
(Bandit) Convex Optimization with Biased Noisy Gradient Oracles

Xiaowei Hu¹

Prashanth L.A.²

András Györfy³

Csaba Szepesvári¹

¹ Department of Computing Science, University of Alberta

² Institute for Systems Research, University of Maryland

³ Department of Electrical and Electronic Engineering, Imperial College London

Abstract

Algorithms for bandit convex optimization and online learning often rely on constructing noisy gradient estimates, which are then used in appropriately adjusted first-order algorithms, replacing actual gradients. Depending on the properties of the function to be optimized and the nature of “noise” in the bandit feedback, the bias and variance of gradient estimates exhibit various tradeoffs. In this paper we propose a novel framework that replaces the specific gradient estimation methods with an abstract oracle. With the help of the new framework we unify previous works, reproducing their results in a clean and concise fashion, while, perhaps more importantly, the framework also allows us to formally show that to achieve the optimal root- n rate either the algorithms that use existing gradient estimators, or the proof techniques used to analyze them have to go beyond what exists today.

1 Introduction

Convex optimization is widely studied under different models concerning what can be observed about the objective function. These models range from accessing the full gradient to only observing samples from the objective function (cf. [27, 13, 18, 29, 17, 1, 2, 28, 3, 22, 24, 33, 34, 14, 5, 15]). In this paper, we present and analyze a novel framework for convex optimization with biased gradient oracles, which encompasses most of what has been done in the bandit framework where the algorithms observe noisy point-evaluations of the objective function and use these to construct gradient estimators.

In our framework, an optimization algorithm can query an oracle repeatedly to get a noisy and biased version of the gradient (or a subgradient for non-differentiable functions), where the algorithm querying the oracle also sets a parameter that controls the tradeoff between the bias and the variance of the gradient estimate.

Gradient oracles have been considered in the literature before: Several previous works assume that the accuracy requirements hold with probability one [10, 4, 13] or consider adversarial noise [31]. Gradient oracles with stochastic noise, which is central to our development, were also considered [21, 19, 16]; however, these papers assume that the bias and the variance are controlled separately, and consider the performance of special algorithms (in some cases in special setups).

The main feature of our model is that we allow stochastic noise, control of the bias and the variance, and we also consider lower bounds on the achievable error rates. Our gradient oracle model applies to several gradient estimation techniques extensively used in the literature, mostly for the case when the gradient is estimated only based on noisy observations of the objective function [22, 24, 33, 34, 14, 5, 15]. A particularly interesting application of our model is the widely studied bandit convex optimization problem, mentioned above, where most previous algorithms essentially use gradient estimates and first-order methods [29, 17, 1, 2, 28, 3, 18].

In this paper, we consider the optimization accuracy in both the stochastic and online bandit convex optimization (BCO) setting. We provide upper and lower bounds on the minimax optimization error (or regret) for several oracle models, which correspond to different ways of quantifying the bias-variance tradeoff of the gradient estimate. In particular, we provide matching upper and lower bounds for optimizing smooth, convex functions. We do not claim to invent methods for proving upper bounds, as the methods we use have been known previously for special cases (see the references above), but our main contribution lies in abstracting away the properties

of gradient estimation procedures, thereby unifying previous analysis, providing a concise summary and an explanation of differences between previous works. More importantly, our framework also allows to prove lower bounds for any algorithm that relies on gradient estimation oracles of the type our framework captures (earlier work of Chen [9] considered a related lower bound on the converge of the iterate instead of the function value).

Note that our oracle model does not capture the full strength of the gradient estimates used in previous work, but it fully describes the properties of the estimates that *so far have been used in their analysis*. As a consequence, our lower bounds show that the known minimax regret of $\sqrt{T}$ [7, 8, 32] of online and stochastic bandit convex optimization cannot be shown to hold for any algorithm that uses current gradient estimation procedures, unless the proof exploited finer properties of the gradient estimators than used in prior works. In particular, our lower bounds even invalidate the claimed (weaker) upper bound of Dekel et al. [11].

The organization of the paper is as follows: We introduce the oracle model in Section 2, upper and lower bounds are provided in Section 3, with applications to online and stochastic BCO in Sections 4 and 5. All proofs can be found in the extended version of the paper [20].

2 Problem Setup

Notation: Capital letters will denote random variables. For $i \leq j$ positive integers, we use the notation $a_{i:j}$ to denote the sequence $(a_i, a_{i+1}, \dots, a_j)$.

We let $\|\cdot\|$ denote some norm on $\mathbb{R}^d$, whose dual is denoted by $\|\cdot\|_*$. Let $\mathcal{K} \subset \mathbb{R}^d$ be a non-empty closed convex set. Given the function $f : \mathcal{K} \rightarrow \mathbb{R}$ which is differentiable¹ in $\mathcal{K}^\circ$,² f is said to be μ -strongly convex w.r.t. $\|\cdot\|$ ($\mu \geq 0$) if $\frac{\mu}{2}\|x - y\|^2 \leq \mathcal{D}_f(x, y) \doteq f(x) - f(y) - \langle \nabla f(y), x - y \rangle$, for all $x \in \mathcal{K}, y \in \mathcal{K}^\circ$. Similarly, f is μ -strongly convex w.r.t. a function $\mathcal{R}$ if $\frac{\mu}{2}D_{\mathcal{R}}(x, y) \leq \mathcal{D}_f(x, y)$ for all $x \in \mathcal{K}, y \in \mathcal{K}^\circ$, where $\mathcal{K}^\circ \subseteq \text{dom}(\mathcal{R})$ and $\mathcal{R}$ is differentiable over $\mathcal{K}^\circ$. A function f is L -smooth w.r.t. $\|\cdot\|$ for some $L > 0$ if $D_f(x, y) \leq \frac{L}{2}\|x - y\|^2$, for all $x \in \mathcal{K}, y \in \mathcal{K}^\circ$. This latter condition is equivalent to that ∇f is L -Lipschitz [27, Theorem 2.1.5]. We let $\mathcal{F}_{L,\mu,\mathcal{R}}(\mathcal{K})$ denote the class of functions that are μ -strongly convex w.r.t. $\mathcal{R}$ and L -smooth w.r.t. some norm $\|\cdot\|$ on the set $\mathcal{K}$. We also let $\mathcal{F}_{L,\mu}(\mathcal{K})$ be $\mathcal{F}_{L,\mu,\mathcal{R}}(\mathcal{K})$ with $\mathcal{R}(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. Then, the set of convex and L -smooth functions is $\mathcal{F}_{L,0}(\mathcal{K})$.

¹Here we assume the differentiability for simplicity of definition, actually it is easy to generalize to non-differentiable functions by using sub-gradient.

²For $A \subset \mathbb{R}^d$, A° denotes the interior of A .

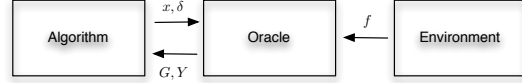


Figure 1: The Interaction of the algorithms with the gradient estimation oracle and the environment. For more information, see the text.

In this paper, we consider convex optimization in a novel setting, both for stochastic as well as online BCO. In the online BCO setting, the environment chooses a sequence of loss functions $f_1, \dots, f_n$ belonging to a set $\mathcal{F}$ of convex functions over a common, non-empty convex closed domain $\mathcal{K} \subset \mathbb{R}^d$. In the stochastic BCO setting, a single fixed loss function $f \in \mathcal{F}$ is chosen. An algorithm chooses a sequence of points $X_1, \dots, X_n \in \mathcal{K}$ in a serial fashion. The novelty of our setting is that the algorithm, upon selecting point X_t , receives a noisy and potentially biased estimate $G_t \in \mathbb{R}^d$ of the gradient of the loss function f (more generally, an estimate of a subgradient of f , in case f is not differentiable at X_t). To control the bias and the variance, the algorithm can choose a *tolerance parameter* $\delta_t > 0$ (in particular, we allow the algorithms to choose the tolerance parameter sequentially). A smaller δ_t results in a smaller “bias” (for the precise meaning of bias, we will consider two definitions below), while typically with a smaller δ_t , the “variance” of the gradient estimate increases. Notice that in the online BCO setting, the algorithm suffers the loss $f_t(Y_t)$ in round t , where $Y_t \in \mathcal{K}$ is guaranteed to be in the δ_t -vicinity of X_t . The goal in the online BCO setting is to keep the expected *regret*, $R_n = \mathbb{E}[\sum_{t=1}^n f_t(Y_t)] - \inf_{x \in \mathcal{K}} \sum_{t=1}^n f_t(x)$, small. In the stochastic BCO setting, the algorithm is also required to select a point $\hat{X}_n \in \mathcal{K}$ once the n th round is over (in both settings, n is given to the algorithms) and the algorithm’s performance is quantified using the *optimization error*, $\Delta_n = \mathbb{E}[f(\hat{X}_n)] - \inf_{x \in \mathcal{K}} f(x)$.

The main novelty of the model is that the information flow between the algorithm and the environment (holding f , or $f_{1:n}$) is mediated by a gradient estimation oracle. As we shall see, numerous existing approaches to online learning and optimization based on noisy pointwise information fit in this framework.

We will use two classes of oracles. In both cases, the oracles are specified by two functions $c_1, c_2 : [0, \infty) \rightarrow [0, \infty)$, which will be assumed to be continuous, monotonously increasing (resp., decreasing) with $\lim_{\delta \rightarrow 0} c_1(\delta) = 0$ and $\lim_{\delta \rightarrow 0} c_2(\delta) = +\infty$. Typical choices for c_1, c_2 are $c_1(\delta) = C_1 \delta^p$, $c_2(\delta) = C_2 \delta^{-q}$ with $p, q > 0$. Our type-I oracles are defined as follows:

³For simplicity, in some cases we allow f to be defined outside of $\mathcal{K}$ and allow Y_t to be in a small vicinity of $\mathcal{K}$.

Definition 1 ((c_1, c_2) type-I oracle) We say that γ is a (c_1, c_2) type-I oracle for $\mathcal{F}$, if for any function $f \in \mathcal{F}$, $x \in \mathcal{K}$, $0 < \delta \leq 1$, γ returns $G \in \mathbb{R}^d$ and $Y \in \mathcal{K}$ random elements such that $\|x - Y\| \leq \delta$ almost surely (a.s.),

1. $\|\mathbb{E}[G] - \nabla f(x)\|_* \leq c_1(\delta)$ (bias); and
2. $\mathbb{E}[\|G - \mathbb{E}[G]\|_*^2] \leq c_2(\delta)$. $\square$

The upper bound on δ is arbitrary: by changing the norm, any other value can also be accommodated. Also, the upper bound only matters when $\mathcal{K}$ is bounded and the functions in $\mathcal{F}$ are defined only in a small vicinity of $\mathcal{K}$.

The second type of oracles considered is as follows:

Definition 2 ((c_1, c_2) type-II oracle) We say that γ is a (c_1, c_2) type-II oracle for $\mathcal{F}$, if for any function $f \in \mathcal{F}$, $x \in \mathcal{K}$, $0 < \delta \leq 1$, γ returns $G \in \mathbb{R}^d$ and $Y \in \mathcal{K}$ random elements such that $\|x - Y\| \leq \delta$ a.s.,

1. There exists $\tilde{f} \in \mathcal{F}$ such that $\|\tilde{f} - f\|_\infty \leq c_1(\delta)$ and $\mathbb{E}[G] = \nabla \tilde{f}(x)$ (bias); and
2. $\mathbb{E}[\|G - \mathbb{E}[G]\|_*^2] \leq c_2(\delta)$ (variance). $\square$

Note that our definition allows the same oracle γ to respond to the same inputs (x, δ, f) with a differently constructed pair (e.g., to have memory), though most often the oracles constructed in practice will map the triples (x, δ, f) to a gradient estimate-point pair using a fixed stochastic mapping.⁴ Examples will be discussed in Section 4. We will denote the set of type-I (type-II) oracles satisfying the (c_1, c_2)-requirements given a function $f \in \mathcal{F}$ by $\Gamma_1(f, c_1, c_2)$ (resp., $\Gamma_2(f, c_1, c_2)$).

Here, we will study the minimax regret in the online convex optimization setting, while we study the minimax error in the stochastic convex optimization setting (sometimes, also called as the “simple regret”). Both are defined with respect to a class of loss functions $\mathcal{F}$, and the bias/variance control functions c_1, c_2 . The *minimax expected regret* for $(\mathcal{F}, c_1, c_2)$ with type-I oracles is

$$R_{\mathcal{F},n}^*(c_1, c_2) = \inf_{\mathcal{A}} \sup_{f_{1:n} \in \mathcal{F}^n} \sup_{\gamma_t \in \Gamma_1(f_t, c_1, c_2)} \sup_{1 \leq t \leq n} R_n^{\mathcal{A}},$$

where $\mathcal{A}$ ranges through all algorithms that interact with the loss sequence $f_{1:n} = (f_1, \dots, f_n)$ through the oracles $\gamma_{1:n}$ (in round t , oracle γ_t is used), and we use $R_n^{\mathcal{A}}$ to denote the expected regret of $\mathcal{A}$ (against $f_{1:n}, \gamma_{1:n}$). The minimax regret for type-II oracles is defined analogously.

⁴For oracles with memory, in the definition the expectation should be replaced with an expectation that is conditioned on the past.

In the stochastic BCO setting, the minimax error is defined through

$$\Delta_{\mathcal{F},n}^*(c_1, c_2) = \inf_{\mathcal{A}} \sup_{f \in \mathcal{F}} \sup_{\gamma \in \Gamma_1(f, c_1, c_2)} \Delta_n^{\mathcal{A}}(f, \gamma), \quad (1)$$

where, again, $\mathcal{A}$ ranges through all algorithms that interact with an oracle and $\Delta_n^{\mathcal{A}}(f, \gamma)$ is the optimization error that $\mathcal{A}$ suffers after n rounds of interaction with f through (a single) γ as described earlier. The minimax error for type-II oracles is defined analogously.

When the set $\mathcal{K}$ is bounded and the function set $\mathcal{F}$ is invariant to linear shifts, every (c_1, c_2) type-I oracle is also an (Rc_1, c_2) type-II oracle, where $R = \sup_{x \in \mathcal{K}} \|x\|$: simply consider $\tilde{f}(y) = f(y) + (\mathbb{E}[G] - \nabla f(x))^T y$, where G is the gradient estimate returned by the oracle. Thus, if for some set $\Delta_n^{\text{type-I}}, \Delta_n^{\text{type-II}}$ denote the appropriate minimax errors and $R = 1$ then $\Delta_{\mathcal{F},n}^{\text{type-I}}(c_1, c_2) \leq \Delta_{\mathcal{F},n}^{\text{type-II}}(c_1, c_2)$. As a result, when proving lower bounds, we shall consider type-I, while when proving upper bounds we will consider type-II oracles. Also note that for either type of oracles, $\Delta_{\mathcal{F},n}^*(c_1, c_2) \leq R_{\mathcal{F},n}^*(c_1, c_2)/n$. This follows by the well known construction that turns an online convex optimization method $\mathcal{A}$ for regret minimization into an optimization method by running the method and at the end choosing as $\hat{X}_n$ the average of the points $X_1, \dots, X_n$ queried by $\mathcal{A}$ during the n rounds. Indeed, then $f(\hat{X}_n) \leq \frac{1}{n} \sum_{t=1}^n f(X_t)$ by Jensen’s inequality, hence the average regret of $\mathcal{A}$ will upper bound the error of choosing $\hat{X}_n$ at the end. A consequence of this relation is that a lower bound for $\Delta_{\mathcal{F},n}^*(c_1, c_2)$ will also be a lower bound for $R_{\mathcal{F},n}^*(c_1, c_2)/n$ and an upper bound on $R_{\mathcal{F},n}^*(c_1, c_2)$ leads to an upper bound on $\Delta_{\mathcal{F},n}^*(c_1, c_2)$. This explains why we allowed taking supremum over time-varying oracles in the definition of the regret and why we used a static oracle for the optimization error: to maximize the strength of the bounds we obtain.

