OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Hessian-Dependent Sample Complexity in Zeroth-Order Stochastic Optimization: Nonconvex Support Sampling Is Necessary for Optimality

Mengtian Hong 2960081H@STUDENT.GLA.AC.UK
University of Glasgow

Jason D. Lee JASONDLEE@BERKELEY.EDU
University of California, Berkeley

Qian Yu QIANYU02@UCSB.EDU
University of California, Santa Barbara

Abstract

Zeroth-order stochastic optimization is fundamental formulation that arises in real-world design problems where gradients are inaccessible. A central challenge in this pursuit is to design gradient estimators and optimization algorithms under noisy, function-only feedback that leverages the local Hessian-based geometry to achieve optimal sample efficiency. We introduce the Spectrally Grouped Estimator (SGE), a novel gradient-estimation scheme that samples over a non-convex set formed by a union of sphere sections, and utilize it to build an algorithm that achieves order-wise improved Hessian-dependent simple regrets over second-order smooth, strongly convex functions compared to ones based on conventional baseline methods which sample over convex sets. We complement these results with the first tight analyses of the baseline schemes, formally demonstrating their shared performance bottleneck and thus emphasizing the necessity of non-convex sampling for optimality. We conjecture that our achieved Hessian-dependent rates are universally-optimal.

1. Introduction

Stochastic zeroth-order optimization investigates the learning process when the only available resource is a noisy function value oracle. The goal is to develop an algorithm that interacts with an oracle, sequentially acquiring noisy function evaluations to identify a point that minimizes the simple regret. The study under this topic has largely been directed towards clarifying the behavior of minimax sample complexities under a variety of structural conditions. Foundational research has investigated the optimal rates for both general and strongly convex functions [9, 16], as well as the comprehension of optimal rates for functions characterized by Lipschitz continuity [3, 7, 12] or Lipschitz gradients [2, 5, 11, 15]. While optimal rates are well-studied for such function classes, recent attention has shifted to higher-order smoothness [2, 3, 5, 7, 11, 12].

This research is centered on the scenario within the second-order smoothness conditions and builds upon the latest developments in this area, where Yu et al. [17] have precisely defined the minimax simple regret for this class, establishing an optimal rate of $\mathcal{O}\big(dT^{-2/3}\big)$. We go beyond the worst-case minimax scenarios and target a significant objective on characterizing optimal instance-dependent rates that are dependent on the Hessian spectrum at the global optimizer.

The principal challenge in obtaining these instance-dependent rates is designing gradient estimator that fully exploiting the Hessian geometry. The conventional approach is to sample within a

single convex body (or on its boundary), a strategy that encompasses both isotropic sampling on a hypersphere [2, 4, 14, 16] and more adaptive, non-isotropic sampling on a hyperellipsoid [1, 17]. However, our analysis reveals that this entire class of methods shares a fundamental limitation: both spherical sampling and the generalized single-ellipsoid scheme have simple regrets that are invariably limited by the smallest eigenvalue of the Hessian, $H_{\min}$, reflecting a small-eigenvalue bottleneck in which a single hard direction limits progress in all dimensions. An alternative, the entry-wise estimator, decomposes the problem along the standard basis, but its prohibitive sample complexity renders it highly suboptimal [6, 8]. This common failure mode suggests rethinking the sampling geometry and going beyond single convex supports. Our results indicate that nonconvex constructions are a promising path.

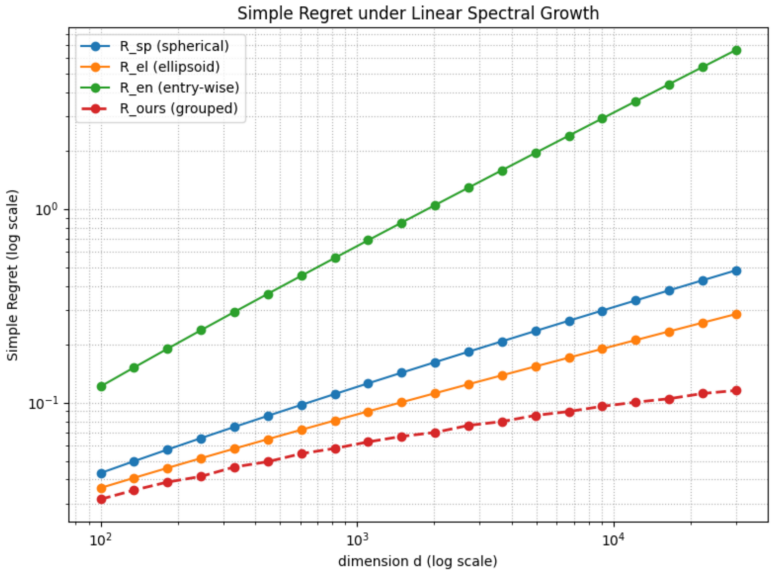

Figure 1: An illustration of the order-wise gain of the proposed scheme over baseline designs for a representative instance where the Hessian eigenvalues at the global minimizer are equally spaced ($H_i = iM$ for $i = 1, \ldots, d$). The simple regrets are normalized by a common factor $\rho^{2/3}T^{-2/3}M^{-1}$.

