STOCHASTIC ZEROTH-ORDER OPTIMIZATION UNDER STRONGLY CONVEXITY AND LIPSCHITZ HESSIAN: MINIMAX SAMPLE COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimization of convex functions under stochastic zeroth-order feedback has been a major and challenging question in online learning. In this work we consider the problem of optimizing second-order smooth and strongly convex functions where the algorithm is only accessible to noisy evaluations of the objective function it queries. We provide the first tight characterization for the rate of the minimax simple regret by developing matching upper and lower bounds. We propose an algorithm that features a combination of a bootstrapping stage and a mirror-descent stage. The main innovation of our approach is the usage of a gradient estimation scheme that exploits the local geometry of the objective function, and we provide sharp analysis for the corresponding estimation bounds.

1 INTRODUCTION

Stochastic optimization of an unknown function with access to only noisy function evaluations is a fundamental problem in operations research, optimization, simulation and bandit optimization research, commonly known under the names of *zeroth-order optimization* (Chen et al., 2017), *derivative-free optimization* (Conn et al., 2009; Rios & Sahinidis, 2013) or *bandit optimization* (Bubeck et al., 2021). At a higher level, in stochastic zeroth-order optimization problems an optimization algorithm interacts sequentially with an oracle and obtains noisy function evaluations at queried points every time. The algorithm then produces an approximately optimal solution after T such evaluations, with its performance evaluated by the expected difference between the function values at the approximate optimal solution produced and the optimal solution. A more rigorous formulation of the problem is given in Sec. 2 below.

Existing works and results on stochastic zeroth-order optimization could be broadly categorized into two classes:

1. **Concave functions**. In the first thread of research, the unknown objective function to be optimized is assumed to be *concave* (for maximization problems) or *convex* (for minimization problems). For these problems, with minimal smoothness (e.g. objective function being Lipschitz continuous) it is possible to achieve a sample complexity of $\widetilde{O}(\varepsilon^{-2})$ for an expected optimization error or ε, which is also a polynomial function of domain dimension d; see for example the works of Agarwal et al. (2013); Lattimore & Gyorgy (2021); Bubeck et al. (2021);
2. **Smooth functions**. In the second thread of research,the unknown objective function to be optimized is assumed to be highly *smooth*, but not necessary concave/convex. Typical results assume the objective function is Hölder smooth of order $k \geq 1$, meaning that the $(k-1)$-th derivative of the objective function is Lipschitz continuous. Without additional conditions, the optimal sample complexity with such smoothness assumptions is $\widetilde{O}(\varepsilon^{-(2+d/k)})$ (Wang et al., 2019), which scales exponentially with the domain dimension d.

In this paper, we study the optimal sample complexity of stochastic zeroth-order optimization when the objective function exhibits both convexity and a high degree of smoothness. As we have remarked in the first bullet point above, with convexity and Hölder smoothness of order $k = 1$ (equivalent to

the objective function being Lipschitz continuous), the works of Agarwal et al. (2013); Lattimore & Gyorgy (2021); Bubeck et al. (2021) established an $\widetilde{O}(\varepsilon^{-2})$ upper bound. With higher order of Hölder smoothness, such as $k = 2$ (equivalent to the gradient of the objective being Lipschitz continuous), it is shown that simpler algorithms exist but the sample complexity remains $\widetilde{O}(\varepsilon^{-2})$ (Besbes et al., 2015; Agarwal et al., 2010; Hazan & Levy, 2014), which seemingly suggests the relatively smaller role smoothness plays in the presence of convexity. In this paper we show that with even higher order of Hölder smoothness $k = 3$ (specifically, the Hessian of the objective being Lipschitz continuous), the optimal sample complexity is improved to $O(\varepsilon^{-1.5})$ which is significantly smaller than the sample complexity of the convex-without-smoothness setting $\widetilde{O}(\varepsilon^{-2})$, or the smooth-without-convexity setting $\widetilde{O}(\varepsilon^{-(2+d/3)})$.

We developed several important techniques in this paper to achieve an improved sample complexity of $O(\varepsilon^{-1.5})$ in the case of the objective function being both strongly convex and has Lipschitz Hessian, which might be of interest to other stochastic optimization problems as well. First, we show that when estimating the gradient under a stochastic environment, it could be beneficial to sample with non-isotropic distributions (as opposed to conventional standard Gaussian, or uniform distributions on hyperspheres). Second, we present a new approach to analyze the bias and variance of gradient estimation under hyperellipsoid sampling, which enables obtaining sharp bounds. Third, we present a two-stage bootstrap-type framework for algorithmic designs, which extends steps that requires perturbative assumptions to the full regime. We fully complete the characterization of the asymptotic minimax regret by deriving a lower bound using the KL-divergence approach.

Our results are also related to a special case discussed in Shamir (2013), which shows that for *quadratic* functions it is possible to achieve a sample complexity of $\widetilde{O}(\varepsilon^{-1})$. As quadratic functions are infinitely differentiable with bounded derivatives on orders, they are Hölder smooth of any arbitrary order $k \to \infty$, which could be regarded as an extreme of the results established in this paper which only require $k = 3$.

lower bound	Bach & Perchet (2016)	Zhang et al. (2020)
$\boldsymbol{\Omega(\mathbf{dT^{-\frac{2}{3}}M^{-1}})}$	$O(d^{1.5}T^{-\frac{1}{2}}M^{-\frac{1}{2}})$	$O(dT^{-\frac{1}{2}}M^{-\frac{1}{2}})$
Akhavan et al. (2020)	Novitskii & Gasnikov (2021)	**Ours**
$O(d^2T^{-\frac{2}{3}}M^{-1})$	$O(d^{\frac{5}{3}}T^{-\frac{2}{3}}M^{-1})$	$\boldsymbol{\mathbf{O}(\mathbf{dT^{-\frac{2}{3}}M^{-1}})}$

Table 1: The dependence of simple regret on T (number of function evaluations), d (dimension) and M (parameter describing strong convexity).