3 Main Results

In this section we provide our main results in forms of upper and lower bounds on the minimax error. First we give an upper bound by analyzing a mirror-descent algorithm given in Algorithm 1. In the algorithm, we assume that the regularizer function $\mathcal{R}$ is α -strongly convex and the target function f is smooth or smooth and strongly convex. A standard analysis yields the following upper bounds:

Theorem 1 (Upper bound) Consider the class $\mathcal{F}_{L,0}$ of convex, L -smooth functions on the bounded, convex domain $\mathcal{K} \neq \emptyset, \mathcal{K} \subset \mathbb{R}^d$. Assume that the regularization function $\mathcal{R}$ is α -strongly convex w.r.t. some norm $\|\cdot\|$, and $\mathcal{K}^\circ \subseteq \text{dom}(\mathcal{R})$. For any (c_1, c_2) type-II

Oracle type	Convex + Smooth		Strongly Convex + Smooth	
	Upper bound	Lower bound	Upper bound	Lower bound
δ -bias, δ^{-2} -variance ($p = 1, q = 2$)	$\left(\frac{C_1^2 C_2 D}{n}\right)^{1/4}$	$\left(\frac{C_1^2 C_2 d^2}{n}\right)^{1/4}$	$\left(\frac{C_1^2 C_2}{n}\right)^{1/3}$	$\left(\frac{C_1^2 C_2}{n}\right)^{1/2}$
δ^2 -bias, δ^{-2} -variance ($p = 2, q = 2$)	$\left(\frac{C_1 C_2 D}{n}\right)^{1/3}$	$\left(\frac{C_1 C_2 \sqrt{d^3}}{n}\right)^{1/3}$	$\left(\frac{C_1 C_2}{n}\right)^{1/2}$	$\left(\frac{C_1 C_2}{n}\right)^{2/3}$

Table 1: Summary of upper and lower bounds for different smooth function classes and gradient oracles for the settings of Theorem 1 and Theorem 2. Note that when $\mathcal{R}$ is the squared norm and $\mathcal{K}$ is the hypercube (as in the lower bounds), $D = \theta(d)$ in the upper bounds and also that C_1, C_2 may hide dimension-dependent quantities for the common gradient estimators, as will be discussed later.

Algorithm 1 Mirror Descent with Type-I/II Oracle.

Input: Closed convex set $\mathcal{K} \neq \emptyset$, regularization function $\mathcal{R} : \text{dom}(\mathcal{R}) \rightarrow \mathbb{R}$, $\mathcal{K}^\circ \subset \text{dom}(\mathcal{R})$, tolerance parameter δ , learning rates $\{\eta_t\}_{t=1}^{n-1}$.
Initialize $X_1 \in \mathcal{K}$ arbitrarily.

for $t = 1, 2, \dots, n-1$ **do**

 Query the oracle at X_t to receive G_t, Y_t .

 Set $X_{t+1} = \arg\min_{x \in \mathcal{K}} [\eta_t \langle G_t, x \rangle + D_{\mathcal{R}}(x, X_t)]$.

Return: $\hat{X}_n = \frac{1}{n} \sum_{t=1}^n X_t$.

oracle with $c_1(\delta) = C_1 \delta^p$, $c_2(\delta) = C_2 \delta^{-q}$, $p, q > 0$, if Algorithm 1 is run with $\eta_t = \alpha/(a_t + L)$ and $\delta = \left(\frac{C_2}{4aC_1} \frac{2p+q}{p} \frac{n}{n+1}\right)^{\frac{1}{p+q}} n^{-\frac{1}{2p+q}}$, where $a^{2p+q} = 2^{q-p} \left(2 + \frac{q}{p}\right)^p \left(1 + \frac{1}{n}\right)^q \left(\frac{\alpha}{D}\right)^{p+q} C_1^q C_2^p$, $a_t = at^{\frac{p+q}{2p+q}}$, for $t = 1, 2, \dots, n-1$, then the regret and the minimax error can be bounded as

$$\Delta_{\mathcal{F}_{L,0,n}}^*(c_1, c_2) \leq \frac{R_{\mathcal{F}_{L,0,n}}^*(c_1, c_2)}{n} = O\left(\left(\frac{DC_1^{\frac{q}{p}} C_2}{n}\right)^{\frac{p}{2p+q}}\right),$$

where $D = \sup_{x,y \in \mathcal{K}} D_{\mathcal{R}}(x, y)$.

For the class $\mathcal{F} = \mathcal{F}_{L,\mu,\mathcal{R}}$ of μ -strongly convex (w.r.t. $\mathcal{R}$) and L -smooth functions, with $\alpha > 2L/\mu$, $\eta_t = 2/(\mu t)$, and $\delta^{p+q} = \frac{C_2(\log n + 1 + \frac{\alpha\mu}{2L})}{2\alpha\mu C_1(n+1)}$, the regret and the minimax error satisfy

$$\begin{aligned} & \Delta_{\mathcal{F}_{L,\mu,\mathcal{R},n}}^*(c_1, c_2) \\ & \leq \frac{R_{\mathcal{F}_{L,\mu,\mathcal{R},n}}^*(c_1, c_2)}{n} = \tilde{O}\left(\left(\frac{C_1^{\frac{q}{p}} C_2}{n}\right)^{\frac{p}{p+q}}\right). \end{aligned}$$

In the bounds $O(\cdot)$ hides a constant that is a function of p, q, α, L and μ .⁵ $\square$

⁵In the second bound, $\tilde{O}(f(n))$ denotes $O(\log(n)f(n))$.

We next state lower bounds for both convex as well as strongly convex function classes. In particular, we observe that for convex and smooth functions the upper bound for the mirror descent scheme matches the lower bound, up to constants, whereas there is a gap for strongly convex+smooth functions. Filling the gap is left for future work.

Theorem 2 (Lower bound) Let $n > 0$ be an integer, $p, q > 0$, $C_1, C_2 > 0$, $\mathcal{K} \subset \mathbb{R}^d$ convex, closed, with $[+1, -1]^d \subset \mathcal{K}$. Then, for any algorithm that observes n random elements from a (c_1, c_2) type-I oracle with $c_1(\delta) = C_1 \delta^p$, $c_2(\delta) = C_2 \delta^{-q}$, the minimax error (and hence the regret) satisfies the following bounds:

$\mathcal{F}_{L,0}(\mathcal{K})$ (Convex and smooth) with $L \geq \frac{1}{2}$:

$$\Delta_{\mathcal{F}_{L,0,n}}^*(c_1, c_2) \geq K_1 \sqrt{d} \left(\frac{C_1^{\frac{q}{p}} C_2}{n}\right)^{\frac{p}{2p+q}},$$

where $K_1 = \frac{(2p+q)^2}{2q^{\frac{q}{2p+q}} (4p+q)^{\frac{4p+q}{2p+q}}}$.

$\mathcal{F}_{L,1}(\mathcal{K})$ (1-strongly convex and smooth) with $L \geq 1$:

$$\Delta_{\mathcal{F}_{L,1,n}}^*(c_1, c_2) \geq K_2 \left(\frac{C_1^{\frac{q}{p}} C_2}{n}\right)^{\frac{p}{p+q/2}},$$

where $K_2 = 2^{\frac{2p-q}{2p+q}} \frac{(2p+q)^3}{q^{\frac{2q}{2p+q}} (6p+q)^{\frac{6p+q}{2p+q}}}$. $\square$

By continuity, the above claim can be extended to cover the case of $q = 0$ (constant variance). For the special case of $p = 0$ and $C_1 > 0$, which implies a constant bias, it is possible to derive an $\Omega(1)$ lower bound by tweaking the proof. On the other hand, the case of $p = 0$ and $C_1 = 0$ (no bias) leads to an $\Omega(d/\sqrt{n})$ lower bound. The proof of the lower bound is obtained in the usual way by presenting a family of functions and a type-I oracle such that any algorithm suffers at least the stated error on one of the functions. In particular, for $\mathcal{F}_{L,0}$ with $L \geq 1/2$

we use $f_{v,\epsilon}(x) = \epsilon(x - v) + 2\epsilon^2 \ln(1 + e^{-\frac{x-v}{\epsilon}})$ with $v = \pm 1$, $\epsilon > 0$, and $x \in \mathcal{K} \subset \mathbb{R}$ for appropriate ϵ . Note that for any $\epsilon > 0$, $f_{v,\epsilon} \in \mathcal{F}_{1/2,0} \setminus \cup_{0 < \lambda < 1/2} \mathcal{F}_{\lambda,0}$.

Remark 1 (Scaling) For any function class $\mathcal{F}$, by the definition of the minimax error (1), it is easy to see that

$$\Delta_n^*(\mu\mathcal{F}, c_1, c_2) = \mu\Delta_n^*(\mathcal{F}, c_1/\mu, c_2/\mu^2),$$

where $\mu\mathcal{F}$ denotes the function class comprised of functions in $\mathcal{F}$, each scaled by $\mu > 0$. In particular, this relation implies that the bound for μ -strongly convex function class is only a constant factor away from the bound for 1-strongly convex function class. $\square$

Table 2 presents the upper and lower bounds for two specific choices of p and q (relevant in applications, as we shall see later). These bounds can be inferred from the results in Theorems 1 and 2. Some specific examples will be discussed in the next section.

4 Applications to Stochastic BCO

The main application of the biased noisy gradient oracle based convex optimization of the previous section is bandit convex optimization. We introduce here briefly the stochastic version of the problem, while online BCO will be considered in Section 5. Readers familiar with these problems and the associated gradient estimation techniques, may skip this description to jump directly to Theorem 4, and come back here only if clarifications are needed.

In the *stochastic BCO* setting, the algorithm sequentially chooses the points $X_1, \dots, X_n \in \mathcal{K}$ while observing the loss function at these points in noise. In particular, in round t , the algorithm chooses X_t based on the earlier observations $Z_1, \dots, Z_{t-1} \in \mathbb{R}$ and $X_1, \dots, X_{t-1}$, after which it observes Z_t , where Z_t is the value of $f(X_t)$ corrupted by “noise”.

Previous research considered several possible constraints connecting Z_t and $f(X_t)$. One simple assumption is that $\{Z_t - f(X_t)\}_t$ is an $\{\mathcal{F}_t\}_t = \{\sigma(X_{1:t}, Z_{1:t-1})\}_t$ -adapted martingale difference sequence (with favorable tail properties). A specific case is when $Z_t - f(X_t) = \xi_t$, where (ξ_t) is a sequence of independent and identically distributed (i.i.d.) variables. A stronger assumption, common in stochastic programming, is that

$$Z_t = F(X_t, \Psi_t), \quad f(x) = \int F(x, \psi) P_\Psi(d\psi), \quad (2)$$

where $\Psi_t \in \mathbb{R}$ is chosen by the algorithm and in particular the algorithm can draw Ψ_t at random from P_Ψ . As in [15], we assume that the function $F(\cdot, \psi)$ is L_ψ -smooth P_Ψ -a.s. and the quantity $\bar{L}_\Psi = \sqrt{\mathbb{E}[L_\Psi^2]}$ is finite. Note that the algorithm is aware of P_Ψ , but does

not know how different values of ψ affect the noise $\xi(x, \psi) = F(x, \psi) - f(x)$. Nevertheless, as the algorithm can control ψ and thus ξ , we refer to this as *controlled noise* setting and to the others as the case of *uncontrolled noise*. As we will see, and is well known in the simulation optimization literature [23, 15], this extra structure allows the algorithm to reduce the variance of the noise of its gradient estimates by reusing the same Ψ_t in consecutive measurements, while measuring the gradient at the same point, an instance of the method of the method of common random variables. As creating an estimate from K points (which is equivalent to the so-called “multi-point feedback setup” from the literature where K points are queried in each round) changes the number of rounds from n to n/K , which does not change the convergence rate as long as K is fixed.

4.1 Estimating the Gradient

A common popular idea in bandit convex optimization is to use the bandit feedback to construct noisy (and biased) estimates of the gradient. In the following, we provide a few examples for oracles that construct gradient estimates for function classes that are increasingly general: from smooth, convex to non-differentiable functions.

One-point feedback Given $x \in \mathcal{K}$, $0 < \delta \leq 1$, common gradient estimates that are based on a single query to the function evaluation oracle (the so-called “one-point feedback”) take the form

$$G = \frac{Z}{\delta} V, \text{ where } Z = f(x + \delta U) + \xi, \quad (3)$$

where $(U, V, \xi) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ are jointly distributed random variables, ξ is the function evaluation noise whose distribution may depend on $x + \delta U$ but $\mathbb{E}[\xi|V] = 0$, and G is the estimate of $\nabla f(x)$ ($f : \mathcal{K} \rightarrow \mathbb{R}$).

In all oracle constructions we will use the following assumption:

Assumption 1 Let $\mathcal{K} \subset \mathcal{D}^\circ \subset \mathbb{R}^d$, where $f : \mathcal{D} \rightarrow \mathbb{R}$. For any $x \in \mathcal{K}$, $x + \delta U \in \mathcal{D}$ a.s., and $\mathbb{E}[\|V\|_*^2], \mathbb{E}[\|U\|^3] < +\infty$.

Note that here the function domain $\mathcal{D}$ can be larger than or equal to the set $\mathcal{K}$, where the algorithm chooses x . This is to ensure that the oracle will not receive invalid inputs, that is, queries where f is not defined. When the functions are defined over $\mathcal{K}$ only and $\mathcal{K}$ is bounded, the above constructions only work for δ small enough. In this case, the best approach perhaps is to use Dikin ellipsoids to construct the oracles, as done by Hazan and Levy [18].

The next proposition, whose proof is based on ideas from Spall [33] shows that the above one-point gradient esti-

mator leads to a type-I (and, hence, also type-II) oracle.

Proposition 1 *Let Assumption 1 hold and let γ be the one-point feedback oracle defined in (3). Assume further that U is symmetrically distributed, $V = h(U)$, where $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is an odd function, $\mathbb{E}[V] = 0$, and $\mathbb{E}[VU^\top] = I$. Then, in the uncontrolled noise case, γ is a $(c_1(\delta), c_2(\delta))$ type-I oracle given in Table 2, where $C_2 = 4\mathbb{E}[\|V\|_*^2] (\text{ess sup } \mathbb{E}[\xi^2|V] + \sup_{x \in \mathcal{D}} f^2(x))$, and $C_1 = \frac{L}{2}\mathbb{E}[\|V\|_*\|U\|^2]$ when $f \in \mathcal{F}_{L,0}$ and $C_1 = \frac{B_3}{6}\mathbb{E}[\|V\|_*\|U\|^3]$ for $f \in \mathcal{C}^3$ where $B_3 = \sup_{x \in \mathcal{D}} \|\nabla^3 f(x)\|_T$ where $\|\cdot\|_T$ denotes the implied norm for rank-3 tensors.*

Another possibility is to use the so-called smoothing technique [29, 17, 18] to obtain type-II oracles. Following the analysis in Flaxman et al. [17], one gets the following result, which improves the bias of the previous result from $O(\delta)$ to $O(\delta^2)$ in the smooth+convex case:

Proposition 2 *Let Assumption 1 hold and let γ be the one-point feedback oracle defined in (3). Define $V = n_W(U) \frac{|\partial W|}{|W|}$, where $W \subset \mathbb{R}^n$ is a convex body with boundary ∂W , U is uniformly distributed on ∂W , $n_W(U)$ denotes the normal vector of ∂W at U , and $|\cdot|$ denotes the appropriate volume. Let $C_2 > 0$ be defined as in Proposition 1. Then, if f is L_0 -Lipschitz, γ is a type-II oracle with $c_1(\delta) = C_1\delta$, $c_2(\delta) = C_2/\delta^2$ where $C_1 = L_0 \sup_{w \in W} \|w\|$. Further, assuming W is symmetric w.r.t. the origin, if f is L -smooth, then γ is a type-I (and type-II oracle) with $c_1(\delta) = C_1\delta^2$, $c_2(\delta) = C_2/\delta^2$ where $C_1 = (L/|W|) \int_W \|w\|^2 dw$, and, if in addition f is also convex (i.e., $f \in \mathcal{F}_{L,0}$) then γ is a type-I oracle with $c_1(\delta) = C_1\delta^2/2$ and $c_2(\delta) = C_2/\delta^2$. $\square$*

Note that the improvement did not even require convexity. Also, the bias is smaller for smoother functions, a property that will be enjoyed by all the gradient estimators.