Summary of technical contributions. **(i)** We introduce the Spectrally Grouped Estimator, which partitions the Hessian spectrum into dyadic groups and samples over a nonconvex support formed as a union of spherical sections. Our main theorem establishes that this estimator achieves a strictly lower instance-dependent simple regret than all convex-support baselines. A detailed comparison of the regret bounds is provided in Table 1. Notably, for polynomial spectra, the estimator demonstrates orderwise improved dependence on the dimension d, as illustrated in Figure 1. **(ii)** We establish the tight worst-case regret bounds for three classes of baseline estimators, highlighting a universal constraint related to small eigenvalues, as determined by $H_{\min}$. **(iii)** In contrast to previous methodologies that employ convex sampling support, to the best of knowledge, our estimator is the first to intentionally utilize a non-convex geometry.

Notations. In accordance with the conventions of machine learning theory, we denote the Hessian of a function f at a point x as $\nabla^2 f(x)$. The parameters for smoothness and strong convexity are represented by ρ and M, respectively, as formally defined in our assumptions. We utilize $\|\cdot\|_2$ to signify the vector ℓ_2 norm and $\|\cdot\|_{\mathrm{F}}$ for the matrix Frobenius norm. We let $H \triangleq \nabla^2 f(x^*)$ represent the Hessian at the global minimizer x^*. Finally, we employ conventional notations (i.e., O, Ω) to describe regret bounds concerning the total number of samples, denoted by T.

Spherical	Ellipsoidal	Entry-wise	Ours (Proposed)
$O\left(\frac{\rho^{\frac{2}{3}} d^{\frac{1}{3}}}{T^{\frac{2}{3}} H_{\min}^{\frac{1}{3}}} \left(\sum_i H_i^{-1}\right)^{\frac{2}{3}}\right)$	$O\left(\frac{\rho^{\frac{2}{3}}}{T^{\frac{2}{3}} H_{\min}^{\frac{1}{3}}} \left(\sum_i H_i^{-\frac{2}{3}}\right)\right)$	$O\left(\frac{\rho^{\frac{2}{3}}}{T^{\frac{2}{3}}} \left(\sum_i H_i^{-\frac{3}{5}}\right)^{\frac{5}{3}}\right)$	$\mathbf{O}\left(\frac{\rho^{\frac{2}{3}}}{\mathbf{T}^{\frac{2}{3}}} \left(\sum_{\mathbf{i}} \mathbf{d_i}^{-\frac{2}{5}} \mathbf{H_i}^{-\frac{3}{5}}\right)^{\frac{5}{3}}\right)$

Table 1: Comparison of achieved Hessian-dependent simple regrets between the proposed scheme and baseline designs, where T is the total number of function evaluations, ρ is the smoothness parameter, d is the dimension, $H_{\min}$ is the smallest eigenvalue of the Hessian at the global minimizer, H_i are the corresponding Hessian eigenvalues, and d_i counts eigenvalues within a factor of 2 of each H_i. Our proposed method achieves strictly better simple regrets than all baseline schemes.

2. Problem Formulation

We consider an optimization setting where an algorithm $\mathcal{A}$ interacts with an unknown objective f over T iterations. In each round $t \in [T] := \{1, \ldots, T\}$, the algorithm selects $x_t \in \mathbb{R}^d$ and observes

$$y_t = f(x_t) + w_t,$$

where the noises $\{w_t\}_{t=1}^T$ are independent, mean zero, and with bounded variances, i.e., $\mathbb{E}[w_t \mid \{(x_\tau, y_\tau)\}_{\tau<t}, x_t] = 0$ and $\mathrm{Var}[w_t|\{(x_\tau, y_\tau)\}_{\tau<t}, x_t] \leq 1$. For simplicity, we additionally assume sub-Gaussianity of the noise. However, this assumption is needed only in the bootstrapping stage of our algorithm and can be removed with more elaborated designs [17].

We assume the algorithm $\mathcal{A}$ can be adaptive; formally, the choice of each x_t can be based on the entire history of prior observations $\{(x_\tau, y_\tau)\}_{\tau<t}$. Our analysis focuses on a broad class of twice-differentiable functions that exhibit well-behaved local geometry. Specifically, we assume the objective function f satisfies the following three standard conditions.