Additional Works Recent years have seen increasing attention on exploiting higher order smoothness in bandit optimization. We list our results together with the most relevant work in Table 1. While this line of work also demonstrates the benefit of higher order smoothness in improving the sample complexity, their setting is related but slightly different from what we considered in this work. (See reference therein: Bach & Perchet (2016); Zhang et al. (2020); Akhavan et al. (2020); Novitskii & Gasnikov (2021)). On one hand, the prior work concentrates on projected gradient-alike algorithms, which requires Lipschitz gradient (and we do not). On the other hand, they use generalized Holder condition instead of bounding the Frobenius norm for Lipschitz Hessian as in this paper, which makes the results not directly comparable.

Notations. We use $\nabla\nabla f(\boldsymbol{x})$ to denote the Hessian of f at point $\boldsymbol{x}$. This should not be confused with the notation $\nabla^2 f(\boldsymbol{x})$, which denotes the trace of the Hessian. We use $||\cdot||_2$ to denote vector L2 norms, and $||\cdot||_F$ to denote matrix Frobenius norms. We use I_d to denote the identity matrix, and S^{d-1} to denote the unit hypersphere centered at the origin, both for the d-dimensional Euclidean space $\mathbb{R}^d$. We adopt the conventional notations (i.e., O, Ω, o, and ω) to describe regret bounds in the asymptotic sense with respect to the total number of samples (denoted by T).

2 PROBLEM FORMULATION

We consider the stochastic optimization problem under the class of functions that are strongly convex and have Lipschitz Hessian. The goal in this setting is to design learning algorithms to achieve approximately the global minimum of an unknown objective function $f : \mathbb{R}^d \to \mathbb{R}$.

A learning algorithm $\mathcal{A}$ can interact with the function by adaptively sampling their value for T times, and receive noisy observations. At each time $t \in [T]$, the algorithm selects $\boldsymbol{x}_t \in \mathbb{R}^d$, and receives the following observation,

$$y_t = f(\boldsymbol{x}_t) + w_t, \tag{1}$$

where $\{w_t\}_{t=1}^T$ are independent random variables with zero mean and bounded variance. Formally, the algorithm can be described by a list of conditional distributions where each $\boldsymbol{x}_t$ is selected based on all historical data $\{\boldsymbol{x}_\tau, y_\tau\}_{\tau<t}$ and the corresponding distribution. Then for any t, we assume that $\mathbb{E}[w_t|\{\boldsymbol{x}_\tau, y_\tau\}_{\tau<t}] = 0$ and $\text{Var}[w_t|\{\boldsymbol{x}_\tau, y_\tau\}_{\tau<t}] \leq 1$ for any t.[1] We also adopt a common assumption that the additive noises are subguassian, particularly, $\mathbb{P}[|w_t| > s] \leq 2e^{-s^2}$ for all $s > 0$ and $t \in [T]$.

We assume that the objective function f is second-order differentiable. Furthermore, we impose the following conditions.

- (A1) (Lipschitz Hessian). There exist a constant $\rho \in (0, +\infty)$ such that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, it holds that $\|\nabla\nabla f(\boldsymbol{x}) - \nabla\nabla f(\boldsymbol{x}')\|_F \leq \rho\|\boldsymbol{x}' - \boldsymbol{x}\|_2$, where $||\cdot||_F$ denotes the Frobenius norm;
- (A2) (Strong Convexity). There exists a constant $M \in (0, +\infty)$ such that for any $\boldsymbol{x} \in \mathbb{R}^d$, the minimum eigenvalue of the Hessian $\nabla\nabla f(\boldsymbol{x})$ is greater than M.
- (A3) (Bounded Distance from Initialization to Optimum Point). There exists a constant $R \in (0, +\infty)$ such that the infimum of $f(\boldsymbol{x})$ within the hyperball $\|\boldsymbol{x}\|_2 \leq R$ is identical to the infimum of $f(\boldsymbol{x})$ over the entire $\mathbb{R}^d$.

In the rest of this paper, we let $\mathcal{F}(\rho, M, R)$ denote the set of all second-order differentiable functions that satisfy the above conditions, with corresponding constants given by ρ, M, and R. We aim to find algorithms to achieve asymptotically the following minimax simple regret, which measures the expected difference of the objective function on x_T and the optimum.

$$\mathfrak{R}(T; \rho, M, R) := \inf_{\mathcal{A}} \sup_{f \in \mathcal{F}(\rho, M, R)} \mathbb{E}\left[f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*)\right],$$

where $\boldsymbol{x}^*$ denotes the global minimum point of f.

3 MAIN RESULTS

Theorem 1. *For any dimension d and constants ρ, M, R, the minimax simple regrets are upper bounded by $\mathfrak{R}(T; \rho, M, R) = O\left(\frac{\rho^{\frac{2}{3}}}{M} d T^{-\frac{2}{3}}\right)$ for sufficiently large T.*

Theorem 2. *For any fixed dimension d and constants ρ, M, R, the minimax simple regrets are lower bounded by $\mathfrak{R}(T; \rho, M, R) = \Omega\left(\frac{\rho^{\frac{2}{3}}}{M} d T^{-\frac{2}{3}}\right)$ for sufficiently large T when the additive noises $w_1, ..., w_T$ are standard Gaussian variables.*

4 PROOF OF THEOREM 1

The proposed algorithm operates in two stages (see Algorithm 3). In the first stage, the algorithm uses half of the samples to obtain a rough estimation of the global minimum point. We ensure that the estimation in the first state is sufficiently accurate with high probability, so that in the following final stage, the objective function can be approximated by a quadratic function and the resulting approximation error can be bounded using tensor analysis.