Two-point feedback While the one-point estimators are intriguing, in the optimization setting one can also always group two consecutive observations and obtain similar smoothing-type estimates at the price of reducing the number of rounds by a factor of two only, which does not change the rate of convergence. Next we present an oracle that uses two function evaluations to obtain a gradient estimate. As will be discussed later, this oracle encompasses several simultaneous perturbation methods (see 5): Given the inputs $x \in \mathcal{K}$, $0 < \delta \leq 1$, the gradient estimate is

$$G = \frac{Z^+ - Z^-}{2\delta} V, \quad (4)$$

Noise $\rightarrow$ Function $\downarrow$	Controlled (see (2))	Uncontrolled (see (5))
Convex + Smooth	$(C_1\delta, C_2)$	Props 1,3: $(C_1\delta, \frac{C_2}{\delta^2})$ Prop 2: $(C_1'\delta^2, \frac{C_2}{\delta^2})$
$f \in \mathcal{C}^3$	$(C_1\delta^2, \frac{C_2}{\delta^2})$	Props 1,3: $(C_1\delta^2, \frac{C_2}{\delta^2})$

Table 2: Gradient oracles for different function classes and noise categories. The constants C_1, C_1', C_2 are defined in Propositions 1–3.

where $Z^\pm = f(X^\pm) + \xi^\pm$, $X^\pm = x \pm \delta U$, $U, V \in \mathbb{R}^d$, $\xi^\pm \in \mathbb{R}$ are random, jointly distributed random variables, U, V chosen by the oracle in the uncontrolled case and chosen by the algorithm in the controlled case from some fixed distribution characterizing the oracle, and $\xi^\pm$ being the noise of the returned feedback $Z^\pm$ at points $X^\pm$. For the following proposition we consider $4 = 2 \times 2$ cases. First, the function is either assumed to be L -smooth and convex (i.e., the derivative of f is L -Lipschitz w.r.t. $\|\cdot\|_*$), or it is assumed to be three times continuously differentiable ($f \in \mathcal{C}^3$). The other two options are either the controlled noise setting of (2), or, in the uncontrolled case, we make the alternate assumptions

$$\begin{aligned} \mathbb{E}[\xi^+ - \xi^- | U, V] &= 0 \quad \text{and} \\ \mathbb{E}[(\xi^+ - \xi^-)^2 | V] &\leq \sigma_\xi^2 < \infty. \end{aligned} \quad (5)$$

The following proposition, whose proof is based on [33, Lemma 1] and [15, Lemma 1], provides conditions under which the bias-variance parameters (c_1, c_2) can be bounded as shown in Table 2:

Proposition 3 *Let Assumption 1 hold and let γ be a two-point feedback oracle defined by (4). Suppose furthermore that $\mathbb{E}[VU^\top] = I$. Then γ is a type-I oracle with the pair $(c_1(\delta), c_2(\delta))$ given by Table 2. For uncontrolled noise and for controlled noise with $f \in \mathcal{C}^3$, C_1 is as in Proposition 1 and C_2 is $4C_2$ from Proposition 1. For the controlled noise case with $f \in \mathcal{F}_{L,0}$, $C_1 = \frac{\bar{L}_\Psi}{2}\mathbb{E}[\|V\|_*\|U\|^2]$ and $C_2 = 2B_1^2 + \frac{\bar{L}_\Psi^2}{2}\mathbb{E}[\|V\|_*^2\|U\|^4]$, with $B_1 = \sup_{x \in \mathcal{K}} \|\nabla f(x)\|_*$. $\square$*

Popular choices for U and V :

- If we set U_i to be independent, symmetric ± 1 -valued random variables and $V_i = 1/U_i$, then we recover the popular SPSA scheme proposed by Spall [33]. It is easy to see that $\mathbb{E}[VU^\top] = I$ holds in this case. When the norm $\|\cdot\|$ is the 2-norm, $C_1 = O(d^2)$ and $C_2 = O(d)$. If we set $\|\cdot\|$ to be the max-norm, $C_1 = O(\sqrt{d})$ and $C_2 = O(d)$.
- If we set $V = U$ with U chosen uniform at random on the surface of a sphere with radius $\sqrt{d}$, then we recover the RDSA scheme proposed by Kushner and Clark

[24, pp. 58–60]. In particular, the (U_i) are identically distributed with $\mathbb{E}[U_i U_j] = 0$ if $i \neq j$ and $\mathbb{E}[U^T U] = d$, hence $\mathbb{E}[U_i^2] = 1$. Thus, if we choose $\|\cdot\|$ to be the 2-norm, $C_1 = O(d^2)$ and $C_2 = O(d)$.

- If we set $V = U$ with U the standard d -dimensional Gaussian with unit covariance matrix, we recover the smoothed functional (SF) scheme proposed by Katkovnik and Kulchitsky [22]. Indeed, in this case, by definition, $\mathbb{E}[V U^T] = \mathbb{E}[U U^T] = I$. When $\|\cdot\|$ is the 2-norm, $C_1 = O(d^2)$ and $C_2 = O(d)$. This scheme can also be interpreted as a smoothing operation that convolves the gradient of f with a Gaussian density.

4.2 Achievable Results for Stochastic BCO

We now consider stochastic BCO with L -smooth functions over a convex, closed non-empty domain $\mathcal{K}$. Let $\mathcal{F}$ denote the set of these functions. Duchi et al. [15] proves that the minimax expected optimization error for the functions $\mathcal{F}$ with uncontrolled noise is lower bounded by $\Omega(n^{-1/2})$. They also give an algorithm which uses two-point gradient estimates which matches this lower bound for the case of *controlled noise*. For controlled noise, the constructions in the previous section give that for two-point estimators $c_1(\delta) = C_1 \delta^p$ and $c_2(\delta) = C_2 \delta^{-q}$ with $p = 1$ and $q = 0$. Plugging this into Theorem 1 we get the rate $O(n^{-1/2})$ (which is unsurprising given that the algorithms and the upper bound proof techniques are essentially the same as that of Duchi et al. [15]). However, when the noise is uncontrolled, the best that we get is $p = 2$ and $q = 2$. From Theorem 2 we get that with such oracles, no algorithm can get better rate than $\Omega(n^{-1/3})$, while from Theorem 1 we get that these rates are matched by mirror descent. We can summarize these findings as follows:

Theorem 3 Consider $\mathcal{F}_{L,0}$, the space of convex, L -smooth functions over a convex, closed non-empty domain $\mathcal{K}$. Then, we have the following:

Uncontrolled noise: Take any (δ^2, δ^{-2}) type-I oracle γ . There exists an algorithm that uses γ and achieves the rate $O(n^{-1/3})$. Furthermore, no algorithm using γ can achieve better error than $\Omega(n^{-1/3})$ for every (δ^2, δ^{-2}) type-I oracle γ .

Controlled noise: Take any $(\delta, 1)$ type-I oracle γ . There exists an algorithm that uses γ and achieves the rate $O(n^{-1/2})$. Furthermore, no algorithm using γ can achieve better error than $\Omega(n^{-1/2})$ for every $(\delta, 1)$ type-I oracle γ . $\square$

For stochastic BCO with uncontrolled noise, Agarwal et al. [3] analyze a variant of the well-known ellipsoid method and provide regret bounds for the case of convex, 1-Lipschitz functions over the unit ball. Their regret bound implies a minimax error (1) bound of or-

der $O(\sqrt{d^{32}/n})$. Liang et al. [25] provide an algorithm based on random walks (and not using gradient estimates) for the setting of convex, bounded functions whose domain is contained in the unit cube and their algorithm results in a bound of the order $O((d^{14}/n)^{1/2})$ for the minimax error. These bounds decrease faster in n than the bound available in Theorem 3, while showing a much worse dependence on the dimension. However, what is more interesting is that our results also shows that an $O(n^{-1/2})$ upper bound *cannot* be achieved solely based on the oracle properties of the gradient estimates considered. Since the analysis of all gradient algorithms for stochastic BCO does this, it is no wonder that the best known upper bound for convex+smooth functions is $O(n^{-1/3})$ [30]. (We will comment on the recent paper of Dekel et al. [11] later.)

The above result also shows that the gradient oracle based algorithms are optimal for smooth problems, under a controlled noise setting. While Duchi et al. [15] suggests that it is the power of two-point gradient estimators that helps to achieve this, we need to add that having controlled noise is also critical.

Finally, let us make some remarks on the early literature on this problem. A finite time lower bound for stochastic, smooth BCO is presented by Chen [9] for convex functions on the real line. When applied to our setting in the uncontrolled noise case, his results imply that $\mathbb{E}[|\hat{X}_n - x^*|]$, that is, the distance of the estimate to the optimum, is at least $\Omega(n^{-1/3})$. Note that this is larger than the error achieved by the algorithms of Liang et al. [25], Bubeck et al. [7], Bubeck and Eldan [8], but the apparent contradiction is easily resolved by noticing the difference in their error measure: distance to the optimum vs. error in the function value (in particular, compressing the range of functions makes locating the minimizer harder). Polyak and Tsybakov [29], who also considered distance to optimum, proved that mirror descent with gradient estimation achieves asymptotically optimal rates for functions that enjoy high order smoothness.

5 Applications to Online BCO

In the *online BCO* setting a learner sequentially chooses the points $X_1, \dots, X_n \in \mathcal{K}$ while observing the losses $f_1(X_1), \dots, f_n(X_n)$. More specifically, in round t , having observed $f_1(X_1), \dots, f_{t-1}(X_{t-1})$ of the previous rounds, the learner chooses $X_t \in \mathcal{K}$, after which it observes $f_t(X_t)$. The learner's goal is to minimize its expected regret $\mathbb{E}[\sum_{t=1}^n f_t(X_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^n f_t(x)]$. This problem is also called online convex optimization with one-point feedback. A slightly different problem is obtained if we allow the learner to choose multiple

points in every round, at which points the function f_t is observed. The loss is suffered at X_t . The points where the function is observed (“observation points” for short) may or may not be tied to X_t . One possibility is that X_t is one of the observation points. Another possibility is that X_t is the average of the observation points (e.g., Agarwal et al. [2]). Yet another possibility is that there is no relationship between them.

The oracle constructions from the previous section also apply to the online BCO setting where the algorithm is evaluated at Y_t , though in this case one cannot employ two-point feedback as the functions change between rounds. This also rules out the controlled noise case. Thus, for the online BCO setting, one should consider type-I (and II) oracles with $c_1(\delta) = C_1\delta^p$ and $c_2(\delta) = C_2\delta^{-q}$ with $p = q = 2$. For these type of oracles, the results from Theorem 2 give the following result:

Theorem 4 *Let $\mathcal{F}_{L,0}$ be the space of convex, L -smooth functions over a convex non-empty domain $\mathcal{K}$. No algorithm that relies on (δ^2, δ^{-2}) type-I oracles can achieve better regret than $\Omega(n^{2/3})$.* $\square$

With a noisy gradient oracle of Proposition 2, Theorem 4 implies that this regret rate is achievable, essentially recovering, and in some sense proving optimality of the result of Saha and Tewari [30]:

Theorem 5 *For zeroth order noisy optimization with smooth convex functions, the gradient estimator of Proposition 2 together with mirror descent (see Algorithm 1) achieve $\mathcal{O}(n^{2/3})$ regret.* $\square$

This optimality result shows that with the usual analysis of the current gradient estimation techniques, no gradient method can achieve the optimal regret $\mathcal{O}(n^{1/2})$ for online bandit convex optimization, established by Bubeck et al. [7], Bubeck and Eldan [8]. Note that this shows a contradiction to the recent result of Dekel et al. [11], who claimed to achieve $\tilde{\mathcal{O}}(n^{5/8})$ regret with the same (δ^2, δ^{-2}) type-II gradient oracle as Saha and Tewari [30], but their proof only used the (δ^2, δ^{-2}) tradeoff in the bias and variance properties of the oracle.

6 Related Work

Gradient oracle models have been studied in a number of previous papers [10, 4, 31, 13]. A full comparison between these oracle models is given by Devolder et al. [13]. For illustration, here we only review the model of this latter paper as a typical example of these previous works. The model of Devolder et al. [13] assumes a first-order approximation to the function with parameters (δ, L) . In particular, given (x, δ, L) and the convex function f , the oracle gives a pair $(t, g) \in \mathbb{R} \times \mathbb{R}^d$ such that

$t + \langle g, \cdot - x \rangle$ is a linear lower approximation to $f(\cdot)$ in the sense that $0 \leq f(y) - \{t + \langle g, y - x \rangle\} \leq \frac{L}{2} \|y - x\|^2 + \delta$. Devolder et al. [13] argue that this notion appears naturally in several optimization problems and study whether the so-called accelerated gradient techniques are still superior to their non-accelerated counterparts (and find a negative answer). The authors study both lower and upper rates of convergence, similarly to our paper. A major difference between the previous and our settings is that we allow stochastic noise (and bias), which the algorithms can control, while the oracle in these previous paper must guarantee that the accuracy requirements hold in each time step with probability one. This is a much stronger requirement, which may be impossible to satisfy in some problems, such as when the only information available about the functions is noise contaminated.

Some works, such as Schmidt et al. [31] allow arbitrary sequences of errors and show error bounds as a function of the accumulated errors. Our proof technique is actually essentially the same (as can be expected). However, the noisy case requires special care. For example, Proposition 3 of Schmidt et al. [31] bounds the optimization error for the smooth, convex case by $\mathcal{O}(1/n^2(\|x_1 - x^*\|^2 + A_n^2))$ where $A_n = \mathcal{O}(\sum_{t=1}^n t\|e_t\|)$, e_t being the error of the approximate gradient. This expression becomes $\Theta(\frac{1}{n^2} \sum_{t=1}^n t^2) \approx n$ assuming that errors’ noise level is a positive constant (in all our result, this holds). This clearly shows that the noisy case requires (some-what) special treatment.

Similar, but simpler noisy oracle models were introduced [21, 19, 16], but these models lack the bias-variance tradeoff central to this paper (i.e., they assume the variance and bias can be controlled independently of each other). The results in these papers are upper bounds on the error of certain gradient methods (also to some very specific problem for [19]), and they correspond to the bounds we obtained with $q = 0$.

7 Conclusions

We presented a novel noisy gradient oracle model for convex optimization. The oracle model covers several gradient estimation methods in the literature designed for algorithms that can observe only noisy function values, while allowing to handle explicitly the bias-variance tradeoff of these estimators. The framework allows to derive sharp upper and lower bounds on the minimax optimization error and the regret in the online case. From our lower bounds it follows that the current state of the art in designing and analyzing noisy gradient methods for stochastic and online smooth bandit convex optimization are suboptimal.

Acknowledgements

This work was supported by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning and RLAI, NSERC, the National Science Foundation (NSF) under Grants CMMI-1434419, CNS-1446665, and CMMI-1362303, and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-15-10050.

References

- [1] Abernethy, J., Hazan, E., and Rakhlin, A. (2008). Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pages 263–274.
- [2] Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT*, pages 28–40.
- [3] Agarwal, A., Foster, D. P., Hsu, D., Kakade, S. M., and Rakhlin, A. (2013). Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240.
- [4] Baes, M. (2009). Estimate sequence methods: Extensions and approximations. Technical report, IFOR Internal report, ETH Zurich, Switzerland.
- [5] Bhatnagar, S., Prasad, H. L., and Prashanth, L. A. (2013). *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods (Lecture Notes in Control and Information Sciences)*, volume 434. Springer.
- [6] Bubeck, S. (2014). Theory of convex optimization for machine learning. Technical report, Microsoft Research.
- [7] Bubeck, S., Dekel, O., Koren, T., and Peres, Y. (2015). Bandit convex optimization: $o(\sqrt{T})$ regret in one dimension. In *COLT*, pages 266–278.
- [8] Bubeck, S. and Eldan, R. (2015). Multi-scale exploration of convex functions and bandit convex optimization. Technical report, Microsoft Research.
- [9] Chen, H. (1988). Lower rate of convergence for locating a maximum of a function. *The Annals of Statistics*, 16(3):1330–1334.
- [10] d’Aspremont, A. (2008). Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19:1171–1183.
- [11] Dekel, O., Eldan, R., and Koren, T. (2015). Bandit smooth convex optimization: Improving the bias-variance tradeoff. In *NIPS*, pages 2926–2934.
- [12] Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. (2012). Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1):165–202.
- [13] Devolder, O., Glineur, F., and Nesterov, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75.
- [14] Dippon, J. (2003). Accelerated randomized stochastic optimization. *The Annals of Statistics*, 31(4):1260–1281.
- [15] Duchi, J. C., Jordan, M., Wainwright, M. J., and Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806.
- [16] Dvurechensky, P. and Gasnikov, A. (2015). Stochastic intermediate gradient method: Convex and strongly convex cases. arXiv:1411.2876.
- [17] Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *SODA*, pages 385–394.
- [18] Hazan, E. and Levy, K. (2014). Bandit convex optimization: Towards tight bounds. In *NIPS*, pages 784–792.
- [19] Honorio, J. (2012). Convergence rates of biased stochastic optimization for learning sparse ising models. In *ICML*, pages 257–264, New York, NY, USA. Omnipress.
- [20] Hu, X., L.A., P., György, A., and Szepesvári, C. (2016). (Bandit) convex optimization with biased noisy gradient oracles. arXiv.
- [21] Juditsky, A. and Nemirovski, A. (2011). First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. In Sra, S., Nowozin, S., and Wright, S., editors, *Optimization for Machine Learning*, pages 121–147. MIT press.
- [22] Katkovnik, V. Y. and Kulchitsky, Y. (1972). Convergence of a class of random search algorithms. *Automation Remote Control*, 8:1321–1326.
- [23] Kleinman, N. L., Spall, J. C., and Naiman, D. Q. (1999). Simulation-based optimization with stochastic approximation using common random numbers. *Management Science*, 45(11):1570–1578.
- [24] Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer Verlag, New York.
- [25] Liang, T., Narayanan, H., and Rakhlin, A. (2014). On zeroth-order stochastic convex optimization via random walks. arXiv preprint 1402.2667.