(A1) (Lipschitz Hessian). There exists a constant $\rho \in (0, +\infty)$ such that for all $x, x' \in \mathbb{R}^d$, the Hessian of f is ρ-Lipschitz continuous with respect to the Frobenius norm:

$$\|\nabla^2 f(x) - \nabla^2 f(x')\|_{\mathrm{F}} \leq \rho \|x' - x\|_2.$$

(A2) (Strong Convexity). There exists a constant $M \in (0, +\infty)$ such that for any $x \in \mathbb{R}^d$, the minimum eigenvalue of the Hessian $\nabla^2 f(x)$ is greater than M. Formally, $\nabla^2 f(x) \succeq MI$.

(A3) (Bounded Distance to Optimum). We assume the unique global minimizer x^* of the function f is located within a ball of a known radius $R > 0$ from the origin, i.e., $\|x^*\|_2 \leq R$.

Let H be a positive definite matrix. We define the function subclass $\mathcal{F}_H$ as the set of all functions f that satisfy $\nabla^2 f(x^*) = H$, where x^* is the global optimum of f. We are interested in the instance-dependent simple regret, characterized by the minimax simple regret over this subclass:

$$\mathcal{R}(T; H) := \inf_{\mathcal{A}} \sup_{f \in \mathcal{F}_H} \mathbb{E}[f(x_T) - f(x^*)]. \tag{1}$$

We also aim to design a universal algorithm that achieves this bound without knowledge of H.

3. Main Results

Our first contribution is to introduce a new gradient-free algorithm that achieves the following sample complexity guarantees.

Theorem 1 (Achievability of Spectrally Grouped Estimator) *There exists an algorithm $\mathcal{A}$ that does not depend on the parameter H and achieves the following simple regret guarantees over all classes $\mathcal{F}_H$*

$$\limsup_{T\to\infty} \sup_{f\in\mathcal{F}_H} \mathbb{E}[f(x_T) - f(x^*)] \cdot T^{\frac{2}{3}} \leq R(H) \triangleq C \cdot \rho^{\frac{2}{3}} \left(\sum_{k=0}^{\lceil \log_2(H_{\max}/M) \rceil} \left(\frac{d_k}{M \cdot 2^k} \right)^{\frac{3}{5}} \right)^{\frac{5}{3}},$$

where C is a universal constant, $H_{\max}$ is the largest eigenvalue of H, and d_k is the number of eigenvalues in the interval $[M \cdot 2^k, M \cdot 2^{k+1})$.

Central to the proof of Theorem 1 is the development of a new gradient estimator, termed the *spectrally grouped estimator*. The following result establishes a sharp bound on the performance of this estimator under a general class of bilinear cost functions that approximate simple regret, matching the rate and spectral dependence observed in Theorem 1.

Theorem 2 ($\hat{H}^{-1}$-Risk Characterization) *For any positive-definite matrix $\hat{H}$ and any point $\boldsymbol{x}_{\mathrm{B}}$, there exists a gradient estimation algorithm that uses n samples and returns a gradient estimate $\hat{\boldsymbol{g}}$ such that for all n greater than a threshold $n_{\mathrm{th}} \leq 2d^3$, the following holds uniformly over all function f that satisfy the Lipschitz Hessian condition*

$$\mathbb{E}\left[\left(\hat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}}) \right)^{\intercal} \hat{H}^{-1} (\hat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}})) \right]$$
$$\leq C \cdot \frac{\rho^{\frac{2}{3}}}{n^{\frac{2}{3}}} \left(\sum_{k=0}^{\lceil \log_2(\hat{H}_{\max}/\hat{H}_{\min}) \rceil} \left(\frac{\hat{d}_k}{\hat{H}_{\min} \cdot 2^k} \right)^{\frac{3}{5}} \right)^{\frac{5}{3}} \left(1 + ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2 \left(\frac{1}{n\rho^2} \right)^{\frac{1}{3}} \right),$$

where C is a universal constant, $\hat{H}_{\max}, \hat{H}_{\min}$ are the largest and smallest eigenvalues of $\hat{H}$, and $\hat{d}_k$ is the number of eigenvalues in the interval $[M \cdot 2^k, M \cdot 2^{k+1})$.

We conjecture that the rates in Theorem 1 are optimal. The formal statement is as follows.

Conjecture 3 (Optimality) *There exists a universal constant $c > 0$, such that for all H, the following inequality holds when the additive noise variables w_t are i.i.d. standard Gaussian.*

$$\liminf_{T\to\infty} \mathcal{R}(T; H) \cdot T^{\frac{2}{3}} \geq cR(H).$$

4. Proof of Main Results

4.1. Algorithm Overview

Our proposed algorithm follows a two-stage structure (see Algorithm 1). First, a bootstrapping function is called to return a preliminary estimate $\boldsymbol{x}_{\mathrm{B}}$ of the global minimizer. Then, a refinement

stage performs a gradient-clipped Newton step based on the gradient and Hessian estimators $\hat{\boldsymbol{g}}$ and $\hat{H}$. The simple regret achieved by this algorithm depends on the qualities of these individual steps, which is characterized in the following proposition and proved in Appendix C.