[1] If the variances of w_t's are bounded by a different constant, all our results can be reproduced by normalizing the values of f.

4.1 KEY TECHNIQUES AND THE FINAL STAGE

We first present the key steps of our algorithm, which relies on two subroutines presented in Algorithm 1 and Algorithm 2, i.e., GradientEst and HessianEst. These subroutines estimate the (linearly transformed) gradients and Hessian functions of f at any given point by randomly sampling the values of f on hyperellipsoids. The key ingredient of our proof is the sharp characterizations for the biases and variances of the GradientEst estimator, stated in Theorem 3.

Algorithm 1

procedure GRADIENTEST($\boldsymbol{x}, Z, n$) ▷ Z can be a $d \times d$ matrix
  for $k \leftarrow 1$ to n **do**
    Let $\boldsymbol{u}_k$ be a point sampled uniformly randomly from the standard hypersphere S^{d-1}, let y_+, y_- be samples of f at $\boldsymbol{x} + Z\boldsymbol{u}_k$ and $\boldsymbol{x} - Z\boldsymbol{u}_k$, respectively
    Let $\boldsymbol{g}_k = \frac{d}{2}(y_+ - y_-)\boldsymbol{u}_k$
  end for
  return $\widehat{\boldsymbol{g}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{g}_k$
end procedure

Algorithm 2

procedure HESSIANEST($\boldsymbol{x}, r, n$)
  for $k \leftarrow 1$ to n **do**
    Let $\boldsymbol{u}_k$ be a point sampled uniformly randomly from the standard hypersphere S^{d-1}, let y^+, y_-, y be be samples of f at $\boldsymbol{x} + r\boldsymbol{u}_k$, $\boldsymbol{x} - r\boldsymbol{u}_k$, and $\boldsymbol{x}$, respectively
    $H_k = d(d+2)(\boldsymbol{u}\boldsymbol{u}^\intercal - \frac{I_d}{d+2})\frac{(y_+ + y_- - 2y)}{4r^2}$
  end for
  return $\widehat{H} = \frac{1}{n}\sum_{k=1}^{n} H_k$
end procedure

Theorem 3. *For any fixed inputs $\boldsymbol{x}$, Z, n, and any function f satisfying the Lipschitz Hessian condition with parameter ρ, the output $\widehat{\boldsymbol{g}}$ returned by the GradientEst subroutine satisfies the following properties*

$$||\mathbb{E}[\widehat{\boldsymbol{g}}] - Z\nabla f(\boldsymbol{x})||_2 \leq \frac{\lambda_Z^3 \rho \sqrt{d}}{2(d+2)}, \tag{2}$$

$$\mathrm{Tr}\left(\mathrm{Cov}[\widehat{\boldsymbol{g}}]\right) \leq \frac{2d}{n}||Z\nabla f(\boldsymbol{x})||_2^2 + \frac{d^2}{18n}\left(\rho\lambda_Z^3\right)^2 + \frac{d^2}{2n}, \tag{3}$$

where λ_Z is the largest singular value of Z.

We also provide a rough estimate for the errors of the Hessian Estimator in Theorem 4.

Theorem 4. *For any fixed inputs $\boldsymbol{x}$, H, n, and any function f satisfying the Lipschitz Hessian condition with parameter ρ, the output $\widehat{H}$ returned by the HessianEst subroutine satisfies the following conditions.*

$$\left|\left|\left(\mathbb{E}[\widehat{H}] - \nabla\nabla f(\boldsymbol{x})\right)\right|\right|_F \leq C_{\rho,d} r, \tag{4}$$

$$\mathrm{Tr}\left(\mathrm{Cov}[\widehat{H}]\right) \leq \frac{C_{\rho,d}^2}{nr^4}(1 + r^6), \tag{5}$$

where $C_{\rho,d}$ depends polynomially on ρ and d.

We postpone the proofs of the above theorems to Section 4.3 and Appendix A and proceed to describe how these results are used in the final stage.

For brevity, let $\epsilon \triangleq \frac{\rho^{\frac{2}{3}}}{M} d T^{-\frac{2}{3}}$ to be the minimax regret we aim to achieve, and let $\boldsymbol{x}_0$ denote the estimator $\boldsymbol{x}$ stored at the end of the first stage. For now we assume that the first stage is designed