- [26] Mahdavi, M. (2014). *Exploiting Smoothness in Statistical Learning, Sequential Prediction, and Stochastic Optimization*. PhD thesis, Michigan State University.
- [27] Nesterov, Y. (2004). *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media.
- [28] Nesterov, Y. and Spokoiny, V. (2011). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pages 1–40.
- [29] Polyak, B. and Tsybakov, A. (1990). Optimal orders of accuracy for search algorithms of stochastic optimization. *Problems in Information Transmission*, pages 126–133.
- [30] Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *AISTATS*, pages 636–642.
- [31] Schmidt, M. W., Roux, N. L., and Bach, F. R. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, pages 1458–1466.
- [32] Shamir, O. (2012). On the complexity of bandit and derivative-free stochastic convex optimization. In *COLT*.
- [33] Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341.
- [34] Spall, J. C. (1997). A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112.
- [35] Yao, A. C. C. (1977). Probabilistic computations: Toward a unified measure of complexity. In *FOCS*, pages 222–227.

A Reduction of Oracles

As is mentioned in Section 2, a Type-I oracle is also a Type-II oracle. But it does not hold in the opposite direction. The following result formalizes the relationship between Type-I and II oracles.

Theorem 6 *Definition 1 is a sufficient condition of Definition 2, given a bounded $\mathcal{K}$. That is, for any (c_1, c_2) Type-I oracle γ , there exist some constant R such that γ is a (c'_1, c_2) Type-II oracle, with $c'_1(\delta) = Rc_1(\delta)$.*

Similarly, Definition 2 is sufficient for Definition 1 if the alternative condition

$$\sup_{x \in \mathcal{K}} \|\nabla \tilde{f} - \nabla f\|_* \leq c_1(\delta) \quad (6)$$

holds. □

PROOF Given $\|\mathbb{E}[G] - \nabla f(x)\|_* \leq c_1(\delta)$ in Definition 1, we can always construct a function $\tilde{f} : \mathcal{K} \rightarrow \mathbb{R}$,

$$\tilde{f}(y) = \mathbb{E} \left[f(y) + (G - \nabla f(x))^\top y \right],$$

where the expectation is over the randomness of G , satisfying

$$\|\tilde{f}(y) - f(y)\|_\infty = \|\mathbb{E}[(G - \nabla f(x))^\top y]\|_\infty \leq \|\mathbb{E}[G] - \nabla f(x)\|_* \|y\| \leq Rc_1(\delta),$$

by choosing $R = \sup_{y \in \mathcal{K}} \|y\|$.

Also notice that

$$\nabla \tilde{f}(x) = \mathbb{E}[\nabla f(x) + (G - \nabla f(x))] = \mathbb{E}[G].$$

Therefore, γ is also a (c'_1, c_2) Type-II oracle.

Next, given (6) and $\mathbb{E}[G] = \nabla \tilde{f}(x)$ in Definition 2, we can immediately get the bias condition of Definition 1. ■

B Gradient Estimation

Proof of Proposition 1

PROOF

Case 1 ($f \in \mathcal{C}^3$):

We use the proof technique of [34]. We start by bounding the bias. Since by assumption $\mathbb{E}[\xi|V] = 0$, we have

$$\mathbb{E} \left[V \left(\frac{\xi}{\delta} \right) \right] = 0,$$

implying that

$$\mathbb{E}[G] = \mathbb{E} \left[V \left(\frac{f(x + \delta U)}{\delta} \right) \right].$$

By Taylor's theorem, we obtain, a.s.,

$$f(x + \delta U) = f(x) + \delta U^\top \nabla f(x) + \frac{\delta^2}{2} U^\top \nabla^2 f(x) U + \frac{\delta^3}{2} R^+(x, \delta, U) (U, U, U),$$

where

$$R^+(x, \delta, U) = \int_0^1 \nabla^3 f(x + s\delta U) (1-s)^2 ds. \quad (7)$$

In the above, $\nabla^3 f(\cdot)$ is considered as a rank-3 tensor. Letting $B_3 = \sup_{x \in D} \|\nabla^3 f(x)\|$,⁶ we have $\|R^+(x, \delta, U)\| \leq B_3/3$ a.s. Now,

$$\begin{aligned} \mathbb{E} \left[V \frac{f(x + \delta U)}{\delta} \right] &= \mathbb{E} \left[V \frac{f(x)}{\delta} \right] + \mathbb{E} [V U^\top \nabla f(x)] + \mathbb{E} \left[\frac{\delta}{2} V U^\top \nabla^2 f(x) U \right] \\ &\quad + \mathbb{E} \left[\frac{\delta^2}{2} V R^+(x, \delta, U) (U \otimes U \otimes U) \right] \\ &= \nabla f(x) + \mathbb{E} \left[\frac{\delta^2}{2} V R^+(x, \delta, U) (U \otimes U \otimes U) \right]. \end{aligned}$$

The final equality above follows from the facts that $\mathbb{E}[V] = 0$, $\mathbb{E}[V U^\top] = I$ and for any $i, j = 1, \dots, d$, $\mathbb{E}[V_i U_j^2] = 0$ since V is a deterministic odd function of U , with U having a symmetric distribution. Using the fact that $|R^+(x, \delta, U)(U \otimes U \otimes U)| \leq \|R^+(x, \delta, U)\| \|U\|^3$, we obtain

$$\|\mathbb{E}[G] - \nabla f(x)\|_* \leq C_1 \delta^2,$$

where $C_1 = \frac{B_3 \mathbb{E}[\|V\|_* \|U\|^3]}{6}$.

Let us now bound the variance of G : Using the identity $\mathbb{E} \|X - E[X]\|^2 \leq 4\mathbb{E} \|X\|^2$, which holds for any random variable X ,⁷ we bound $\mathbb{E} \|G - \mathbb{E}G\|_*^2$ as follows:

$$\begin{aligned} \mathbb{E} \|G - \mathbb{E}G\|_*^2 &\leq 4\mathbb{E} \|G\|_*^2 \\ &= 4\mathbb{E} \left(\|V\|_*^2 \left(\left(\frac{\xi}{\delta} \right)^2 + 2 \left(\frac{\xi}{\delta} \right) \left(\frac{f(x + \delta U)}{\delta} \right) + \left(\frac{f(x + \delta U)}{\delta} \right)^2 \right) \right) \\ &= 4\mathbb{E} \left(\|V\|_*^2 \left(\frac{\xi}{\delta} \right)^2 \right) + 4\mathbb{E} \left(\|V\|_*^2 \right) \left(\frac{f(x + \delta U)}{\delta} \right)^2 \\ &\leq \frac{C_2}{\delta^2}, \end{aligned} \tag{8}$$

where $C_2 = 4\mathbb{E} [\|V\|_*^2] (\sigma_\xi^2 + B_0^2)$, where $\sigma_\xi^2 = \text{ess sup } \mathbb{E} [\xi^2 | V]$ and $B_0 = \sup_{x \in D} f(x)$. The equality in (8) follows from $\mathbb{E} [\xi | V] = 0$.

Therefore, for $f \in \mathcal{C}^3$, γ defined by (3) is a $(C_1 \delta^2, C_2/\delta^2)$ type-I oracle.

Case 2 (f is convex and L -smooth):

Since f is convex and L -smooth, for any $0 < \delta < 1$,

$$0 \leq \frac{f(x + \delta u) - f(x)}{\delta} - \langle \nabla f(x), u \rangle \leq \frac{L\delta \|u\|^2}{2}.$$

Denote $\phi(x, \delta, u) := \frac{f(x + \delta u) - f(x)}{\delta} - \langle \nabla f(x), u \rangle$, we have $|\phi(x, \delta, u)| \leq \frac{L\delta}{2} \|u\|^2$. Then, given $\mathbb{E}[V U^\top] = I$, $\mathbb{E}[V] = 0$, we obtain

$$\begin{aligned} \|\mathbb{E}[G] - \nabla f(x)\|_* &= \left\| \mathbb{E} \left[\frac{f(x + \delta U)}{\delta} V \right] - \mathbb{E}[V U^\top \nabla f(x)] \right\|_* \\ &= \left\| \mathbb{E} \left[V \left(\frac{f(x + \delta U)}{\delta} - U^\top \nabla f(x) \right) \right] \right\|_* \\ &= \left\| \mathbb{E} \left[V \left(\phi(x, \delta, U) + \frac{f(x)}{\delta} \right) \right] \right\|_* \\ &= \|\mathbb{E}[V \phi(x, \delta, U)]\|_* \\ &\leq C_1 \delta, \end{aligned} \tag{9}$$

⁶Here, $\|\cdot\|$ is the implied norm: For a rank-3 tensor T , $\|T\| = \sup_{x, y, z \neq 0} \frac{|T(x, y, z)|}{\|x\| \|y\| \|z\|}$.

⁷When $\|\cdot\|$ is defined from an inner product, $\mathbb{E} \|X - E[X]\|^2 = \mathbb{E} [\|X\|^2] - \|\mathbb{E}[X]\|^2 \leq \mathbb{E} [\|X\|^2]$ also holds, shaving off a factor of four from the inequality below.

where $C_1 = \frac{L}{2} \mathbb{E} [\|V\|_* \|U\|^2]$. The claim regarding the variance of G follows in a similar manner as in Case 1, i.e., $f \in \mathcal{C}^3$.

Therefore, for f convex and L -smooth, γ defined by (3) is a $(C_1\delta, C_2/\delta^2)$ type-I oracle, where C_1 is given by (9) and C_2 as defined in Case 1. ■

Proof of Proposition 2

Before the proof, we introduce a fundamental theorem of vector calculus, which is commonly known as the Gauss-Ostrogradsky theorem or the divergence theorem. A special case of the theorem for real-valued functions in $\mathbb{R}^n$ can be stated as follows.

Lemma 1 *Suppose $W \subset \mathbb{R}^n$ is an open set with the boundary ∂W . At each point of ∂W there is a normal vector n_W such that n_W (i) has unit norm, (ii) is orthogonal to ∂W , (iii) points outward from W . Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a function of class C^1 defined at least on the closure of W , then we have*

$$\int_W \nabla f \, dW = \int_{\partial W} f n_W \, d\partial W. \quad \square$$

PROOF Given that $\mathbb{E} [\|V\|_*^2]$ and $\mathbb{E} [\xi^2]$ are bounded, the variance of G remains the same as stated in Proposition 1.

As to the bias, let $\tilde{f}$ be a smoothed version of f , i.e., $\forall x \in \mathcal{K}$,

$$\tilde{f}(x) = \mathbb{E} [f(x + \delta V)] = \int_{v \in W} f(x + \delta v) \frac{dw}{|W|},$$

where the expectation is w.r.t. V , which is a random variable uniformly chosen from W . The second equality interprets the expectation as integral. Now we want to prove that for any given $x \in \mathcal{K}$, G is an unbiased gradient estimate of $\tilde{f}$ at x . Since U is uniformly distributed over ∂W , the expectation of G can be written as

$$\mathbb{E} [G] = \frac{|\partial W|}{|W|} \int_{\partial W} \frac{1}{\delta} f(x + \delta U) n_W(U) \frac{dU}{|\partial W|} = \int_W \nabla f(x + \delta U) \frac{dU}{|W|},$$

where the second equality follows from Lemma 1, by replacing the gradient of $\hat{f}(u) = \frac{1}{\delta} f(x + \delta u)$ with $\nabla f(x + \delta u)$. Then, the order of the gradient and the integral can be exchanged, because $\int_W f(x + \delta U) \, dU$ exists. Consequently, we obtain $\mathbb{E} [G] = \nabla \tilde{f}(x)$.

Moreover, $\tilde{f}$ and f are actually close. In particular, for any $x \in \mathcal{K}$,

$$\tilde{f}(x) - f(x) = \int_W f(x + \delta w) - f(x) \frac{dw}{|W|}. \quad (10)$$

When f is L_0 -Lipschitz, $|f(x + \delta w) - f(x)| \leq L_0 \delta \|w\|$, which combined with (10) gives that γ is a type-II oracle with $c_1(\delta) = C_1 \delta$, where $C_1 = L_0 \sup_{w \in W} \|w\|$.

When f is convex and L -smooth, $0 \leq f(x + \delta w) - f(x) - \langle \nabla f(x), \delta w \rangle \leq \frac{L}{2} \delta^2 \|w\|^2$. Given that W is symmetric, $\int_W \langle \nabla f(x), \delta w \rangle \, dw = 0$. Hence, one can easily get that γ is a type-II oracle with $c_1(\delta) = C'_1 \delta^2$, where $C'_1 = \frac{L}{2|W|} \int_W \|w\|^2 \, dw$.

Finally, if f is L -smooth,

$$\|\nabla \tilde{f}(x) - \nabla f(x)\|_* \leq \int_W \|\nabla f(x + \delta w) - \nabla f(x)\|_* \frac{dw}{|W|} \leq L \delta^2 \int_W \|w\|^2 \frac{dw}{|W|} = 2C'_1 \delta^2$$

with the same value of C'_1 as before. So γ is also a type-I oracle with $c_1(\delta) = 2C'_1 \delta^2$. ■

Proof of Proposition 3

PROOF

Case 1 ($f \in \mathcal{C}^3$):

We use the proof technique of [33] (in particular, Lemma 1 there). We start by bounding the bias. Since by assumption $\mathbb{E}[\xi^+ - \xi^- | V] = 0$, we have

$$\mathbb{E} \left[V \left(\frac{\xi_n^+ - \xi_n^-}{2\delta} \right) \right] = 0,$$

implying that

$$\mathbb{E}[G] = \mathbb{E} \left[V \frac{f(X^+) - f(X^-)}{2\delta} \right].$$

By Taylor's theorem, using that $f \in \mathcal{C}^3$, we obtain, a.s.,

$$f(x \pm \delta U) = f(x) \pm \delta U^\top \nabla f(x) + \frac{\delta^2}{2} U^\top \nabla^2 f(x) U \pm \frac{\delta^3}{2} R^\pm(x, \delta, U)(U, U, U),$$

where, as in the proof of Proposition 1, $R^\pm(x, \delta, U)$ is defined as follows:

$$R^\pm(x, \delta, U) = \int_0^1 \nabla^3 f(x \pm s\delta U)(1-s)^2 ds. \quad (11)$$

Letting $B_3 = \sup_{x \in D} \|\nabla^3 f(x)\|$, we have $\|R^\pm(x, \delta, U)\| \leq B_3/3$ a.s. Now,

$$\begin{aligned} V \frac{f(X^+) - f(X^-)}{2\delta} &= V \frac{f(x + \delta U) - f(x - \delta U)}{2\delta} \\ &= V U^\top \nabla f(x) + \frac{\delta^2}{4} V (R^+(x, \delta, U) + R^-(x, \delta, U))(U \otimes U \otimes U). \end{aligned} \quad (12)$$

and therefore, by taking expectations of both sides, using $\mathbb{E}[V U^\top] = I$ and then $|R^\pm(x, \delta, U)(U \otimes U \otimes U)| \leq \|R^\pm(x, \delta, U)\| \|U\|^3$, we get that

$$\|\mathbb{E}[G] - \nabla f(x)\|_* \leq C_1 \delta^2,$$

where $C_1 = \frac{B_3 \mathbb{E}[\|V\|_* \|U\|^3]}{6}$.