Algorithm 1: Achievability Framework for Theorem 1

Input: T, ρ, M

Let $n_{\mathrm{B}} = n_{\mathrm{H}} = n_{\mathrm{G}} = \lfloor (T-1)/3 \rfloor$; `// constant fraction of samples to each routine`

Let $\boldsymbol{x}_{\mathrm{B}} = \text{Bootstrapping}\,(n_{\mathrm{B}}, \rho, M)$; `// Bootstrapping stage`

Let $\hat{H} = \text{HessianEst}(\boldsymbol{x}_{\mathrm{B}}, n_{\mathrm{H}})$, $\hat{\boldsymbol{g}} = \text{GroupedEst}(\boldsymbol{x}_{\mathrm{B}}, n_{\mathrm{G}})$; `// Hessian and Gradient Estimation`

Let $\boldsymbol{r} = -\hat{H}^{-1}\hat{\boldsymbol{g}} / \max\left\{1, \rho M^{-\frac{3}{2}} \cdot \sqrt{\hat{\boldsymbol{g}}^{\intercal}\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{g}}}\right\}$; `// Clipped Newton`

return $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{B}} + \boldsymbol{r}$;

Proposition 4 *Consider any function f that satisfies strong convexity and Lipschitz Hessian conditions. If $\boldsymbol{x}$ is generated by the final stage of Algorithm 1 conditioned on any fixed $\boldsymbol{x}_{\mathrm{B}}$ and any symmetric $\hat{H}$ with all eigenvalues lower bounded by M, then*

$$f(\boldsymbol{x}) - f^* \le (\hat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}}))^{\intercal} \hat{H}^{-1} (\hat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}})) (1 + \epsilon_H/M) + 2\frac{\epsilon_H ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{M^2} + (6 + \epsilon_H/M)\frac{\rho ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3}. \tag{2}$$

where $\hat{\boldsymbol{x}} \triangleq \boldsymbol{x}_{\mathrm{B}} - \hat{H}^{-1}\nabla f(\boldsymbol{x}_{\mathrm{B}})$, $\epsilon_H \triangleq ||\hat{H} - \nabla^2 f(\boldsymbol{x}_{\mathrm{B}})||_{\mathrm{F}}$, and $\hat{f}(\boldsymbol{x})$ denote the quadratic approximation $\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}})^{\intercal}\hat{H}(\boldsymbol{x} - \hat{\boldsymbol{x}})$.

Essentially, to minimize the simple regret, we need carefully designed routines to minimize the gradient estimation error (measured in a bilinear form characterized by $\hat{H}^{-1}$), Hessian estimation error (measured in Frobenius norm), and the gradient norm at the end of the bootstrapping stage.

4.2. Proof of Theorem 2

To minimize the gradient estimation error, we proposed an algorithm that samples the objective function on a union of spherical intersections (see Algorithm 2). Specifically, we partition the coordinates under the eigenbasis of $\hat{H}$ into groups according to their associated eigenvalues, and within each group we jointly estimate the corresponding gradient components by isotropic sampling on a spherical section.

By decoupling the measurements across groups, the algorithm avoids the dependence on the smallest eigenvalue $\hat{H}_{\min}$ in the global factor of the achieved $\hat{H}^{-1}$ error. At the same time, the joint spherical estimation within each group yields a better bias-variance tradeoff than fully entrywise methods. We conjecture that our spectrally grouped estimator attains the order-optimal $\hat{H}^{-1}$-norm error in the regime of $\|\nabla f(\boldsymbol{x}_{\mathrm{B}})\| \to 0$ and $n \to \infty$. We refer readers to Appendix E for further details.

Algorithm 2: GroupedEst

Input: $\boldsymbol{x}_B$, $\hat{H}$, n

// $\boldsymbol{x}_B$: query point; n: number of samples

Let $\hat{H}_1, ..., \hat{H}_d$ be the eigenvalues of $\hat{H}$, and let $\boldsymbol{e}_1, ..., \boldsymbol{e}_d$ be an associated eigenbasis;

Let $\hat{H}_{\min} = \min_j \hat{H}_j$ and $\hat{H}_{\max} = \max_j \hat{H}_j$. Partition indices into groups G_i:

$$G_i = \{\, k \mid \hat{H}_k \in [\hat{H}_{\min} 2^{i-1},\, \hat{H}_{\min} 2^i)\,\}, \quad i = 1, \ldots, \lceil \log_2(\hat{H}_{\max}/\hat{H}_{\min}) \rceil;$$