Algorithm 3

procedure OPT(T, ρ, M)
Let $\boldsymbol{x} = \mathbf{0}$
The First Stage:
for $k \leftarrow 1$ to $\lfloor T^{0.1} \rfloor$ **do**
Let $n = \lfloor \frac{T^{0.9}}{10} \rfloor$
Let $\widehat{H}_k = \text{HessianEst}(\boldsymbol{x}, n^{-\frac{1}{6}}, n)$, Z_{H_k} be any symmetric matrix such that $Z_{H_k}^2 = \widehat{H}_k^{-1}$
Let $\widehat{\boldsymbol{g}} = \text{GradientEst}(\boldsymbol{x}, n^{-\frac{1}{6}} Z_{H_k}, n)$, $\boldsymbol{r}_k = n^{\frac{1}{6}} Z_{H_k} \widehat{\boldsymbol{g}}$
$\boldsymbol{x} = \text{Aggregate}(\boldsymbol{x}, \boldsymbol{r}_k, \frac{M}{2\rho})$
end for
The Final Stage:
Let $n_{\mathrm{f}} = \lfloor \frac{T}{10} \rfloor$
Let $\widehat{H} = \text{HessianEst}(\boldsymbol{x}, n_{\mathrm{f}}^{-\frac{1}{6}}, n_{\mathrm{f}})$, Z_H be any symmetric matrix such that $Z_H^2 = \widehat{H}^{-1}$, λ_{Z_H} be the largest singular value of Z_H
Let $Z = n_{\mathrm{f}}^{-\frac{1}{6}} \rho^{-\frac{1}{3}} d^{\frac{1}{2}} Z_H / \lambda_{Z_H}$
Let $\widehat{\boldsymbol{g}} = \text{GradientEst}(\boldsymbol{x}, Z, n_{\mathrm{f}})$, $\boldsymbol{r} = \widehat{H}^{-1} Z^{-1} \widehat{\boldsymbol{g}}$
return $\boldsymbol{x} + \boldsymbol{r}$
end procedure

procedure AGGREGATE($\boldsymbol{x}, \boldsymbol{r}, L$)
if $||\boldsymbol{r}_k||_2 \le L$ **then**
return $\boldsymbol{x} = \boldsymbol{x} + \boldsymbol{r}_k$
else
return Let $\boldsymbol{x} = \boldsymbol{x} + \frac{L \boldsymbol{r}_k}{||\boldsymbol{r}_k||_2}$
end if
end procedure

such that $f(\boldsymbol{x}_0) \le f(\boldsymbol{x}^*) + \frac{M}{2}\epsilon^{0.6}$ w.p. $1 - o(\epsilon)$, which will be proved in Section 4.2. By strong convexity, this assumption implies that $||\boldsymbol{x}_0 - \boldsymbol{x}^*||_2 \le \epsilon^{0.3}$.

We perform a Hessian estimation near $\boldsymbol{x}_0$ using the HessianEst subroutine with $r = n_{\mathrm{f}}^{-\frac{1}{6}} = O(T^{-\frac{1}{6}})$ and $n = n_{\mathrm{f}} = O(T)$ samples. From Theorem 4, this results in an expected estimation error of $O(C_{\rho,d} T^{-\frac{1}{6}})$ for sufficiently large T. Note that asymptotically we have $\epsilon^{0.24} = \omega(C_{\rho,d} T^{-\frac{1}{6}})$. From the subgaussian condition of the noise variables and the Lipschitz Hessian condition, we have the upper bound $||\widehat{H} - \nabla\nabla f(\boldsymbol{x})||_F = O(\epsilon^{0.24})$ for all $||\boldsymbol{x} - \boldsymbol{x}_0|| \le \frac{1}{\rho}\epsilon^{0.24}$ with high probability (i.e., $1 - o(\epsilon)$). Recall that we have $||\boldsymbol{x}_0 - \boldsymbol{x}^*||_2 = O(\epsilon^{0.3})$. The same bound also applies to all points $\boldsymbol{x}$ with $||\boldsymbol{x} - \boldsymbol{x}^*|| \le ||\boldsymbol{x}_0 - \boldsymbol{x}^*||_2$ for sufficiently large T.

The above statements are used to show that with high probability, $||\nabla f(\boldsymbol{x}_0) - \widehat{H}(\boldsymbol{x}_0 - \boldsymbol{x}^*)||_2^2 = O(\epsilon^{1.08})$, where the RHS is bounded by $o(\epsilon)$ for large T. This is obtained by integrating the second-order differential of f along the line segment from $\boldsymbol{x}^*$ to $\boldsymbol{x}_0$. Formally, let $\boldsymbol{x}_\alpha \triangleq (1-\alpha)\boldsymbol{x}^* + \alpha \boldsymbol{x}_0$, we have

$$\nabla f(\boldsymbol{x}_0) - \widehat{H}(\boldsymbol{x}_0 - \boldsymbol{x}^*) = \int_0^1 \left(\nabla\nabla f(\boldsymbol{x}_\alpha) - \widehat{H} \right) \cdot (\boldsymbol{x}_0 - \boldsymbol{x}^*) d\alpha.$$

Thus, by triangle inequality of L_2 norms,

$$\begin{aligned} ||\nabla f(\boldsymbol{x}_0) - \widehat{H}(\boldsymbol{x}_0 - \boldsymbol{x}^*)||_2 &\le \int_0^1 \left|\left| \left(\nabla\nabla f(\boldsymbol{x}_\alpha) - \widehat{H} \right) \cdot (\boldsymbol{x}_0 - \boldsymbol{x}^*) \right|\right|_2 d\alpha \\ &\le \int_0^1 \left|\left| \nabla\nabla f(\boldsymbol{x}_\alpha) - \widehat{H} \right|\right|_F \cdot ||\boldsymbol{x}_0 - \boldsymbol{x}^*||_2 d\alpha \\ &= O(\epsilon^{0.54}) \end{aligned}$$

with high probability. Given this fact, if we update the estimator by inverting the measured Hessian, by strong convexity, the resulting regret is dominated by the errors from gradient estimation.