Using arguments similar to that in the proof of Proposition 1, the variance of G is bounded as follows:

$$\begin{aligned} \mathbb{E} \|G - \mathbb{E}G\|_*^2 &\leq 4\mathbb{E} \|G\|_*^2 \\ &= 4\mathbb{E} \left(\|V\|_*^2 \left(\left(\frac{\xi^+ - \xi^-}{2\delta} \right)^2 + 2 \left(\frac{\xi^+ - \xi^-}{2\delta} \right) \left(\frac{f(X^+) - f(X^-)}{2\delta} \right) + \left(\frac{f(X^+) - f(X^-)}{2\delta} \right)^2 \right) \right) \\ &= 4\mathbb{E} \left(\|V\|_*^2 \left(\frac{\xi^+ - \xi^-}{2\delta} \right)^2 \right) + 4\mathbb{E} \left(\|V\|_*^2 \right) \left(\frac{f(X^+) - f(X^-)}{2\delta} \right)^2 \\ &\leq \frac{C_2}{\delta^2}, \end{aligned} \quad (13)$$

where $C_2 = 4\mathbb{E}[\|V\|_*^2] \left(\sigma_\xi^2 + \text{span}(f) \right)$ and $\text{span}(f) = \sup_{x \in \mathcal{D}} f(x) - \inf_{x \in \mathcal{D}} f(x)$. The equality in (13) follows from $\mathbb{E}[\xi^+ - \xi^- | U, V] = 0$.

Therefore, for $f \in \mathcal{C}^3$, γ defined by (4) is a $(C_1 \delta^2, C_2 / \delta^2)$ type-I oracle.

Case 2 (Controlled noise and F is convex and L_ψ -smooth):

The proof follows by parallel arguments to that used in the proof of Lemma 1 in [15] and we give it here for the sake of completeness.

For any convex function f with an L -Lipschitz gradient, for any $\delta > 0$ it holds that

$$\frac{\langle \nabla f(x), \delta u \rangle}{2\delta} \leq \frac{f(x + \delta u) - f(x)}{2\delta} \leq \frac{\langle \nabla f(x), \delta u \rangle + (L/2)\|\delta u\|^2}{2\delta}.$$

Using similar inequalities for $f(x - \delta u)$, we obtain

$$\langle \nabla f(x), u \rangle - \frac{L\delta\|u\|^2}{2} \leq \frac{f(x + \delta u) - f(x - \delta u)}{2\delta} \leq \langle \nabla f(x), u \rangle + \frac{L\delta\|u\|^2}{2}.$$

Letting $\phi(x, \delta, u) := \frac{1}{\delta} \left(\frac{f(x + \delta u) - f(x - \delta u)}{2\delta} - \langle \nabla f(x), u \rangle \right)$, we get

$$|\phi(x, \delta, u)| \leq \frac{L}{2}\|u\|^2.$$

Using $\mathbb{E}[VU^\top] = I$, we obtain

$$\begin{aligned} \mathbb{E} \left[V \left(\frac{f(x + \delta U) - f(x - \delta U)}{2\delta} \right) \right] &= \mathbb{E} [VU^\top \nabla f(x) + \delta \phi(x, \delta, U)V] \\ &= \nabla f(x) + \delta \hat{\phi}(x, \delta), \end{aligned}$$

where $\hat{\phi}(x, \delta)$ satisfies $\|\hat{\phi}(x, \delta)\|_* \leq \frac{L}{2}\mathbb{E}[\|V\|_*\|U\|^2]$.

Applying the above expression to $F(\cdot, \Psi)$ and recalling that $G = V \left(\frac{F(X^+, \psi) - F(X^-, \psi)}{2\delta} \right)$, we have, for P -almost every ψ ,

$$\mathbb{E}[G] = \nabla F(x, \psi) + \delta \hat{\phi}(x, \delta),$$

where, as before, $\hat{\phi}(x, \delta)$ satisfies $\|\hat{\phi}(x, \delta)\|_* \leq \frac{L_\psi}{2}\mathbb{E}[\|V\|_*\|U\|^2]$.

Using the fact that $E[\nabla F(x, \Psi)] = \nabla f(x)$, we obtain

$$\begin{aligned} \|\mathbb{E}[G] - \nabla f(x)\|_* &= \left\| \mathbb{E} \left[V \left(\frac{f(x + \delta U) - f(x - \delta U)}{2\delta} \right) - VU^\top \nabla f(x) \right] \right\|_* \leq \delta \|\mathbb{E}[V\phi(x, \delta, U)]\|_* \\ &\leq \frac{\delta \bar{L}_\Psi}{2} \mathbb{E}[\|V\|_*\|U\|^2], \end{aligned}$$

and the claim for the bias follows by setting $C_1 = \frac{\bar{L}_\Psi}{2}\mathbb{E}[\|V\|_*\|U\|^2]$.

We now bound $\mathbb{E}[\|G\|_*^2]$ as follows:

$$\begin{aligned} \mathbb{E}\|G\|^2 &= \mathbb{E}\|V(\delta \phi(x, \delta, U) + U^\top \nabla f(x))\|^2 \leq \mathbb{E} \left[\left(\|VU^\top \nabla f(x)\|_* + \frac{\delta L}{2}\|V\|_*\|U\|^2 \right)^2 \right] \\ &\leq 2\mathbb{E}[\|VU^\top \nabla f(x)\|_*^2] + \frac{\delta^2 \bar{L}_\Psi^2}{2} \mathbb{E}[\|V\|_*^2\|U\|^4], \end{aligned}$$

and the claim for the variance follows by setting $C_2 = 2B_1^2 + \frac{\bar{L}_\Psi^2}{2}\mathbb{E}[\|V\|_*^2\|U\|^4]$ with $B_1 = \sup_{x \in \mathcal{K}} \|\nabla f(x)\|_*$.

Therefore, for the case of controlled noise with a convex and L_ψ -smooth F , we have that γ defined by (4) is a $(C_1\delta, C_2)$ type-I oracle. $\blacksquare$

C Upper Bound Proof

In the proof, we use Lemma 2 to deal with noisy and biased gradient, which is essentially Theorem C.4 of [26], and also identical to Theorem 6.3 of [6], who cites Dekel et al. [12] as the source. The lemma and its proof can be found below.

PROOF (THEOREM 1) By the discussion in Section 2, we have $\Delta_{\mathcal{F},n}^*(c_1, c_2) \leq R_{\mathcal{F},n}^*(c_1, c_2)/n$ by Jensen's inequality, so it is enough to bound the regret for the online case. From the bias condition in Definition 2, we get

$$\frac{1}{n} \mathbb{E} \left[\sum_{t=1}^n f_t(X_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^n f_t(x) \right] \leq \frac{1}{n} \mathbb{E} \left[\sum_{t=1}^n \tilde{f}_t(X_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^n \tilde{f}_t(x) \right] + 2C_1\delta^p.$$

Given $\bar{G}_t = \mathbb{E}[G_t] = \nabla \tilde{f}(X_t)$, where $\tilde{f} \in \mathcal{F}_{L,0}$ is convex and smooth, the result immediately follows by applying Lemma 2 to $\tilde{f}$. Substituting $\beta_t = 0$, $\sigma_t^2 = C_2\delta^{-q}$, $a_t = at^r$ for some $0 < r < 1$, respectively, we obtain

$$\begin{aligned} & \frac{1}{n} \mathbb{E} \left[\sum_{t=1}^n f(X_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^n f(x) \right] \\ & \leq \frac{1}{n} \left(\mathbb{E} \left[f(X_1) - \inf_{x \in \mathcal{K}} f(x) \right] + \frac{DL}{\alpha} \right) + \frac{Da}{\alpha} n^{r-1} + \frac{C_2\delta^{-q}}{2a(1-r)} n^{-r} + \left(2 + \frac{2}{n}\right) C_1\delta^p. \end{aligned} \quad (14)$$

Choosing $r = \frac{p+q}{2p+q}$, and a, δ as stated in Theorem 1, the last 3 terms in (14) are optimized to

$$K_1 C_1^{q/(2p+q)} C_2^{p/(2p+q)} n^{-p/(2p+q)},$$

where $(K_1/3)^{2p+q} \leq 2^{2q-p} (2 + q/p)^p (D/\alpha)^p$.

When $\tilde{f}_t \in \mathcal{F}_{L,\mu,\mathcal{R}}$ is L -smooth and μ -strongly convex, similarly, we obtain

$$\begin{aligned} & \frac{1}{n} \mathbb{E} \left[\sum_{t=1}^n f(X_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^n f(x) \right] - \frac{1}{n} \mathbb{E} \left[f(X_1) - \inf_{x \in \mathcal{K}} f(x) \right] \\ & \leq \left(2 + \frac{2}{n}\right) C_1\delta^p + \frac{C_2\delta^{-q}}{\alpha\mu n} \sum_{t=1}^{n-1} \frac{1}{t - \frac{2L}{\alpha\mu}} \\ & \leq \left(2 + \frac{2}{n}\right) C_1\delta^p + \frac{C_2}{\alpha\mu} \delta^{-q} \frac{\log n + 1 + \alpha\mu/(\alpha\mu - 2L)}{n} \\ & \leq K_2 C_1^{\frac{q}{p+q}} C_2^{\frac{p}{p+q}} \left(\frac{\log n + 1 + \frac{\alpha\mu}{\alpha\mu - 2L}}{n} \right)^{\frac{p}{p+q}}. \end{aligned}$$

The last step optimizes the bound via letting $\delta^{p+q} = \frac{C_2 \left(\log n + 1 + \frac{\alpha\mu}{\alpha\mu - 2L} \right)}{2\alpha\mu C_1(n+1)}$, where $K_2^{p+q} = 2^q \alpha^{-p} \mu^{-p}$. ■

Lemma 2 Let $(\mathcal{F}_t)_t$ be a filtration such that X_t is $\mathcal{F}_t$ -measurable. Let $\bar{G}_t = \mathbb{E}[G_t|\mathcal{F}_t]$ and assume that the nonnegative real-valued deterministic sequence $(\beta_t)_{1 \leq t \leq n}$ is such that $\|\bar{G}_t - \nabla f(X_t)\|_* \leq \beta_t$ holds almost surely. Further, assume that $D = \sup_{x,y \in \mathcal{K}} D_{\mathcal{R}}(x,y)$ and let $\eta_t = \frac{\alpha}{a_t + L}$ for some increasing sequence $(a_t)_{t=1}^{n-1}$ of numbers. Then, if f is convex and L -smooth,

$$\mathbb{E} \left[\sum_{t=1}^n f(X_t) - f(x) \right] \leq \mathbb{E} [f(X_1) - f(x)] + \sqrt{\frac{2D}{\alpha}} \sum_{t=1}^{n-1} \beta_t + \frac{D(a_{n-1} + L)}{\alpha} + \sum_{t=1}^{n-1} \frac{\sigma_t^2}{2a_t},$$

where $\sigma_t^2 = \mathbb{E} [\|G_t - \bar{G}_t\|_*^2]$ is the “variance” of G_t .

If f is also μ -strongly convex w.r.t. $\mathcal{R}$, which is α -strongly convex, $\alpha > \frac{2L}{\mu}$. Let $\eta_t = \frac{2}{\mu t}$, $a_t = \frac{\alpha\mu}{2}t - L > 0$,

$$\mathbb{E} \left[\sum_{t=1}^n f(X_t) - f(x) \right] \leq \mathbb{E} [f(X_1) - f(x)] + \sqrt{\frac{2D}{\alpha}} \sum_{t=1}^{n-1} \beta_t + \sum_{t=1}^{n-1} \frac{\sigma_t^2}{2a_t}. \quad \square$$

Before the proof, we introduce a well-known bound on the instantaneous linearized regret of Mirror Descent.

Lemma 3 For any $x \in \mathcal{K}$ and any $t \geq 1$,

$$\langle G_t, X_{t+1} - x \rangle \leq \frac{1}{\eta_t} (D_{\mathcal{R}}(x, X_t) - D_{\mathcal{R}}(x, X_{t+1}) - D_{\mathcal{R}}(X_{t+1}, X_t)),$$

where X_{t+1} is selected as in Algorithm 1. $\square$

PROOF The point X_{t+1} is the minimizer of $\Psi_{t+1}(x) = \eta_t \langle G_t, x \rangle + D_{\mathcal{R}}(x, X_t)$ over $\mathcal{K}$. The gradient of $\Psi_{t+1}(x)$ is

$$\nabla \Psi_{t+1}(x) = \eta_t G_t + \nabla \mathcal{R}(x) - \nabla \mathcal{R}(X_t).$$

By the optimality condition, for any $x \in \mathcal{K}$,

$$\langle \eta_t G_t + \nabla \mathcal{R}(x) - \nabla \mathcal{R}(X_t), x - X_{t+1} \rangle \geq 0,$$

which is equivalent to the result by substituting the definition of Bregman divergence $D_{\mathcal{R}}$. $\blacksquare$

PROOF (LEMMA 2) Writing $\nabla f = (\nabla f - \bar{G}_t) + \xi_t + G_t$ where $\xi_t = \bar{G}_t - G_t$ is the “noise”, from the convexity of f we obtain

$$\begin{aligned} f(X_t) - f(x) &\leq \langle \nabla f(X_t), X_t - x \rangle \\ &= \langle (\nabla f - \bar{G}_t) + \xi_t + G_t, X_t - x \rangle \\ &\leq \beta_t \sqrt{\frac{2D}{\alpha}} + \langle \xi_t, X_t - x \rangle + \langle G_t, X_t - x \rangle, \end{aligned} \quad (15)$$

where in the last step we used the Cauchy-Schwarz inequality and the strong convexity of $\mathcal{R}$. After we plug this into (15), the plan is to take the conditional expectation of both sides w.r.t. $\mathcal{F}_t$. As X_t is $\mathcal{F}_t$ -measurable, and by the definition of ξ_t and $\bar{G}_t$, $\mathbb{E}[\xi_t | \mathcal{F}_t] = 0$, we have

$$\mathbb{E}[\langle \xi_t, X_{t+1} - x \rangle | \mathcal{F}_t] = \underbrace{\mathbb{E}[\langle \xi_t, X_t - x \rangle | \mathcal{F}_t]}_{=0} + \mathbb{E}[\langle \xi_t, X_{t+1} - X_t \rangle | \mathcal{F}_t].$$

We bound the second term inside the expectation from the Fenchel-Young inequality:

$$\langle \xi_t, X_{t+1} - X_t \rangle \leq \frac{1}{2} \left(\frac{\|\xi_t\|_*^2}{a_t} + a_t \|X_{t+1} - X_t\|^2 \right).$$

Using again that $\mathcal{R}$ is α -strongly convex, we get

$$\langle \xi_t, X_{t+1} - X_t \rangle \leq \frac{1}{2} \left(\frac{\|\xi_t\|_*^2}{a_t} + \frac{2a_t}{\alpha} D_{\mathcal{R}}(X_{t+1}, X_t) \right).$$

Applying Lemma 3 to bound $\langle G_t, X_{t+1} - x \rangle$, and putting everything together gives

$$\begin{aligned} \mathbb{E}[f(X_{t+1}) - f(x) | \mathcal{F}_t] &\leq \beta_t \sqrt{\frac{2D}{\alpha}} + \frac{1}{2a_t} \mathbb{E}[\|\xi_t\|_*^2 | \mathcal{F}_t] + \frac{1}{\eta_t} (D_{\mathcal{R}}(x, X_t) - D_{\mathcal{R}}(x, X_{t+1})) \\ &\quad + \underbrace{\left(\frac{a_t + L}{\alpha} - \frac{1}{\eta_t} \right) D_{\mathcal{R}}(X_{t+1}, X_t)}_{=0}. \end{aligned} \quad (16)$$

Given $\sigma_t^2 = \mathbb{E} [\|\xi_t\|_*^2]$, we will sum up these inequalities for $t = 1, \dots, n-1$. To bound the right-hand side of the result, we use

$$\begin{aligned}
 & \sum_{t=1}^{n-1} \frac{1}{\eta_t} (D_{\mathcal{R}}(x, X_t) - D_{\mathcal{R}}(x, X_{t+1})) \\
 &= D_{\mathcal{R}}(x, X_1) \frac{1}{\eta_1} + D_{\mathcal{R}}(x, X_2) \left(\frac{1}{\eta_2} - \frac{1}{\eta_1} \right) + \dots + D_{\mathcal{R}}(x, X_{n-1}) \left(\frac{1}{\eta_{n-1}} - \frac{1}{\eta_{n-2}} \right) - \frac{1}{\eta_{n-1}} D_{\mathcal{R}}(x, X_n) \\
 &\leq D \frac{1}{\eta_1} + D \sum_{t=2}^{n-1} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \\
 &= \frac{D}{\eta_{n-1}}.
 \end{aligned} \tag{17}$$