// Optimal allocation of the samples

Set $d_i = |G_i|$, $c_i = d_i / \left(\hat{H}_{\min} 2^i\right)$, and $n_i = \left\lceil (n - 2d) \cdot \frac{c_i^{3/5}}{2 \sum_j c_j^{3/5}} \right\rceil$;

foreach $G_i \neq \varnothing$ **do**

  Compute the estimate $\hat{g}_i$ of the gradient component within the subspace spanned by $\{\boldsymbol{e}_k\}_{k \in G_i}$ using the spherical estimator with $2n_i$ samples and radius $r_i = d_i^{\frac{1}{2}} n_i^{-\frac{1}{6}} \rho^{-\frac{1}{3}}$;

  // see Algorithm 3 in Appendix A

end

return $\hat{g} \leftarrow \sum_{i=1}^{I} \hat{g}_i$; // Return $\hat{g}$ as an estimator of $\nabla f(x)$

4.3. Proof of Theorem 1

According to the earlier analysis, the gradient estimation guarantee provided by Theorem 2 implies the following achievability result for Algorithm 1 when T is sufficiently large.

$$\mathbb{E}[f(\boldsymbol{x}) - f^*] \leq \mathbb{E}\left[C \cdot \frac{\rho^{\frac{2}{3}}}{T^{\frac{2}{3}}} \left(\sum_{k=0}^{\lceil \log_2(\hat{H}_{\max}/\hat{H}_{\min}) \rceil} \left(\frac{\hat{d}_k}{\hat{H}_{\min} \cdot 2^k} \right)^{\frac{3}{5}} \right)^{\frac{5}{3}} \left(1 + \frac{||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{(n\rho^2)^{\frac{1}{3}}} \right) (1 + \epsilon_H/M) \right.$$
$$\left. + 2\frac{\epsilon_H ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{M^2} + (6 + \epsilon_H/M) \frac{\rho ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3} \right]. \tag{3}$$

Note that in the regime of ϵ_H and $||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2 \to 0$, the bound above matches the rate in Theorem 1 up to the $\hat{H}$-dependent factor. Since the corresponding factors coincide when $\hat{H} = H$ and $H_{\min} = M$, a crucial step in our proof is to show that this factor is stable under our choice of the Hessian estimator (even though their constituent terms are not), and that the extra $(1 + \epsilon_H/M)$ multiplier does not induce higher-order dependence. We present the proof details in Appendix D, as well as the convergence guarantees for ϵ_H and $||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2$.

5. Conclusion and Future Work

In this work, we established a new gradient estimator for stochastic zeroth-order optimization, named Spectrally Grouped Estimator, and utilized it to achieve orderwise improvements on Hessian-dependent rates compared to conventional convex-support sampled methods. While our focus was on tight characterization of asymptotic rates, improved characterization for the non-asymptotic regime (e.g., through an improved analysis of n_{th}) remains future work. Finally, we recently resolved the conjecture on the optimality of SGE, which will be presented in the extended version.

References

[1] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, number 110, 2009.

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

[3] Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems*, 24, 2011.

[4] 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.

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

[6] Julius R Blum. Multidimensional stochastic approximation methods. *The annals of mathematical statistics*, pages 737–744, 1954.

[7] 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.

[8] Liyi Dai. Convergence rates of finite difference stochastic approximation algorithms part ii: implementation via common random numbers. In *Sensing and Analysis Technologies for Biomedical and Cognitive Applications 2016*, volume 9871, pages 138–156. SPIE, 2016.

[9] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[10] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

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

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

[13] Abdelkader Mokkadem and Mariane Pelletier. A companion for the kiefer–wolfowitz–blum stochastic approximation algorithm. 2007.

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

[15] Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on learning theory*, pages 3–24. PMLR, 2013.

[16] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

[17] Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason D Lee. Stochastic zeroth-order optimization under strongly convexity and lipschitz hessian: minimax sample complexity. *Advances in Neural Information Processing Systems*, 37:99564–99600, 2024.

Appendix A. Analysis for Baseline Gradient Estimators

In this appendix, we provide a detailed analysis of the three baseline zeroth-order gradient estimators. For each estimator, we present its algorithm block and a theorem stating its performance in terms of the $\hat{H}^{-1}$-Risk, which leads to tight simple-regret guarantees following the steps in Section 4.3. This analysis serves to precisely characterize its limitations and formally establish the shared performance bottleneck that motivates our main algorithm.

A.1. Spherical Sampling

The spherical gradient estimator is a foundational and widely-used technique in zeroth-order optimization [4, 10], relying on isotropic sampling along random directions drawn from the unit hypersphere.