We sample $\widehat{\boldsymbol{g}}$ using the GradientEst subroutine to estimate $Z\nabla f(\boldsymbol{x}_0)$. By directly applying Theorem 3 with the design parameters specified in the Opt function in Algorithm 1, one can show that the overall resulting regret is still $O(\epsilon)$. In particular, this relies on a possibly non-trivial observation that the assumption $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x}^*) + \frac{M}{2}\epsilon^{0.6}$ implies that the L2 norm of $Z_H\nabla f(\boldsymbol{x}_0)$ is $o(1)$ with high probability, so that the covariance of $\widehat{\boldsymbol{g}}$ is only dominated by the third term for large T.

4.2 The First Stage

Now we illustrate the intuition that one can achieve a simple regret of $\epsilon^{0.6}$ with $O(T)$ samples with high probability. Note that here we have a relaxed requirement compared to the final stage. The design parameters we choose enables that once the estimator has a function value that is sufficiently close to $f(\boldsymbol{x}^*)$, the simple regret decays exponentially. Besides, one can show that it takes finitely many iterations for $\boldsymbol{x}$ to get arbitrarily close to $\boldsymbol{x}^*$. Given these two facts, it remains to check the subclass of function instances where the Hessian contains arbitrarily large eigenvalues. These instances can be treated by exploiting the Lipschitz Hessian condition and the perturbative matrix inversion formula. One can show that in fact large eigenvalues improve the convergence of the proposed algorithm as the function f in the corresponding dimensions can be better approximated by quadratic functions, and the optimization process in those dimension approximately do not interfere with ones on other dimensions.

4.3 Proof of Theorem 3

Proof. To prove inequality (2), we investigate the following function

$$\boldsymbol{G}(r;\boldsymbol{x}) \triangleq \mathbb{E}_{\boldsymbol{u}\sim\text{Unif}(S^{d-1})}\left[\frac{d}{2r}(f(\boldsymbol{x}+r\boldsymbol{u}) - f(\boldsymbol{x}-r\boldsymbol{u}))\boldsymbol{u}\right],$$

where $\text{Unif}(S^{d-1})$ denotes the uniform distribution on S^{d-1}. Recall that in our algorithm we have $\mathbb{E}[\widehat{\boldsymbol{g}}] = r\boldsymbol{G}(r;\boldsymbol{x})$ if $Z = rI_d$ for some $r \in (0, +\infty)$, and by differentiability we have $\nabla f(\boldsymbol{x}) = \lim_{z\to 0^+}\boldsymbol{G}(z;\boldsymbol{x})$. Under this condition, we can bound $||\mathbb{E}[\widehat{\boldsymbol{g}}] - r\nabla f(\boldsymbol{x})||_2$ by integration, i.e.,

$$||\mathbb{E}[\boldsymbol{g}] - r\nabla f(\boldsymbol{x})||_2 = r\left|\left|\boldsymbol{G}(r;\boldsymbol{x}) - \lim_{z\to 0^+}\boldsymbol{G}(z;\boldsymbol{x})\right|\right|_2 \leq r\int_{0^+}^{r}\left|\left|\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x})\right|\right|_2 dz. \tag{6}$$

Note that $\boldsymbol{G}(z;\boldsymbol{x})$ can be written into the following equivalent form.

$$\boldsymbol{G}(z;\boldsymbol{x}) = \frac{\int_{S^{d-1}}\frac{d}{2z}(f(\boldsymbol{x}+z\boldsymbol{u}) - f(\boldsymbol{x}-z\boldsymbol{u}))\mathbf{dA}}{\int_{S^{d-1}}||\mathbf{dA}||_2},$$

where the integration is with respect to $\boldsymbol{u}$ and the surface S^{d-1} is oriented with normal vectors pointing outward. The differential of $\boldsymbol{G}(z;\boldsymbol{x})$ over z can be written as

$$\begin{aligned}\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x}) &= \frac{\int_{S^{d-1}}\frac{\partial}{\partial z}\left(\frac{d}{2z}(f(\boldsymbol{x}+z\boldsymbol{u}) - f(\boldsymbol{x}-z\boldsymbol{u}))\right)\mathbf{dA}}{\int_{S^{d-1}}||\mathbf{dA}||_2} \\ &= \frac{\int_{S^{d-1}}\left(-\frac{d}{2z^2}(f(\boldsymbol{x}+z\boldsymbol{u}) - f(\boldsymbol{x}-z\boldsymbol{u})) + \frac{d}{2z}\boldsymbol{u}\cdot(\nabla f(\boldsymbol{x}+z\boldsymbol{u}) + \nabla f(\boldsymbol{x}-z\boldsymbol{u}))\right)\mathbf{dA}}{\int_{S^{d-1}}||\mathbf{dA}||_2}.\end{aligned}$$

The gist of this proof is to note that for any $\boldsymbol{u} \in S$ we have $\boldsymbol{u}$ and $\mathbf{dA}$ are parallel, so the second term in the integral above on the numerator can be written as $\int_{S^{d-1}}\frac{d}{2z}\boldsymbol{u}(\nabla f(\boldsymbol{x}+z\boldsymbol{u}) + \nabla f(\boldsymbol{x}-z\boldsymbol{u}))\cdot\mathbf{dA}$. Hence, by divergence theorem, we have

$$\begin{aligned}\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x}) &= \frac{\int_{B^d}\nabla_{\boldsymbol{u}}\cdot\left(-\frac{d}{2z^2}I_d\left(f(\boldsymbol{x}+z\boldsymbol{u}) - f(\boldsymbol{x}-z\boldsymbol{u})\right) + \left(\nabla f(\boldsymbol{x}+z\boldsymbol{u}) + \nabla f(\boldsymbol{x}-z\boldsymbol{u})\right)\frac{d}{2z}\boldsymbol{u}\right)\mathbf{dV}}{\int_{S^{d-1}}||\mathbf{dA}||_2} \\ &= \frac{d}{2}\cdot\frac{\int_{B^d}\boldsymbol{u}(\nabla^2 f(\boldsymbol{x}+z\boldsymbol{u}) - \nabla^2 f(\boldsymbol{x}-z\boldsymbol{u}))\mathbf{dV}}{\int_{S^{d-1}}||\mathbf{dA}||_2},\end{aligned} \tag{7}$$

where B^d denotes the standard hyperball.