The inequality results from the fact that $\{\eta_t\}$ is non-increasing. Hence, by the tower rule,

$$\begin{aligned}
 \mathbb{E} \left[\sum_{t=1}^n f(X_t) - f(x) \right] &\leq \mathbb{E} [f(X_1) - f(x)] + \sqrt{\frac{2D}{\alpha}} \sum_{t=1}^{n-1} \beta_t + \frac{D}{\eta_{n-1}} + \sum_{t=1}^{n-1} \frac{\sigma_t^2}{2a_t} \\
 &= \mathbb{E} [f(X_1) - f(x)] + \sqrt{\frac{2D}{\alpha}} \sum_{t=1}^{n-1} \beta_t + \frac{D(a_{n-1} + L)}{\alpha} + \sum_{t=1}^{n-1} \frac{\sigma_t^2}{2a_t}.
 \end{aligned}$$

When f is L -smooth and μ -strongly convex, we can rewrite (15) as

$$\begin{aligned}
 & f(X_{t+1}) - f(x) \\
 &\leq f(X_t) + \langle \hat{G}_t, X_{t+1} - X_t \rangle + \frac{L}{2} \|X_{t+1} - X_t\|^2 - \left\{ f(X_t) + \langle \hat{G}_t, x - X_t \rangle + \frac{\mu}{2} D_{\mathcal{R}}(x, X_t) \right\} \\
 &= \langle \hat{G}_t, X_{t+1} - x \rangle + \frac{L}{2} \|X_{t+1} - X_t\|^2 - \frac{\mu}{2} D_{\mathcal{R}}(x, X_t) \\
 &\leq \langle \hat{G}_t, X_{t+1} - x \rangle + \frac{L}{\alpha} D_{\mathcal{R}}(X_{t+1}, X_t) - \frac{\mu}{2} D_{\mathcal{R}}(x, X_t).
 \end{aligned}$$

Again, (16) can be written as

$$\begin{aligned}
 \mathbb{E} [f(X_{t+1}) - f(x) | \mathcal{F}_t] &\leq \beta_t \sqrt{\frac{2D}{\alpha}} + \frac{1}{2a_t} \mathbb{E} [\|\xi_t\|_*^2 | \mathcal{F}_t] \\
 &\quad + \left(\frac{1}{\eta_t} - \frac{\mu}{2} \right) D_{\mathcal{R}}(x, X_t) - \frac{1}{\eta_t} D_{\mathcal{R}}(x, X_{t+1}) + \left(\frac{L + a_t}{\alpha} - \frac{1}{\eta_t} \right) D_{\mathcal{R}}(X_{t+1}, X_t).
 \end{aligned}$$

Given $\frac{1}{\eta_t} = \frac{\mu t}{2} = \frac{L + a_t}{\alpha}$, summing up these inequalities for $t = 1, 2, \dots, n-1$, we get

$$\begin{aligned}
 \mathbb{E} \left[\sum_{t=1}^n f(X_t) - f(x) \right] &\leq \mathbb{E} [f(X_1) - f(x)] + \sqrt{\frac{2D}{\alpha}} \sum_{t=1}^{n-1} \beta_t + \sum_{t=1}^{n-1} \frac{\sigma_t^2}{2a_t} - \frac{1}{\eta_{n-1}} D_{\mathcal{R}}(x, X_n) \\
 &\leq \mathbb{E} [f(X_1) - f(x)] + \sqrt{\frac{2D}{\alpha}} \sum_{t=1}^{n-1} \beta_t + \sum_{t=1}^{n-1} \frac{\sigma_t^2}{2a_t},
 \end{aligned}$$

which finishes the proof. ■

C.1 Online optimization based on Saha and Tewari [30]

Assume that $\tilde{f}$ in the (c_1, c_2) type-II oracle is convex with gradients bounded by M . We also need that $\mathbb{E} [\tilde{f}_t(Y_t) - \tilde{f}_t(X_t)] \leq c_3(\delta) = N\delta^2$. This holds for the smooth case when $\mathbb{E} [Y_t | \mathcal{F}_t] = X_t$, since

then $\mathbb{E} [\tilde{f}_t(Y_t) - \tilde{f}_t(X_t) | \mathcal{F}_t] \leq \mathbb{E} [\langle \nabla \tilde{f}_t(X_t), Y_t - X_t \rangle + \frac{L}{2} \|X_t - Y_t\|^2 | \mathcal{F}_t] \leq L\delta^2/2$, where we used that $\mathbb{E} [Y_t - X_t | \mathcal{F}_t] = 0$. Instead of Lemma 3, the prox-lemma states

$$\langle G_t, X_t - x \rangle \leq \frac{1}{\eta_t} (D_{\mathcal{R}}(x, X_t) - D_{\mathcal{R}}(x, X_{t+1})) + \eta_t \frac{\|G_t\|_*^2}{2\alpha}.$$

Using the linearization $\tilde{f}(X_t) - \tilde{f}(x) \leq \langle \nabla \tilde{f}_t, X_t - x \rangle = \mathbb{E} [\langle G_t, X_t - x \rangle | \mathcal{F}_t]$, summing up the bounds for all t , and applying (17), we get the standard mirror descent bound:

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^n \tilde{f}_t(X_t) - \inf_{x \in \mathcal{K}} \sum_{t=1}^n \tilde{f}_t(x) \right] &\leq \mathbb{E} \left[\sum_{t=1}^n \langle G_t, X_t - x \rangle \right] \\ &\leq \frac{D}{\eta_{n-1}} + \sum_{t=1}^n \eta_t \frac{\mathbb{E} [\|G_t\|_*^2]}{2\alpha} \\ &\leq \frac{D}{\eta_{n-1}} + \sum_{t=1}^n \eta_t \frac{2M^2 + 2c_2(\delta)^2}{2\alpha} \end{aligned} \quad (18)$$

Moreover,

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^n f_t(Y_t) \right] - \inf_{x \in \mathcal{K}} \sum_{t=1}^n f_t(x) &\leq \mathbb{E} \left[\sum_{t=1}^n \tilde{f}_t(Y_t) \right] - \inf_{x \in \mathcal{K}} \sum_{t=1}^n \tilde{f}_t(x) + 2nc_1(\delta) \\ &\leq \mathbb{E} \left[\sum_{t=1}^n \tilde{f}_t(X_t) \right] - \inf_{x \in \mathcal{K}} \sum_{t=1}^n \tilde{f}_t(x) + 2nc_1(\delta) + nc_3(\delta) \\ &\leq \frac{D}{\eta_{n-1}} + \sum_{t=1}^n \eta_t \frac{2M^2 + 2c_2(\delta)^2}{2\alpha} + 2nc_1(\delta) + nc_3(\delta). \end{aligned}$$

This should recover the bound of Saha and Tewari [30].

D Lower Bound Proof

D.1 Proof of Theorem 2 for convex + smooth function class $\mathcal{F}_{L,0}(\mathcal{K})$

We will use a novel technique that will allow us to reduce the d -dimensional case to the 1-dimensional case (see later). Thus, we start with the one-dimensional case.

Proof in one dimension: As the section title suggests, we here prove the theorem for $\mathcal{F} = \mathcal{F}_{L,0}(\mathcal{K})$, where by the assumptions of the theorem, $L \geq 1/2$, $\mathcal{K}$ is convex and $[-1, 1] \subset \mathcal{K}$. For brevity, let Δ_n^* denote the minimax error $\Delta_n^*(\mathcal{F}, c_1, c_2)$. Throughout the proof, a d -dimensional normal distribution with mean μ and covariance matrix Σ is denoted by $N(\mu, \Sigma)$.

We follow the standard proof technique of lower bounds: We define two functions $f_+, f_- \in \mathcal{F}$ with associated type-I gradient oracles γ_+, γ_- such that the expected error of any deterministic algorithm can be bounded from below for the case when the environment is chosen uniformly at random from $\{(f_+, \gamma_+), (f_-, \gamma_-)\}$. By Yao's principle [35], the same lower bound applies to the minimax error Δ_n^* even when randomized algorithms are also allowed.

The proof uses (c_1, c_2) type-I oracles which have no memory. In particular, we restrict the class of oracles to those that on input (x, δ) return a random gradient estimate

$$G(x, \delta) = \bar{\gamma}(x, \delta) + \xi \quad (19)$$

with some map $\bar{\gamma} : \mathcal{K} \times [0, 1) \rightarrow \mathbb{R}$, where ξ is a zero-mean normal random variable with variance $c_2(\delta) := C_2\delta^{-q}$, satisfying the variance requirement, and drawn independently every time the oracle is queried.⁸ The map $\bar{\gamma}$, which

⁸The argument presented below is not hard to extend to the case when all observations are from a bounded set, but this extension is left to the reader.

will be chosen based on f to satisfy the requirement on the bias. The Y value returned by the oracles is made equal to x .

Next we define the two target functions and their associated oracles. With a slight abuse of notation, we will use interchangeably the subscripts $+$ ($-$) and $+1$ (-1) for any quantities corresponding to these two environments, e.g., f_+ and f_{+1} (respectively, f_- and f_{-1}). For $v \in \{\pm 1\}$, let

$$f_v(x) := \epsilon(x - v) + 2\epsilon^2 \ln \left(1 + e^{-\frac{x-v}{\epsilon}} \right), \quad x \in \mathcal{K}. \quad (20)$$

These functions, with the choice $\epsilon = 0.1$, are shown in Fig. 2a. The idea underlying these functions is that they approximate $\epsilon|x - v|$, but with a prescribed smoothness. The first and second derivatives of f_v are

$$f'_v(x) = \epsilon \frac{1 - e^{-\frac{x-v}{\epsilon}}}{1 + e^{-\frac{x-v}{\epsilon}}}, \quad \text{and} \quad f''_v(x) = \frac{2e^{-\frac{x-v}{\epsilon}}}{\left(1 + e^{-\frac{x-v}{\epsilon}}\right)^2}$$

(the functions were designed by choosing f'_v). From the above calculation, it is easy to see that $0 \leq f''(x) \leq 1/2$; thus f_v is $\frac{1}{2}$ -smooth, and so $f_v \in \mathcal{F}$.

For $f_v, v \in \{-1, +1\}$, the gradient oracle we consider is defined as $\gamma_v(x, \delta) = \bar{\gamma}_v(x, \delta) + \xi_\delta$ with $\xi_\delta \sim \mathcal{N}(0, \frac{C_2}{\delta^q})$ selected independently for every query, where $\bar{\gamma}_v$ is a biased estimate of the gradient f'_v . The derivatives of f_+ and f_- are shown in Fig. 2a; we define the "bias" in $\bar{\gamma}_v$ to move the gradients closer to each other: The idea is to shift f'_+ and f'_- towards each other, with the shift depending on the allowed bias $c_1(\delta) = C_1\delta^p$. In particular, since $f'_+ \leq f'_-$, f'_+ is shifted up, while f'_- is shifted down. However, the shifted up version of f'_+ is clipped for positive x so that it never goes above the shifted down version of f'_- , cf. Fig. 2b. By moving the curves towards each other, algorithms which rely on the obtained oracles will have an increasingly harder time (depending on the size of the shift) to distinguish whether the function optimized is f_+ or f_- . Since

$$0 \leq f'_-(x) - f'_+(x) \leq \sup_x f'_-(x) - \inf_x f'_+(x) = 2\epsilon,$$

we don't allow shifts larger than ϵ (so no crossing over happens), leading to the following formal definitions:

$$\bar{\gamma}_+(x, \delta) = \begin{cases} f'_+(x) + \min(\epsilon, C_1\delta^p), & \text{if } x < 0; \\ \min\{f'_+(x) + \min(\epsilon, C_1\delta^p), f'_-(x) - \min(\epsilon, C_1\delta^p)\}, & \text{otherwise,} \end{cases} \quad (21)$$

and

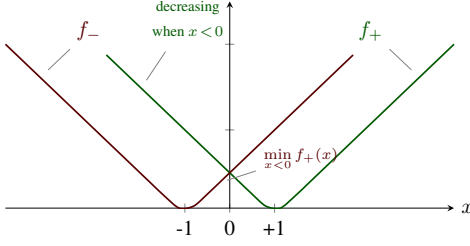
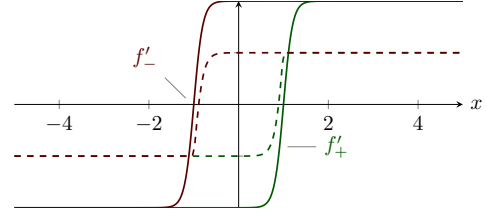
$$\bar{\gamma}_-(x, \delta) = \begin{cases} f'_-(x) - \min(\epsilon, C_1\delta^p), & \text{if } x > 0; \\ \max\{f'_-(x) - \min(\epsilon, C_1\delta^p), f'_+(x) + \min(\epsilon, C_1\delta^p)\}, & \text{otherwise.} \end{cases} \quad (22)$$

We claim that the oracle γ_v based on these functions is indeed a (c_1, c_2) type-I oracle, with $c_1(\delta) = C_1\delta^p$ and $c_2(\delta) = \frac{C_2}{\delta^q}$. The variance condition is trivial. To see that $c_1(\delta) = C_1\delta^p$ works, notice that $\gamma_v(x, \delta) = -\gamma_{-v}(-x, \delta)$ and $f'_v(x) = -f'_{-v}(x)$. Thus, $|\bar{\gamma}_+(x, \delta) - f'_+(x)| = |\bar{\gamma}_-(-x, \delta) - f'_-(-x)|$, hence it suffices to consider $v = +1$. The bias condition trivially holds for $x < 0$. For $x \geq 0$, using that $f'_+(x) \leq f'_-(x)$, we get $f'_+(x) - \min(\epsilon, C_1\delta^p) \leq \bar{\gamma}_+(x, \delta) \leq f'_+(x) + \min(\epsilon, C_1\delta^p)$, showing $|\bar{\gamma}_+(x, \delta) - f'_+(x)| \leq C_1\delta^p$. Thus, γ_v is indeed an oracle with the required properties.

To bound the performance of any algorithm in minimizing $f_v, v \in \{\pm 1\}$, notice that f_v is minimized at $x_v^* = v$, with $f_v(v) = 2\epsilon^2 \ln 2$. Next we show that if x has the opposite sign of v , the difference $f_v(x) - f_v(x_v^*)$ is "large". This will mean that if the algorithm cannot distinguish between $v = +1$ and $v = -1$, it necessarily chooses a highly suboptimal point for either of these cases.

Since vf_v is decreasing on $\{x : xv \leq 0\}$, we have

$$M_v := \min_{x: xv \leq 0} f_v(x) - f_v(v) = f_v(0) - f_v(v) = \epsilon \left(-v + 2\epsilon \ln \frac{1 + e^{\frac{v}{\epsilon}}}{2} \right).$$


 (a) Plot of f_+ and f_- with $\epsilon = 0.1$

 (b) Plot of f'_+ and f'_- with $\epsilon = 0.1$. The dashed lines show $\bar{\gamma}_v(\cdot, \delta)$ for $C_1\delta^p = \epsilon$, $v \in \{\pm 1\}$.