Algorithm 3: SphericalGradEst [16]

`// Spherical two-point gradient estimator;`
Input: x, r, n

`// x: query point; r: sampling radius`
`// n: number of samples`
`// Return` $\hat{g}$ `as an estimator of` $\nabla f(x)$

for $k \leftarrow 1$ **to** $\lfloor n/2 \rfloor$ **do**
  A random direction u_k is drawn uniformly from the unit hypersphere $\mathbb{S}^{d-1}$;
  The oracle is queried for noisy values y_+ and y_- at the points $x + ru_k$ and $x - ru_k$, respectively;
  A single two-point estimate is formed as $g_k = \dfrac{d}{2r}(y_+ - y_-)u_k$;
end
return $\hat{g} = \frac{1}{n}\sum_{k=1}^{n} g_k$;

To characterize its performance in the $\hat{H}^{-1}$ norm, we first prove the following preliminary result.

Proposition 5 *For any fixed x, r, n, the estimator $\hat{g}$ returned by SphericalGradEst satisfies the following inequality for any unit vector $\boldsymbol{e}$.*

$$\mathrm{Var}\Big[\hat{g} \cdot \boldsymbol{e}\Big] \leq \frac{6}{\lfloor n/2 \rfloor}||\nabla f(x)||_2^2 + \frac{d}{18\lfloor n/2 \rfloor}\left(\rho r^2\right)^2 + \frac{d}{2\lfloor n/2 \rfloor r^2}. \tag{4}$$

Proof We first consider each g_k's contribution to the variance. For clarity, let w_+, w_- denote two independent samples of additive noises.

$$\begin{aligned}\text{Var}\Big[g_k \cdot \boldsymbol{e}\Big] &= \mathbb{E}_{\boldsymbol{u}\sim\text{Unif}(S^{d-1}),w_+,w_-}\left[\left(\frac{d}{2r}\right)^2 \left(f(\boldsymbol{x}+r\boldsymbol{u}) - f(\boldsymbol{x}-r\boldsymbol{u}) + w_- - w_-\right)^2 (\boldsymbol{u}\cdot\boldsymbol{e})^2\right] \\ &\leq \frac{d}{4r^2}\mathbb{E}_{\boldsymbol{u}\sim\text{Unif}(S^{d-1})}\left[d\left(f(\boldsymbol{x}+r\boldsymbol{u}) - f(\boldsymbol{x}-r\boldsymbol{u})\right)^2 (\boldsymbol{u}\cdot\boldsymbol{e})^2 + 2\right].\end{aligned}$$

Note that by Lipschitz Hessian condition, $|f(\boldsymbol{x}+r\boldsymbol{u}) - f(\boldsymbol{x}-r\boldsymbol{u})| \leq |2r\nabla f(x)\cdot\boldsymbol{u}| + \frac{1}{3}\rho r^3$. The first term above can be bounded by the AM-GM inequality.

$$\begin{aligned}\text{Var}\Big[g_k \cdot \boldsymbol{e}\Big] &\leq \frac{d^2}{2}\mathbb{E}_{\boldsymbol{u}\sim\text{Unif}(S^{d-1})}\left[\left(4\left(\nabla f(x)\cdot\boldsymbol{u}\right)^2 + \left(\frac{1}{3}\rho r^2\right)^2\right)(\boldsymbol{u}\cdot\boldsymbol{e})^2\right] + \frac{d}{2r^2} \\ &\leq 6||\nabla f(x)||_2^2 + \frac{d}{18}\left(\rho r^2\right)^2 + \frac{d}{2r^2}.\end{aligned}$$

Then, the proposition is proved by incorporating the contributions from all iterations. ■

Given the above proposition, an upper bound on the variance term in the $\hat{H}^{-1}$ norm of the gradient estimation error can be written as a weighted sum of the bounds in inequality (4). On the other hand, the contribution due to the bias of the estimation can be characterized as follows.

$$\begin{aligned}\mathbb{E}\left[\hat{g} - \nabla f(x)\right]^\top \hat{H}^{-1}\mathbb{E}\left[\hat{g} - \nabla f(x)\right] &\leq \hat{H}_{\min}^{-1}\left|\left|\mathbb{E}\left[\hat{g} - \nabla f(x)\right]\right|\right|_2^2 \\ &\leq \frac{r^4\rho^2 d}{4\hat{H}_{\min}(d+2)^2},\end{aligned}$$

where the second inequality is due to the sharp characterization of the estimation bias in [17]. By choosing the optimal sampling radius that balances the two components, we obtain the following characterization.