Now consider any unit vector $\boldsymbol{e}$. Let $\boldsymbol{u_e}$ denote the reflection of $\boldsymbol{u}$ with respect to the hyperplane orthogonal to $\boldsymbol{e}$, i.e., $\boldsymbol{u_e} \triangleq \boldsymbol{u} - 2\boldsymbol{u} \cdot \boldsymbol{e}\boldsymbol{e}$. Because the hyperball B is invariant under the reflection $\boldsymbol{u} \to \boldsymbol{u_e}$, equation (7) can also be written as

$$\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x}) = \frac{d}{2} \cdot \frac{\int_{B^d} \boldsymbol{u_e}(\nabla^2 f(\boldsymbol{x}+z\boldsymbol{u_e}) - \nabla^2 f(\boldsymbol{x}-z\boldsymbol{u_e}))\mathbf{dV}}{\int_{S^{d-1}} ||\mathbf{dA}||_2}. \tag{8}$$

Hence, by averaging equation (7) and (8), we have

$$\begin{aligned}\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x}) \cdot \boldsymbol{e} &= \frac{d}{4}\left(\frac{\int_{B^d}(\boldsymbol{u}(\nabla^2 f(\boldsymbol{x}+z\boldsymbol{u}) - \nabla^2 f(\boldsymbol{x}-z\boldsymbol{u})) \\ + \boldsymbol{u_e}(\nabla^2 f(\boldsymbol{x}+z\boldsymbol{u_e}) - \nabla^2 f(\boldsymbol{x}-z\boldsymbol{u_e})))\mathbf{dV}}{\int_{S^{d-1}} ||\mathbf{dA}||_2}\right) \cdot \boldsymbol{e} \\ &= \frac{d}{4}\left(\frac{\int_{B^d} \boldsymbol{u}\cdot\boldsymbol{e}((\nabla^2 f(\boldsymbol{x}+z\boldsymbol{u}) - \nabla^2 f(\boldsymbol{x}+z\boldsymbol{u_e})) \\ - (\nabla^2 f(\boldsymbol{x}-z\boldsymbol{u}) - \nabla^2 f(\boldsymbol{x}-z\boldsymbol{u_e})))\mathbf{dV}}{\int_{S^{d-1}} ||\mathbf{dA}||_2}\right).\end{aligned} \tag{9}$$

By the Lipschitz Hessian condition and Cauchy's inequality, the difference between the differential terms above can be bounded as follows.

$$\begin{aligned}|\nabla^2 f(\boldsymbol{x}\pm z\boldsymbol{u}) - \nabla^2 f(\boldsymbol{x}\pm z\boldsymbol{u_e})| &\le \sqrt{d}||\nabla\nabla f(\boldsymbol{x}\pm z\boldsymbol{u}) - \nabla\nabla f(\boldsymbol{x}\pm z\boldsymbol{u_e})||_F \\ &\le \rho\sqrt{d}||z\boldsymbol{u} - z\boldsymbol{u_e}||_2 \\ &= 2z\rho\sqrt{d}\boldsymbol{u}\cdot\boldsymbol{e}.\end{aligned} \tag{10}$$

Consequently,

$$\begin{aligned}\left|\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x})\cdot\boldsymbol{e}\right| &\le \frac{z\rho d\sqrt{d}\int_{B^d}(\boldsymbol{u}\cdot\boldsymbol{e})^2\,\mathbf{dV}}{\int_{S^{d-1}}||\mathbf{dA}||_2} \\ &= \frac{z\rho\sqrt{d}}{d+2}.\end{aligned}$$

Note that $\boldsymbol{e}$ can be any unit vector. We have essentially bounded the $L2$ norm of $\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x})$, i.e., $\left|\left|\frac{d}{dz}\boldsymbol{G}(z;\boldsymbol{x})\right|\right|_2 \le \frac{z\rho\sqrt{d}}{d+2}$. As mentioned earlier, when $Z = rI_d$ inequality (2) is obtained by applying this gradient-norm bound to inequality (6) .

For general input matrix Z, we can view GradientEst as a subroutine that operates on the same function f but with a linear transformation applied to the input domain. Formally, let $f'(\boldsymbol{y}) \triangleq f(\boldsymbol{x} + \frac{Z}{\lambda_Z}(\boldsymbol{y}-\boldsymbol{x}))$. We have that f' satisfies the Lipschitz Hessian condition with parameter ρ as well. Therefore, inequality (2) can be obtained following the same analysis by replacing f with f' and Z with $\lambda_Z I_d$.

Now we present the proof for inequality (3). Formally, let w_+, w_- be two independent samples of additive noises. Then the trace of covariance matrix of $\widehat{\boldsymbol{g}}$ can upper bounded using the second moments of single measurements.

$$\begin{aligned}\mathrm{Tr}\left(\mathrm{Cov}[\widehat{\boldsymbol{g}}]\right) &\le \frac{1}{n}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(S^{d-1}),w_+,w_-}\left[\left(\frac{d}{2}\right)^2(f(\boldsymbol{x}+Z\boldsymbol{u}) - f(\boldsymbol{x}-Z\boldsymbol{u}) + w_- - w_-)^2\right] \\ &= \frac{d^2}{4n}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(S^{d-1})}\left[(f(\boldsymbol{x}+Z\boldsymbol{u}) - f(\boldsymbol{x}-Z\boldsymbol{u}))^2 + 2\right].\end{aligned} \tag{11}$$

The identity above uses the fact additive noises are unbiased and have bounded variances.