Let $h(v) = -v + 2\epsilon \ln \frac{1 + e^{\frac{v}{\epsilon}}}{2}$. Simple algebra shows that h is an even function, that is, $h(v) = h(-v)$. Indeed,

$$h(v) = -v + 2\epsilon \ln \left(e^{\frac{v}{\epsilon}} \frac{1 + e^{-\frac{v}{\epsilon}}}{2} \right) = -v + 2\epsilon \frac{v}{\epsilon} + 2\epsilon \ln \frac{1 + e^{-\frac{v}{\epsilon}}}{2} = h(-v).$$

Specifically, $h(1) = h(-1)$ and thus

$$M_+ = M_- = \epsilon \left(-1 + 2\epsilon \ln \frac{1 + e^{\frac{1}{\epsilon}}}{2} \right).$$

From the foregoing, when $xv \leq 0$ and $\epsilon < \frac{1}{4 \ln 2}$, we have

$$f_v(x) - f_v(x_v^*) \geq \epsilon \left(-1 + 2\epsilon \ln \frac{1 + e^{\frac{1}{\epsilon}}}{2} \right) > \frac{\epsilon}{2}.$$

Hence,

$$f_v(x) - f_v(x_v^*) \geq \frac{\epsilon}{2} \mathbf{1}\{xv < 0\}. \quad (23)$$

Given the above definitions and (23), by Yao's principle, the minimax error (1) is lower bounded by

$$\Delta_n^* \geq \inf_{\mathcal{A}} \mathbb{E}[f_V(\hat{X}_n) - \inf_{x \in X} f_V(x)] \geq \inf_{\mathcal{A}} \frac{\epsilon}{2} \mathbb{P}(\hat{X}_n V < 0), \quad (24)$$

where $V \in \{\pm 1\}$ is a random variable, $\hat{X}_n$ is the estimate of the algorithm after n queries to the oracle γ_V for f_V , the infimum is taken over all deterministic algorithms, and the expectation is taken with respect to the randomness in V and the oracle. More precisely, the distribution above is defined as follows:

Consider a fixed (c_1, c_2) type-I oracle γ satisfying (19) and a deterministic algorithm $\mathcal{A}$. Let $x_t^{\mathcal{A}}$ (respectively, $\delta_t^{\mathcal{A}}$) denote the map from the algorithm's past observations that picks the point (respectively, accuracy parameter δ), which are sent to the oracle in round t . Define the probability space $(\Omega, \mathcal{B}, P_{\mathcal{A}, \gamma})$ with $\Omega = \mathbb{R}^n \times \{-1, 1\}$, its associated Borel sigma algebra $\mathcal{B}$, where the probability measure $P_{\mathcal{A}, \gamma}$ takes the form $P_{\mathcal{A}, \gamma} := p_{\mathcal{A}, \gamma} d(\lambda \times m)$, where λ is the Lebesgue measure on $\mathbb{R}^n$, m is the counting measure on $\{\pm 1\}$ and $p_{\mathcal{A}, \gamma}$ is the density function defined by

$$\begin{aligned} p_{\mathcal{A}, \gamma}(g_{1:n}, v) &= \frac{1}{2} \left(p_{\mathcal{A}, \gamma}(g_n | g_{1:n-1}) \cdot \dots \cdot p_{\mathcal{A}, \gamma}(g_{n-1} | g_{1:n-2}) \cdot \dots \cdot p_{\mathcal{A}, \gamma}(g_1) \right) \\ &= \frac{1}{2} \left(p_{\mathcal{N}}(g_n - \bar{\gamma}(x_n^{\mathcal{A}}(g_{1:n-1}), \delta_n^{\mathcal{A}}(g_{1:n-1})), c_2(\delta_n^{\mathcal{A}}(g_{1:n-1}))) \cdot \dots \cdot p_{\mathcal{N}}(g_1 - \bar{\gamma}(x_1^{\mathcal{A}}, \delta_1^{\mathcal{A}}), c_2(\delta_1^{\mathcal{A}})) \right), \end{aligned}$$

where $v \in \{-1, 1\}$ and $p_{\mathcal{N}}(\cdot, \sigma^2)$ is the density function of a $\mathcal{N}(0, \sigma^2)$ random variable. Then the expectation in (24) is defined w.r.t. the distribution $\mathbb{P} := \frac{1}{2} (P_{\mathcal{A}, \gamma_+} \mathbf{1}\{v = +1\} + P_{\mathcal{A}, \gamma_-} \mathbf{1}\{v = -1\})$ and $V : \Omega \rightarrow \{\pm 1\}$ is defined by

$V(g_{1:n}, v) = v$.⁹ Define $\mathbb{P}_+(\cdot) := \mathbb{P}(\cdot \mid V = 1)$, $\mathbb{P}_-(\cdot) := \mathbb{P}(\cdot \mid V = -1)$. From (24), we obtain

$$\Delta_n^* \geq \inf_{\mathcal{A}} \frac{\epsilon}{2} \left(\mathbb{P}_+(\hat{X}_n < 0) + \mathbb{P}_-(\hat{X}_n > 0) \right), \quad (25)$$

$$\geq \inf_{\mathcal{A}} \frac{\epsilon}{2} (1 - \|\mathbb{P}_+ - \mathbb{P}_-\|_{\text{TV}}), \quad (26)$$

$$\geq \inf_{\mathcal{A}} \frac{\epsilon}{2} \left(1 - \left(\frac{1}{2} D_{\text{kl}}(P_+ \| P_-) \right)^{\frac{1}{2}} \right), \quad (27)$$

where (25) uses the definitions of $\mathbb{P}_+$ and $\mathbb{P}_-$, $\|\cdot\|_{\text{TV}}$ denotes the total variation distance, (26) follows from its definition, while (27) follows from Pinsker's inequality. It remains to upper bound $D_{\text{kl}}(P_+ \| P_-)$.

Define G_t to be the t th observation of $\mathcal{A}$. Thus, $G_t : \Omega \rightarrow \mathbb{R}$, with $G_t(g_{1:n}, v) = g_t$. Let $P_+^t(g_1, \dots, g_t)$ denote the joint distribution of $G_1, \dots, G_t$ conditioned on $V = +1$. Let $P_+^t(\cdot \mid g_1, \dots, g_{t-1})$ denote the distribution of G_t conditional on $V = +1$ and $G_1 = g_1, \dots, G_{t-1} = g_{t-1}$. Define $P_-^t(\cdot \mid g_1, \dots, g_{t-1})$ in a similar fashion. Then, by the chain rule for KL-divergences, we have

$$D_{\text{kl}}(P_+ \| P_-) = \sum_{t=1}^n \int_{\mathbb{R}^{t-1}} D_{\text{kl}}(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})) dP_+^t(g_{1:t-1}). \quad (28)$$

By the oracle's definition on $V = +1$ we have $G_t \sim \mathcal{N}(\bar{\gamma}_+(x_t^A(G_{1:t-1}), \delta_t^A(G_{1:t-1})), c_2(\delta_t^A(G_{1:t-1})))$, i.e., $P_+^t(\cdot \mid g_{1:t-1})$ is the normal distribution with mean $\bar{\gamma}_+(x_t^A(G_{1:t-1}), \delta_t^A(G_{1:t-1}))$ and variance $c_2(\delta_t^A(G_{1:t-1}))$. Using the shorthands $x_t^A := x_t^A(g_{1:t-1})$, $\delta_t^A := \delta_t^A(g_{1:t-1})$, we have

$$D_{\text{kl}}(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})) = \frac{(\bar{\gamma}_+(x_t^A, \delta_t^A) - \bar{\gamma}_-(x_t^A, \delta_t^A))^2}{2c_2(\delta_t^A)},$$

as the KL-divergence between normal distributions $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ is equal to $\frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

It remains to upper bound the numerator. For $(x, \delta) \in \mathbb{R} \times (0, 1]$, first note that $\gamma_+(x, \delta) \leq \gamma_-(x, \delta)$. Hence,

$$\begin{aligned} |\gamma_+(x, \delta) - \gamma_-(x, \delta)| &= \gamma_-(x, \delta) - \gamma_+(x, \delta) \\ &< \sup_x \gamma_-(x, \delta) - \inf_x \gamma_+(x, \delta) \\ &= \lim_{x \rightarrow \infty} \gamma_-(x, \delta) - \lim_{x \rightarrow -\infty} \gamma_+(x, \delta) \\ &= \epsilon - \epsilon \wedge C_1 \delta^p - (-\epsilon + \epsilon \wedge C_1 \delta^p) \\ &= 2\epsilon - 2\epsilon \wedge C_1 \delta^p \\ &\leq 2(\epsilon - C_1 \delta^p)^+, \end{aligned} \quad (29)$$

where $(u)^+ = \max(u, 0)$ is the positive part of u .

From the above, using the abbreviations $x_t^A = x_t^A(g_{1:t-1})$ and $\delta_t^A = \delta_t^A(g_{1:t-1})$ (effectively fixing $g_{1:t-1}$ for this step),

$$D_{\text{kl}}(P_+^t(\cdot \mid g_{1:t-1}) \| P_-^t(\cdot \mid g_{1:t-1})) < \frac{2\{(\epsilon - C_1(\delta_t^A)^p)^+\}^2 (\delta_t^A)^q}{C_2} \quad (30)$$

$$\leq \sup_{\delta > 0} \frac{2\{(\epsilon - C_1 \delta^p)^+\}^2 \delta^q}{C_2}, \quad (31)$$

where inequality (30) follows from (29). Notice that the right-hand side of the above inequality does not depend on the algorithm anymore.

⁹Here, we are slightly abusing the notation as $\mathbb{P}$ depends on $\mathcal{A}$, but the dependence is suppressed. In what follows, we will define several other distributions derived from $\mathbb{P}$, which will all depend on $\mathcal{A}$, but for brevity this dependence will also be suppressed. The point where the dependence on $\mathcal{A}$ is eliminated will be called to the reader's attention.

Now, observe that $\sup_{\delta>0} \{(\epsilon - C_1 \delta^p)^+\}^2 \delta^q = \sup_{(\epsilon/C_1)^{1/p} \geq \delta > 0} (\epsilon - C_1 \delta^p)^2 \delta^q$. From this we obtain

$$\delta_* = \left(\frac{\epsilon q}{C_1(2p+q)} \right)^{1/p}. \quad (32)$$

Note that $C_1 \delta_*^p \leq \epsilon$, hence $\max_{\delta>0} \{(\epsilon - C_1 \delta^p)^+\}^2 \delta^q = (\epsilon - C_1 \delta_*^p)^2 \delta_*^q$. Plugging (31) into (28) and using this last observation we obtain

$$D_{\text{kl}}(P_+ \| P_-) \leq \frac{2n}{C_2} (\epsilon - C_1 \delta_*^p)^2 \delta_*^q. \quad (33)$$

Note that the above bound holds uniformly over all algorithms $\mathcal{A}$. Substituting the above bound into (27), we obtain

$$\Delta_n^* \geq \frac{\epsilon}{2} \left(1 - \sqrt{n} \frac{(\epsilon - C_1 \delta_*^p) \delta_*^{q/2}}{\sqrt{C_2}} \right) = \frac{\epsilon}{2} \left(1 - \sqrt{n} K_1 \epsilon^{\frac{2p+q}{2p}} \right), \quad (34)$$

where $K_1 = \frac{2p}{\sqrt{C_2}(2p+q)} \left(\frac{q}{C_1(2p+q)} \right)^{\frac{q}{2p}}$.

By choosing $\epsilon = \left(\frac{2p}{\sqrt{n} K_1 (4p+q)} \right)^{\frac{2p}{2p+q}}$, we see that

$$\Delta_n^* \geq \frac{2p+q}{2(4p+q)} \left(\frac{2p}{\sqrt{n} K_1 (4p+q)} \right)^{\frac{2p}{2p+q}} = \frac{(2p+q)^2}{2q^{\frac{q}{2p+q}} (4p+q)^{\frac{4p+q}{2p+q}}} C_1^{\frac{q}{2p+q}} C_2^{\frac{p}{2p+q}} n^{-\frac{p}{2p+q}}. \quad (35)$$

Now, when $q = 2$ and $p = 1$, the lower bound in (35) simplifies to

$$\Delta_n^* \geq \frac{2}{3\sqrt{3}} C_1^{1/2} C_2^{1/4} n^{-1/4}.$$

On the other hand, for $p = q = 2$, we obtain

$$\Delta_n^* \geq \frac{9}{10} \left(\frac{1}{25} \right)^{1/3} C_1^{1/3} C_2^{1/3} n^{-1/3}.$$

Generalization to d dimensions: To prove the d -dimensional result, we introduce a new device which allows us to relate the minimax error of the d -dimensional problem to that of the 1-dimensional problem. The main idea is to use separable d -dimensional functions and oracles and show that if there exists an algorithm with a small loss for a rich set of separable functions and oracles, then there exists good one dimensional algorithms for the one-dimensional components of the functions and oracles.

This device works as follows: For $1 \leq i \leq d$, let $\mathcal{K}_i \subset \mathbb{R}$ be nonempty sets, and for each $v \in V := \{\pm 1\}^d$, let $f_v^{(i)} : \mathcal{K}_i \rightarrow \mathbb{R}$ (later we will use $V_i = \{\pm 1\}$ and $f_v^{(i)}$ will be the functions constructed for the 1-dimensional case with $\{-1, +1\} \subset \mathcal{K}_i \subset \mathbb{R}$). Without the loss of generality, we assume that $\inf_{x \in \mathcal{K}_i} f_v^{(i)}(x) = 0$, so that the optimization error of the algorithm producing $\hat{X}_n$ as the output is $f_v^{(i)}(\hat{X}_n)$. Further, let $\gamma_v^{(i)}$ be a gradient oracle that the algorithms can use when the function to be optimized is $f_v^{(i)}$.

Let $\mathcal{K} = \times_{i=1}^d \mathcal{K}_i$ and for $v = (v_1, \dots, v_d) \in V := \times_{i=1}^d V_i$ (with a slight abuse of notation), let $f_v : \mathcal{K} \rightarrow \mathbb{R}$ be defined by

$$f_v(x) = \sum_{i=1}^d f_{v_i}^{(i)}(x_i), \quad x \in \mathcal{K}. \quad (36)$$

We also define a γ_v d -dimensional oracle as follows: The oracle is obtained from “composing” the d 1-dimensional oracles, $(\gamma_{v_i}^{(i)})_i$. In particular, the i -th component of the response of γ_v given the history of queries $(x_t, \delta_t, \dots, x_1, \delta_1) \in$

$(\mathcal{K} \times [0, 1))^t$ is defined as the response of $\gamma_{v_i}^{(i)}$ given the history of queries $(x_{t,i}, \delta_t, \dots, x_{1,i}, \delta_1) \in (\mathcal{K}_i \times [0, 1))^t$. This definition is so far unclear about the randomization of the oracles. In fact, it turns out that the one-dimensional oracles can even use the same randomization (i.e., their output can depend on the same single uniformly distributed random variable U), but they could also use separate randomization: Our argument will not depend on this. If $\Gamma^{(i)}(f_{v_i}^{(i)}, c_1, c_2)$ denotes a non-empty set of (c_1, c_2) type-I oracles for objective function $f_{v_i}^{(i)} : \mathcal{K}_i \rightarrow \mathbb{R}$, then we shall denote by $\Gamma_{\text{sep}}(f_v, c_1, c_2)$ the set of oracles for function f_v as defined in the previous paragraph. We also define $\mathcal{F}_{\text{sep}} = \{f : f(x) = \sum_{i=1}^d f_{v_i}^{(i)}(x_i), x \in \mathcal{K}, v_i \in V_i\}$. Note that when $\|\cdot\| = \|\cdot\|_2$ is used in the definition of type-I oracles then $\Gamma_{\text{sep}}(f_v, c_1/d^{1/2}, c_2/d) \subset \Gamma(f_v, c_1, c_2)$.

Let an algorithm $\mathcal{A}$ interact with an oracle γ . We will denote the distribution of the output $\hat{X}_n$ of $\mathcal{A}$ at the end of n rounds by $F_{\mathcal{A}, \gamma}$ (we fix n , hence the dependence of F on n is omitted). Thus, the expected optimization error of $\mathcal{A}$ on a function f with zero optimal value is

$$L^{\mathcal{A}}(f, \gamma) = \int f(x) F_{\mathcal{A}, \gamma}(dx).$$

Note that this definition applies both in the one and the d -dimensional cases. For $v \in V = \times_i V_i$, we introduce the abbreviation

$$L^{\mathcal{A}}(v) = L^{\mathcal{A}}(f_v, \gamma_v).$$

We also define

$$\tilde{L}_i^{\mathcal{A}}(v) = \int f_{v_i}^{(i)}(x_i) F_{\mathcal{A}, \gamma_v}(dx)$$

so that

$$L^{\mathcal{A}}(v) = \sum_{i=1}^d \tilde{L}_i^{\mathcal{A}}(v).$$

Also, for $v \in V_i$ and a one-dimensional algorithm $\mathcal{A}$, we let

$$L_i^{\mathcal{A}}(v) = L^{\mathcal{A}}(f_v^{(i)}, \gamma_v^{(i)}).$$

Note that while the domain of $\tilde{L}_i^{\mathcal{A}}$ is $V = \times_i V_i$, the domain of $L_i^{\mathcal{A}}$ is V_i , while both express an expected error measured against $f_{v_i}^{(i)}$. In fact, $\tilde{L}_i^{\mathcal{A}}$ depends on v because the algorithm $\mathcal{A}$ uses the d -dimensional oracle γ_v , which depends on v (and not only on v_i) and thus algorithm $\mathcal{A}$ could use information returned by $\gamma_{v_j}^{(j)}$, $j \neq i$. In a way our proof shows that using this information cannot help a d -dimensional algorithm on a separable problem, a claim that we find rather intuitive, and which we now formally state and prove.