Theorem 6 ($\hat{H}^{-1}$-Risk Characterization for the Spherical Estimator) *For any positive-definite matrix $\hat{H}$ and any point x, the SphericalGradEst algorithm (Algorithm 3) uses n samples and returns a gradient estimate $\hat{g}$ such that for all sufficiently large n, there exists a choice of r such that the following holds uniformly over all functions f satisfying the Lipschitz Hessian condition*

$$\begin{aligned}\mathbb{E}\Big[(\hat{g} - \nabla f(x))^\top \hat{H}^{-1}(\hat{g} - \nabla f(x))\Big] \leq C\cdot\frac{\rho^{2/3}d^{1/3}}{n^{2/3}}\cdot\frac{1}{\hat{H}_{\min}^{1/3}}\left(\sum_{i=1}^d \hat{H}_i^{-1}\right)^{2/3} \\ \cdot\left(1 + \|\nabla f(x)\|_2^2\left(\frac{1}{n\rho^2}\right)^{1/3}\right),\end{aligned}$$

where C is a universal constant, and $\hat{H}_i$ and $\hat{H}_{\min}$ are the i-th and smallest eigenvalues of $\hat{H}$, respectively.

A.2. Ellipsoidal Sampling

The single-hyperellipsoid estimator represents a significant advancement over isotropic methods by attempting to adapt to the problem instance's geometry [1, 17]. This approach, which forms the basis of the final stage in our prior work to achieve the minimax sample complexities, uses a shaping matrix to create a non-isotropic sampling distribution.

Algorithm 4: EllipsoidGradEst [17]

```
// Non-isotropic gradient estimator
```
Input: x, Z, n

```
// x:  query point; Z:  shaping matrix ∈ R^{d×d}
// n:  number of samples
// The resulting ĝ provides a stochastic estimate of Z^T ∇f(x)
```
for $k \leftarrow 1$ **to** $\lfloor n/2 \rfloor$ **do**

A random direction u_k is drawn uniformly from the unit hypersphere $\mathbb{S}^{d-1}$;

The oracle is queried for noisy values y_+ and y_- at the transformed points $x + Zu_k$ and $x - Zu_k$, respectively;

A single-point estimate is formed as follows, let $g_k = \dfrac{d}{2}(y_+ - y_-)u_k$;

end

return $\hat{g} = \frac{1}{n}\sum_{k=1}^{n} g_k$;

While the achievability guarantee of the Ellipsoidal Estimator is not directly provided by our analysis in Appendix A.1, one can characterize their variance through a similar analysis in a linearly transformed domain where the Lipschitz Hessian Condition is preserved.

Proposition 7 *For any fixed x, Z, n, the estimator $\hat{g}$ returned by EllipsoidGradEst satisfies the following inequality for any unit vector $\boldsymbol{e}$.*

$$\mathrm{Var}\left[\hat{g} \cdot \boldsymbol{e}\right] \leq \frac{6}{\lfloor n/2 \rfloor}||Z^\intercal \nabla f(x)||_2^2 + \frac{d}{18\lfloor n/2 \rfloor}\left(||Z||^2\rho\right)^2 + \frac{d}{2\lfloor n/2 \rfloor}. \tag{5}$$

where $||Z||$ is the largest singular value of Z.

Proof Let $g(y) := f(x + Zy)$, we have $\nabla g(y) = Z^\intercal \nabla f(x + Zy)$. Note that g is $||Z||^2\rho$-Lipschitz Hessian. The proposition is proved by applying the spherical variance bound (Proposition 5) to g. ■

On the other hand, a tight characterization on the bias of the Ellipsoidal Estimator requires a more careful adaptation of the proof methods introduced in [17]. Specifically, as shown in [17], for any given matrix Z and parameter $s \in [0, 1]$, let $G(s, x)$ denote the expectation of the EllipsoidGradEst estimator with the first two inputs given by x and sZ, we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{\boldsymbol{G}(s;\boldsymbol{x})}{s} = \mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(B^d)}\left[\boldsymbol{u}\,\mathrm{Tr}\left(ZZ^\intercal\nabla^2 f(\boldsymbol{x} + sZ\boldsymbol{u})\right)\right], \tag{6}$$

where $\mathrm{Unif}(B^d)$ is the unform distribution over the standard hyperball. Then, if we apply the mirror-symmetric cancellation w.r.t. any unit vector $\boldsymbol{e}$, we instead have

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{\boldsymbol{G}(s;\boldsymbol{x})}{s}\cdot\boldsymbol{e} = \frac{1}{2}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(B^d)}\left[(\boldsymbol{u}\cdot\boldsymbol{e})\,\mathrm{Tr}\left(ZZ^\intercal\nabla^2\left(f(\boldsymbol{x} + sZ\boldsymbol{u}) - f(\boldsymbol{x} + sZ\boldsymbol{u}_{\boldsymbol{e}})\right)\right)\right], \tag{7}$$

where $\boldsymbol{u_e}$ denotes the reflection of $\boldsymbol{u}$ with respect to the hyperplane orthogonal to $\boldsymbol{e}$. By Lipschitz Hessian, we have that

$$\left|\mathrm{Tr}\left(ZZ^\intercal\nabla^2\left(f(\boldsymbol{x}+sZ\boldsymbol{u})-f(\boldsymbol{x}+sZ\boldsymbol{u_e})\right)\right)\right| \le 2\rho||ZZ^\intercal||_{\mathrm{F}}\cdot|\boldsymbol{u}\cdot\boldsymbol{e}|\cdot||sZ\boldsymbol{e}||_2. \tag{8}$$