Note that from the Lipschitz Hessian condition, we have that

$$|f(\boldsymbol{x}\pm Z\boldsymbol{u}) - f_2(\boldsymbol{x}\pm Z\boldsymbol{u})| \le \frac{1}{6}\rho||Z\boldsymbol{u}||_2^3 \le \frac{1}{6}\rho\lambda_Z^3,$$

where f_2 is the Taylor polynomial of f expanded at $\boldsymbol{x}$ up to the quadratic terms. Consequently, inequality (11) implies

$$\begin{aligned}
\mathrm{Tr}\left(\mathrm{Cov}[\widehat{\boldsymbol{g}}]\right) &\leq \frac{d^2}{4n}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(S^{d-1})}\left[\left(|f_2(\boldsymbol{x}+Z\boldsymbol{u})-f_2(\boldsymbol{x}-Z\boldsymbol{u})|+\frac{1}{3}\rho\lambda_Z^3\right)^2+2\right] \\
&= \frac{d^2}{4n}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(S^{d-1})}\left[\left(|2Z\boldsymbol{u}\cdot\nabla f(\boldsymbol{x})|+\frac{1}{3}\rho\lambda_Z^3\right)^2+2\right] \\
&\leq \frac{d^2}{4n}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(S^{d-1})}\left[2\cdot|2Z\boldsymbol{u}\cdot\nabla f(\boldsymbol{x})|^2+2\left(\frac{1}{3}\rho\lambda_Z^3\right)^2+2\right] \\
&= \frac{2d}{n}||Z\nabla f(\boldsymbol{x})||_2^2+\frac{d^2}{18n}\left(\rho\lambda_Z^3\right)^2+\frac{d^2}{2n}.
\end{aligned}$$

□

5 Lower Bounds: Examples for the 1D Case

To illustrate the main proof idea, we first consider the 1D case. The proof for general d can be found in Appendix B. The gist of our proof is to construct a pair of hard-instance functions that needs to be distinguished in order to achieve low simple regret. While we also require them to be sufficiently close to each other so that they are indistinguishable without sufficiently many samples. These requirements are captured quantitatively in the following result, which is proved using an analysis of KL divergence. Here we assume their correctness and focus on the constructions.

Definition 1. *For any function class $\mathcal{F}_{\mathrm{H}}$ and any distribution p defined on $\mathcal{F}_{\mathrm{H}}$, we define the* uniform sampling error *to be*

$$P_\epsilon \triangleq \inf_{\boldsymbol{x}} \mathbb{P}_{f\sim p}[f(\boldsymbol{x})-\inf f\geq\epsilon].$$

We also define the maximum local variance *to be*

$$V \triangleq \sup_{\boldsymbol{x}} \mathrm{Var}_{f\sim p}[f(x)].$$

Lemma 1 (Restatement of Proposition 7 in Yu et al. (2022))**.** *For any sampling algorithm to achieve an expected simple regret of $\epsilon>0$ over a function class $\mathcal{F}_\epsilon$, if $P_{2\epsilon/c}\geq c$ for some universal constant $c\in(0,1)$, and the observation noises are standard Gaussian, then the required sample complexity to achieve a minimax regret of ϵ is at least $\Omega(1/V)$.*

We construct our hard instances using the following function

$$g(x)=\begin{cases}\frac{1}{2}\left(\sin\left(\frac{1}{2}x\right)+1\right) & \text{if } x\in(-\pi,3\pi]\\ -\cos x-1 & \text{if } x\in(-3\pi,-\pi]\\ 0 & \text{otherwise.}\end{cases}$$

Some key properties of $g(x)$ to be used are that its differential $g'(x)$ is 1-Lipschitz, and we have $|g'(x)|\leq 1$ for all x. Our hard instances consist of two functions. We define

$$f_1(x)=Mx^2+y_0\int_{-\pi}^{x/x_0}g(z)dz,\qquad f_2(x)=Mx^2+y_0\int_{-\pi}^{-x/x_0}g(z)dz,$$

where y_0, x_0 are normalization factors given by $y_0=\frac{1}{\pi\sqrt{T}}$, $x_0=\left(\frac{y_0}{\rho}\right)^{\frac{1}{3}}$. The normalization factors are chosen to satisfy the Lipschitz Hessian condition and a maximum local variance bound required for a KL-divergence based approach presented in Lemma 1.

Specifically, the choice of x_0 and the fact that $g'(x)$ is 1-Lipschitz imply that both f_1 and f_2 satisfy the Lipschitz Hessian condition. Then because the absolute value of integration of $g(x)$ is bounded by 2π, one can show that the maximum local variance for the function class $\{f_1,f_2\}$ is no greater

than $\pi^2 y_0^2 = \frac{1}{T}$ for the uniform prior distribution, which is to be used to show the sample complexity lower bound.

