{f}(x) - f(x)\right| \le r^a := \sigma\sqrt{\frac{2}{T}\log\frac{2(2k+1)^d}{\delta}} \text{ for all } x \in G\right\},$$

The next lemma shows that Γ happens with probability at least $1 - \delta$.

Lemma 60 *We have $\mathbb{P}(\Gamma) \ge 1 - \delta$.*

Proof First, for each $x \in G$, denote $\epsilon(x) := \widehat{f}(x) - f(x)$. Then, since by our assumption, the noise $\{\widehat{f}(x) - f(x)\}_{x=1}^{T}$ is mean 0, independent and subgaussian with parameter σ^2, we have that $\epsilon(x)$ is mean 0 and subgaussian with parameter σ^2/T. Therefore, for any fix $x \in G$,

$$\mathbb{P}\left(|\epsilon(x)| \geq r^a\right) \leq 2\exp\left(-\frac{(r^a)^2 T}{2\sigma^2}\right) \leq \delta(2k+1)^{-d},$$

where the first inequality above uses the subgaussianity of $\epsilon(x)$, and the second inequality uses the definition of r^a. Now, the desired claim of the lemma follows from the fact that $|G| = (2k+1)^d$ and the union bound of the above events. ■

Note that, since by definition $\hat{f}(\hat{x}) \leq \hat{f}(\bar{x})$, we get the below upper bound of $f(\hat{x})$ on Γ,

$$f(\hat{x}) - f(x^\star_{f,c,r}) = \underbrace{f(\hat{x}) - \hat{f}(\hat{x})}_{\leq r^a} + \underbrace{\hat{f}(\hat{x}) - \hat{f}(\bar{x})}_{\leq 0} + \underbrace{\hat{f}(\bar{x}) - f(\bar{x})}_{\leq r^a} + \underbrace{f(\bar{x}) - f(x^\star_{f,c,r})}_{\leq \frac{Lr\sqrt{d}}{k}} \leq 2r^a + \frac{Lr\sqrt{d}}{k}.$$

Now, the assumption on k, T shows that, $T \geq 2k^2 \log(2k+1) r^{-2}$ and,

$$r^a = \sigma\left(\sqrt{\frac{2\log\frac{2}{\delta} + 2d\log(2k+1)}{T}}\right) \leq \frac{\sigma r}{k} \cdot \left(\sqrt{2\log\frac{2}{\delta} + d}\right)$$

This gives the desired claim of the proposition. ■

Motivated by Proposition 59, it becomes important to understand when such k exists in Eq (133) and how large it is.

Lemma 61 *Assume n is large enough satisfying $nr^2 \geq 3^{2d+2}$. Denote $k(r) = \left(\gamma\left(nr^2\right)\right)^{\frac{1}{d+2}}$ and $k = \left\lfloor \frac{1}{3}k(r) \right\rfloor$. Then $k^\star \in \mathbb{N}$, and $k^\star$ satisfies Eq (133).*

Proof Note that, $\gamma(x) \geq \sqrt{x}$ whenever $x \geq 3$. Thus, by assumption that $nr^2 \geq 3^{2(d+2)}$, we get that $\gamma(nr^2) \geq 3^{d+2}$ and hence $k^\star \geq 1$. To show that $k^\star$ satisfies Eq (133), note that, when $k = k^\star$, we have,

$$(2k+1)^d \left\lceil 2k^2 \log(2k+1) \right\rceil \leq (3k)^{d+2} \log(3k)^{d+2} \leq (k(r))^{d+2} \log(k(r))^{d+2} \leq nr^2,$$

where the last inequality follows from the fact that, for any $x > 0$, $\gamma(x)\log\gamma(x) \leq x$. ■

Proposition 59 and lemma 61 immediately give us the corollary below.

Corollary 62 *Given any fix $c \in \mathbb{R}^d$ and $r \in (0,1]$, set $k = \left\lfloor \frac{1}{3}(\gamma(nr^2))^{\frac{1}{d+2}} \right\rfloor$ and $T = \left\lfloor \frac{n}{2k+1} \right\rfloor$. Assume n is large enough satisfying $nr^2 \geq 3^{2d+2}$. Then if we denote $\hat{x}$ to be the output from algorithm 14 when we input (c, r, k, T) as the input parameters, we get,*

$$\mathbb{P}\left(f(\hat{x}) - f(x^\star_{f,c,r}) \leq \gamma^\star\right) \geq 1 - \delta,$$

where

$$\gamma^\star = 6r^{\frac{d}{d+2}} n^{-\frac{1}{d+2}} \log(nr^2)^{\frac{1}{d+2}} \cdot \left(2\sigma\sqrt{2\log\frac{2}{\delta} + d} + L\sqrt{d}\right).$$