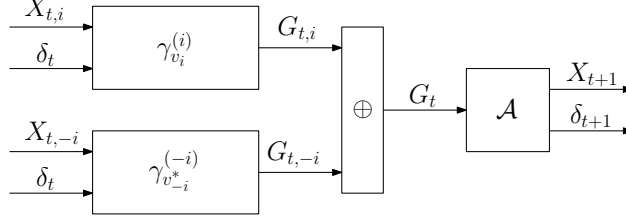
Lemma 4 (“Cross-talk” does not help in separable problems) *Let $V = \times_{i=1}^d V_i$, $(f_v)_{v \in V}$, $(\gamma_v)_{v \in V}$ be separable, $\mathcal{A}$ any d -dimensional algorithm. Then there exist d one-dimensional algorithms, $\mathcal{A}_i^*$, $1 \leq i \leq d$ such that*

$$\max_{v \in V} L^{\mathcal{A}}(v) \geq \max_{v_1 \in V_1} L_1^{\mathcal{A}_1^*}(v_1) + \dots + \max_{v_d \in V_d} L_d^{\mathcal{A}_d^*}(v_d). \quad (37)$$

□

PROOF We will explicitly construct the one-dimensional algorithms, using $\mathcal{A}$. The difficulty is that $\mathcal{A}$ is d -dimensional, and the i th one-dimensional algorithms can only interact with the one-dimensional oracle that depends on v_i but does not depend on $v_{-i} := (v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_d)$. Hence, to use $\mathcal{A}$ we need to supply some values v_{-i}^* replacing v_{-i} so that we can use the full d -dimensional oracle, which $\mathcal{A}$ needs.

Before the construction, we need one more notational convention: Slightly abusing notation, we let $v = (v_i, v_{-i})$ and when writing (v_i, v_{-i}) as the argument of some function g , instead of $g((v_i, v_{-i}))$ we will write $g(v_i, v_{-i})$. The decomposition of a vector into one component and all the others will also be used for other d -dimensional vectors (not only for $v \in V$).


 Figure 3: The construction of algorithm $\mathcal{A}_i^*$ used in the proof of Lemma 4.

To define $\mathcal{A}_i^*$, consider the solution of the following max-min problem:

$$\max_{v_i} \min_{v_{-i}} \tilde{L}_i^{\mathcal{A}}(v_i, v_{-i}).$$

Let the optimal solution of this problem be denoted by $(\hat{v}_i^*, v_{-i}^*)$, thus this max-min problem defines the values replacing the missing values v_{-i} . We also collect $(\hat{v}_i^*)_i$ into the vector $\hat{v}^* \in V$.

Now, algorithm $\mathcal{A}_i^*$ is constructed as illustrated on Fig. 3. Fix $v_i \in V_i$. Then, algorithm $\mathcal{A}_i^*$ interacts with oracle $\gamma_{v_i}^{(v_i)}$ as follows: In each round t , algorithm $\mathcal{A}_i^*$ produces a pair $(X_t, \delta_t) \in \mathcal{K} \times [0, 1)$. In particular, in the first round, X_1, δ_1 is the output of $\mathcal{A}$ in the first round. In round $t + 1$, given the pair X_t, δ_t produced in the previous round, the i th component of X_t and δ_t are fed to oracle $\gamma_{v_i}^{(i)}$, whose output we name $G_{t,i}$. The other components of $X_t, X_{t,-i}$, together with δ_t are fed to oracle $\gamma_{v_{-i}^*}^{(-i)}$, which produces a $d - 1$ -dimensional vector, which we call $G_{t,-i}$. The values $G_{t,i}, G_{t,-i}$ are put together to form the d -dimensional vector $G_t = (G_{t,i}, G_{t,-i})$, which is fed to algorithm $\mathcal{A}$. We then set (X_{t+1}, δ_{t+1}) to be equal to the output of $\mathcal{A}$. At the end of the n rounds, $\mathcal{A}$ is queried to produce $\hat{X}_n$, whose i th component, $\hat{X}_{n,i}$, is returned as the output of $\mathcal{A}_i^*$.

By construction, $L_i^{\mathcal{A}_i^*}(v_i) = \tilde{L}_i^{\mathcal{A}}(v_i, v_{-i}^*)$. Now, notice that

$$\max_{v_i \in V_i} \tilde{L}_i^{\mathcal{A}}(v_i, v_{-i}^*) = \tilde{L}_i^{\mathcal{A}}(\hat{v}_i^*, v_{-i}^*) \leq \tilde{L}_i^{\mathcal{A}}(\hat{v}_i^*, \hat{v}_{-i}^*) = \tilde{L}_i^{\mathcal{A}}(\hat{v}^*),$$

where the equality uses the definition of $\hat{v}_i^*$, while the inequality uses the definition of v_{-i}^* . Thus,

$$\sum_{i=1}^d \max_{v_i \in V_i} L_i^{\mathcal{A}_i^*}(v_i) \leq \sum_{i=1}^d \tilde{L}_i^{\mathcal{A}}(\hat{v}^*) = L^{\mathcal{A}}(\hat{v}^*) \leq \max_{v \in V} L^{\mathcal{A}}(v),$$

which was the claim to be proven. ■

Now, let

$$\mathcal{F}^{(i)} = \{f_{v_i} : v_i \in V_i\}, \quad i = 1, \dots, d$$

and define

$$\Delta_n^{(d)*} = \Delta_n(\mathcal{F}_{\text{sep}}, c_1, c_2).$$

The next result follows easily from the previous lemma:

Lemma 5 *Let $\|\cdot\| = \|\cdot\|_2$ in the definition of the type-I oracles. Then, we have that*

$$\Delta_n^{(d)*} \geq \sum_{i=1}^d \Delta_n^*(\mathcal{F}^{(i)}, c_1/d^{1/2}, c_2/d).$$

□

PROOF By our earlier remark, $\Gamma_{\text{sep}}(f_v, c_1/d^{1/2}, c_2/d) \subset \Gamma(f, c_1, c_2)$. Hence,

$$\Delta_n^{(d)*} = \inf_{\mathcal{A}} \sup_{v \in V} \sup_{\gamma \in \Gamma(f_v, c_1, c_2)} \Delta_n^{\mathcal{A}}(f_v, \gamma) \geq \inf_{\mathcal{A}} \sup_{v \in V} \sup_{\gamma \in \Gamma_{\text{sep}}(f_v, c_1/d^{1/2}, c_2/d)} \Delta_n^{\mathcal{A}}(f_v, \gamma). \quad (38)$$

For each $i = 1, \dots, d$, pick $\gamma_{v_i}^{(i)} \in \Gamma(f_{v_i}, c_1/d^{1/2}, c_2/d)$ such that $\Delta_n^*(\mathcal{F}^{(i)}, c_1/d^{1/2}, c_2/d) = \inf_{\mathcal{A}} \sup_{v_i \in V_i} \Delta_n^{\mathcal{A}}(f_{v_i}, \gamma_{v_i}^{(i)})$. For $v \in V$, let $\gamma_v \in \Gamma_{\text{sep}}(f_v, c_1/d^{1/2}, c_2/d)$ be the oracle whose “components” are $\gamma_{v_i}^{(i)}, i = 1, \dots, d$. Now, by Lemma 4,

$$\sup_{v \in V} \Delta_n^{\mathcal{A}}(f_v, \gamma_v) \geq \sum_{i=1}^d \inf_{\mathcal{A}} \sup_{v_i \in V_i} \Delta_n^{\mathcal{A}}(f_{v_i}, \gamma_{v_i}^{(i)}) = \sum_{i=1}^d \Delta_n^*(\mathcal{F}^{(i)}, c_1/d^{1/2}, c_2/d).$$

This, together with $\sup_{v \in V} \sup_{\gamma \in \Gamma_{\text{sep}}(f_v, c_1/d^{1/2}, c_2/d)} \Delta_n^{\mathcal{A}}(f_v, \gamma) \geq \sup_{v \in V} \Delta_n^{\mathcal{A}}(f_v, \gamma_v)$ and (38) gives the desired result. ■

Main proof: Let $\mathcal{K} \subset \mathbb{R}^d$, such that $\times_i \mathcal{K}_i \subset \mathcal{K}$, $\{\pm 1\} \subset \mathcal{K}_i \subset \mathbb{R}$, $\mathcal{F}_d = \mathcal{F}_{L,0}(\mathcal{K})$, where recall that $L \geq 1/2$. For any $1 \leq i \leq d$, $x_i \in \mathcal{K}_i$,

$$f_{v_i}^{(i)}(x_i) := \epsilon(x_i - v_i) + 2\epsilon^2 \ln\left(1 + e^{-\frac{x_i - v_i}{\epsilon}}\right). \quad (39)$$

i.e., $f_{v_i}^{(i)}$ is like in the one-dimensional lower bound proof (cf. (20)). Note that $f_v \in \mathcal{F}_d$ since f_v is separable, so its Hessian is diagonal and from our earlier calculation we know that $0 \leq \frac{\partial^2}{\partial x_i^2} f_{v_i}^{(i)}(x_i) \leq 1/2$. Let $\Delta_n^{(d)*}$ denote the minimax error $\Delta_n^*(\mathcal{F}_d, C_1 \delta^p, \frac{C_2}{\delta^q})$ for the d -dimensional family of functions $\mathcal{F}_d$. Let $\mathcal{F}^{(i)} = \{f_{-1}^{(i)}, f_{+1}^{(i)}\}$. As it was noted above, $f_v \in \mathcal{F}_d$ for any $v \in \{\pm 1\}^d$. Hence, by Lemma 5,

$$\Delta_n^{(d)*} \geq \sum_{i=1}^d \Delta_n^*\left(\mathcal{F}^{(i)}, \frac{C_1}{\sqrt{d}} \delta^p, \frac{C_2}{d} \delta^{-q}\right). \quad (40)$$

Derivation of rates:

Plugging the bound derived in (35) for 1-dimensional setting into the bound in (40), we obtain the following result for any $p, q > 0$:

$$\Delta_n^{(d)*} \geq \sqrt{d} \Delta_n^*. \quad (41)$$

The above bound simplifies to the following for the case where $p = 1$ and $q = 2$:

$$\Delta_n^{(d)*} \geq \frac{(4C_1^2 C_2)^{1/4}}{3\sqrt{3}} d^{1/2} n^{-1/4}.$$

On the other hand, for the case $p = q = 2$, we obtain

$$\Delta_n^{(d)*} \geq \frac{9}{10} \left(\frac{C_1 C_2}{100}\right)^{1/3} d^{1/2} n^{-1/3}.$$

D.2 Proof of Theorem 2 for strongly convex + smooth function class $\mathcal{F}_{L,1}(\mathcal{K})$.

We follow the notational convention used earlier for convex functions in one dimension. Let $\mathcal{F} = \mathcal{F}_{L,1}(\mathcal{K})$, where $L \geq 1$ and $\mathcal{K}$ includes ± 1 . We consider functions f_v , for $v \in \{-1, +1\}$, defined as

$$f_v(x) := \frac{1}{2}x^2 - v\epsilon x, \quad x \in \mathcal{K}. \quad (42)$$

It is easy to see that $\{f_+, f_-\} \subset \mathcal{F}$.

Clearly, f_v is minimized at $x_v^* = v\epsilon$. By the definition of f_v , we have

$$f_v(x) - f_v(x_v^*) \geq \frac{\epsilon^2}{2} \mathbf{1}\{xv < 0\}. \quad (43)$$

We will consider the oracles γ_v defined as

$$\gamma_v(x) = x - v\epsilon + v \min(\epsilon, C_1 \delta^p) + \xi, \quad (44)$$

where $\xi \sim \mathcal{N}(0, \frac{C_2}{\delta^q})$; as with f_v , we will also use γ_+ (γ_-) to denote γ_{+1} (resp., γ_{-1}). The oracle is indeed a (c_1, c_2) type-I oracle, with $c_1(\delta) = C_1 \delta^p$ and $c_2(\delta) = \frac{C_2}{\delta^q}$.

Using arguments similar to those in the proof of lower bound for convex functions, we obtain

$$\Delta_n^{(1)*} := \Delta_n^* \geq \inf_{\mathcal{A}} \frac{\epsilon^2}{2} \left(1 - \left(\frac{1}{2} D_{\text{kl}}(P_+ \| P_-) \right)^{\frac{1}{2}} \right), \quad (45)$$

Note that P_+ (resp. P_-) is $\mathbb{P}$ conditioned on the event $V = +1$ (resp. $V = -1$).

Observe that, for any $x \in \mathbb{R}$, $f'_-(x) - f'_+(x) = 2\epsilon$ and hence

$$|\gamma_+(x) - \gamma_-(x)| = |f'_+(x) - \min(\epsilon, C_1 \delta^p) - (f'_-(x) + \min(\epsilon, C_1 \delta^p))| = 2(\epsilon - C_1 \delta^p)^+. \quad (46)$$

From the foregoing,

$$D_{\text{kl}}(P_+^t(\cdot | g_{1:t-1}) \| P_-^t(\cdot | g_{1:t-1})) \leq \frac{2\{(\epsilon - C_1 \delta^p)^+\}^2 \delta_t^q}{C_2}, \quad (47)$$

where the inequality (47) follows from (46). Thus, we obtain

$$D_{\text{kl}}(P_+ \| P_-) \leq 2n \sup_{\delta > 0} \frac{\{(\epsilon - C_1 \delta^p)^+\}^2 \delta^q}{C_2}. \quad (48)$$

Substituting the above bound into (45), we obtain

$$\Delta_n^{(1)*} \geq \frac{\epsilon^2}{2} \left(1 - \sqrt{n} \sup_{\delta > 0} \frac{(\epsilon - C_1 \delta^p)^+ \delta^{q/2}}{\sqrt{C_2}} \right). \quad (49)$$

Derivation of the rates uniformly for all δ : As in the proof of the lower bound for $\mathcal{F}_{L,0}(\mathcal{K})$, we replace the positive part function in (49) and optimize over δ to obtain that the right-hand side of (48) is optimized by

$$\delta_* = \left(\frac{\epsilon q}{C_1(2p+q)} \right)^{1/p}. \quad (50)$$

From the above, we have

$$\Delta_n^{(1)*} \geq \frac{\epsilon^2}{2} \left(1 - \sqrt{n} \frac{(\epsilon - C_1 \delta_*^p) \delta_*^{q/2}}{\sqrt{C_2}} \right) = \frac{\epsilon^2}{2} \left(1 - \sqrt{n} K_1 \epsilon^{\frac{p+q}{p}} \right),$$

where $K_1 = \frac{p}{\sqrt{C_2}(p+\frac{q}{2})} \left(\frac{q}{2C_1(p+\frac{q}{2})} \right)^{\frac{q}{2p}}$.

Plugging in $\epsilon = \left(\frac{4p}{(6p+q)\sqrt{n}K_1} \right)^{\frac{2p}{2p+q}}$, we obtain

$$\Delta_n^{(1)*} \geq 2^{\frac{2p-q}{2p+q}} \frac{(2p+q)^3}{q^{\frac{2q}{2p+q}} (6p+q)^{\frac{6p+q}{2p+q}}} C_1^{\frac{2q}{2p+q}} C_2^{\frac{2p}{2p+q}} n^{-\frac{2p}{2p+q}}. \quad (51)$$

Now, when $q = 2$ and $p = 1$, the lower bound in (51) simplifies to

$$\Delta_n^{(1)*} \geq \frac{1}{2} C_1 C_2^{1/2} n^{-1/2}.$$

On the other hand, for $p = q = 2$, we obtain

$$\Delta_n^{(1)*} \geq 27 \left(\frac{2}{7^7} \right)^{\frac{1}{3}} C_1^{2/3} C_2^{2/3} n^{-2/3}.$$

Generalization to d dimensions: Recall that in this result, $\|\cdot\| = \|\cdot\|_2$. The proof in d dimensions for strongly convex functions is the same as that for the case of smooth convex functions with the difference that we use (42) in defining the functions $f_{v_i}^{(i)}$. Then, for any $v \in \{\pm 1\}^d$, $f_v \in \mathcal{F}_{L,1}(\mathcal{K})$. Indeed, $f_v(x) = \sum_{i=1}^d f^{(i)}(x_i)$, hence $\nabla^2 f_v(x) = I_{d \times d}$, where $I_{d \times d}$ is the $d \times d$ identity matrix. Thus, $\lambda_{\min}(\nabla^2 f_v(x)) = \lambda_{\max}(\nabla^2 f_v(x)) = 1$. From (40) and (51) we get

$$\Delta_n^{(d)*} \geq \Delta_n^{(1)*}. \quad (52)$$