We first check that both f_1 and f_2 are within our function class of interests. Note that both $f_1''(x)$ and $f_2''(x)$ belong to the interval $[2M - \frac{5}{4}\frac{y_0}{x_0^2}, 2M - \frac{3}{4}\frac{y_0}{x_0^2}]$. From the fact that $\lim_{T\to\infty} \frac{y_0}{x_0^2} = 0$ and $M > 0$, we have both $f_1''(x) > M$ and $f_2''(x) > M$ for all x for sufficiently large T. So the strong convexity requirement is satisfied. On the other hand, consider any global minimum point x^* of either f_1 or f_2. Because of their differentiability, we must have $f_1'(x) = 0$ or $f_2'(x) = 0$. Note that $f_1'(x) = 2Mx + g\left(\frac{x}{x_0}\right)\frac{y_0}{x_0}$, $f_2'(x) = 2Mx - g\left(\frac{x}{x_0}\right)\frac{y_0}{x_0}$, and $|g(x)| \le 2$ for all x. We must have $|x^*| \le \frac{y_0}{x_0}/M$, where the RHS is $o(1)$ for large T. Combined with strong convexity, this inequality implies that assumption A3 holds for both functions. To conclude, we have proved that $f_1, f_2 \in \mathcal{F}(\rho, M, R)$ for sufficiently large T.

Now we let $\epsilon = \frac{1}{128M}\left(\frac{y_0}{x_0}\right)^2$ and $c = \frac{1}{2}$ to apply Lemma 1. Note that $\liminf_{T\to\infty} T^{\frac{2}{3}}\epsilon = \frac{\rho^{\frac{2}{3}}}{128\pi^{\frac{4}{3}}M}$. The quantity ϵ exactly matches the lower bounds we aim to prove. Therefore, it remains to check that the required condition on uniform sampling errors in Definition 1 are satisfied.

Formally, we need to show that $f_k(0) - \inf_x f_k(x) \ge 4\epsilon$ for $k \in \{1, 2\}$. Without loss of generality, we focus on the case of $k = 1$. Note that $f_1''(x) \le 2M + \frac{y_0}{4x_0^2}$ for all $x \in [-\pi x_0, 0]$. Therefore, we have $f_1(x) - f_1(0) \le f_1'(0)x + \frac{1}{2}x^2 \sup_{z\in[-\pi x_0,0]} f_1''(z) \le \frac{y_0}{2x_0}x + \frac{1}{2}x^2\left(2M + \frac{y_0}{4x_0^2}\right)$ for $x \in [-\pi x_0, 0]$. and $\lim_{T\to\infty} x_0 = 0$. Consider any sufficiently large T such that $\frac{y_0}{4x_0^2} \le 2M$, we can choose $x = -\frac{y_0}{2x_0}\frac{1}{2M + \frac{y_0}{4x_0^2}}$ for the above bound, which falls into the interval of $[-\pi x_0, 0]$. Then we have

$$\inf_x f_1(x) \le f_1\left(-\frac{y_0}{2x_0}\frac{1}{2M + \frac{y_0}{4x_0^2}}\right) \le f_1(0) - \frac{1}{2}\left(\frac{y_0}{2x_0}\right)^2 \frac{1}{2M + \frac{y_0}{4x_0^2}} \le f_1(0) - 4\epsilon.$$

We use this inequality to lower bound the minimum sampling error. Note that f_1 is an increasing function for $x \ge 0$ and $\inf_x f_1(x) = \inf_x f_2(x)$. We have $f_1(x) \ge \inf_x f_2(x) + 4\epsilon$ for $x \ge 0$. Following the same arguments, we also have $f_2(x) \ge \inf_x f_1(x) + 4\epsilon$ for $x \le 0$. Recall the definition of uniform sampling error in Definition 1. We have essentially proved that $P_{4\epsilon} \ge \frac{1}{2}$. According to earlier discussions, this implies that the minimax simple regret is lower bounded by $\epsilon = \Omega\left(\frac{\rho^{\frac{2}{3}}T^{-\frac{2}{3}}}{M}\right)$.

REFERENCES

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pp. 28–40. Citeseer, 2010.

Alekh Agarwal, Dean P. Foster, Daniel Hsu, Sham M. Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

Arya Akhavan, Massimiliano Pontil, and Alexandre Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *Advances in Neural Information Processing Systems*, 33:9017–9027, 2020.

Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pp. 257–283. PMLR, 2016.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Sébastien Bubeck, Ronen Eldan, and Yin Tat Lee. Kernel-based methods for bandit convex optimization. *Journal of the ACM (JACM)*, 68(4):1–35, 2021.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

Elad Hazan and Kfir Levy. Bandit convex optimization: Towards tight bounds. *Advances in Neural Information Processing Systems*, 27, 2014.

Tor Lattimore and Andras Gyorgy. Improved regret for zeroth-order stochastic convex bandits. In *Conference on Learning Theory*, pp. 2938–2964. PMLR, 2021.

Vasilii Novitskii and Alexander Gasnikov. Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit. *arXiv preprint arXiv:2101.03821*, 2021.

Luis Miguel Rios and Nikolaos V Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 3–24, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Shamir13.html.

Yining Wang, Sivaraman Balakrishnan, and Aarti Singh. Optimization of smooth functions with noisy observations: Local minimax rates. *IEEE Transactions on Information Theory*, 65(11): 7350–7366, 2019.

Qian Yu, Yining Wang, Baihe Huang, Qi lei, and Jason D. Lee. Optimal sample complexity bounds for convex and non-convex optimization: From kurdyka-lojasiewicz condition to quadratic bandits. *arxiv preprint*, 2022.

Yan Zhang, Yi Zhou, Kaiyi Ji, and Michael M Zavlanos. Boosting one-point derivative-free online optimization via residual feedback. *arXiv preprint arXiv:2010.07378*, 2020.