Therefore, for any non-singular matrix Z, we conclude that

$$\begin{aligned}
\left|\left|\frac{\mathrm{d}}{\mathrm{d}s}\frac{(Z^\intercal)^{-1}\boldsymbol{G}(s;\boldsymbol{x})}{s}\right|\right|_2 &= \sup_{\boldsymbol{v}:||\boldsymbol{v}||_2=1}\left|\frac{\mathrm{d}}{\mathrm{d}s}\frac{(Z^\intercal)^{-1}\boldsymbol{G}(s;\boldsymbol{x})}{s}\cdot\boldsymbol{v}\right| \\
&= \sup_{\boldsymbol{e}:||\boldsymbol{e}||_2=1}\left|\frac{\mathrm{d}}{\mathrm{d}s}\frac{\boldsymbol{G}(s;\boldsymbol{x})}{s}\cdot\frac{\boldsymbol{e}}{||Z\boldsymbol{e}||_2}\right| \\
&\le \sup_{\boldsymbol{e}:||\boldsymbol{e}||_2=1}\mathbb{E}_{\boldsymbol{u}\sim\mathrm{Unif}(B^d)}\left[\rho||ZZ^\intercal||_{\mathrm{F}}\cdot|\boldsymbol{u}\cdot\boldsymbol{e}|^2\cdot|s|\right] \\
&= \frac{\rho|s|}{d+2}||ZZ^\intercal||_{\mathrm{F}}.
\end{aligned} \tag{9}$$

Consequently, the contribution due to the bias of the ellipsoidal estimator can be characterized as follows.

$$\begin{aligned}
\mathbb{E}\left[(Z^\intercal)^{-1}\hat{g}-\nabla f(x)\right]^\top\hat{H}^{-1}\mathbb{E}\left[(Z^\intercal)^{-1}\hat{g}-\nabla f(x)\right] &\le \hat{H}_{\min}^{-1}\left|\left|\mathbb{E}\left[(Z^\intercal)^{-1}\hat{g}-\nabla f(x)\right]\right|\right|_2^2 \\
&\le \frac{||ZZ^\intercal||_{\mathrm{F}}^2\rho^2}{4\hat{H}_{\min}(d+2)^2},
\end{aligned}$$

By selecting a carefully designed Z matrix to balance the bias-variance tradeoff, we present the following characterization.

Theorem 8 ($\hat{H}^{-1}$-Risk Characterization for the Ellipsoidal Estimator) *For any positive-definite matrix $\hat{H}$ and any point x, there exists a choice of shaping matrix Z for the EllipsoidalGradEst algorithm (Algorithm 4) that uses n samples and returns a gradient estimate $\hat{g}$ such that for all sufficiently large n, the following holds uniformly over all functions f that satisfy the Lipschitz Hessian condition:*

$$\begin{aligned}
\mathbb{E}\Big[((Z^\intercal)^{-1}\hat{g}-\nabla f(x))^\top\hat{H}^{-1}((Z^\intercal)^{-1}\hat{g}-\nabla f(x))\Big] \le C\cdot\frac{\rho^{2/3}}{n^{2/3}}\cdot\frac{1}{\hat{H}_{\min}^{1/3}}\left(\sum_{i=1}^d\hat{H}_i^{-2/3}\right) \\
\cdot\left(1+\|\nabla f(x)\|_2^2\left(\frac{1}{n\rho^2}\right)^{1/3}\right),
\end{aligned}$$

where C is a universal constant, and $\hat{H}_i$ and $\hat{H}_{\min}$ are the i-th and smallest eigenvalues of $\hat{H}$, respectively.

A.3. Entry-wise Sampling

The entry-wise (or coordinate-wise) estimator presents a clear and effective strategy for gradient estimation [6, 13]. This approach simplifies the task by breaking it down into a sequence of one-dimensional finite-difference computations along a choice of standard basis vectors.

Algorithm 5: EntrywiseGradEst [6, 8]

```
// Coordinate-wise finite-difference gradient estimator
```
Input: $x, \{r_k, e_k, n_k\}_{k\in\{1,\dots,d\}}$

```
// x: query point; r_k: step sizes for coordinate k
// n_k: number of samples for coordinate k
// {e_k}_{k∈{1,...,d}}: an orthonormal basis of R^d
// Return ĝ as an estimator of ∇f(x)
```

for $k \leftarrow 1$ **to** d **do**

The oracle is queried for the empirical averages of noisy values $y_{+,k}$ and $y_{-,k}$ at the points $x + r_k e_k$ and $x - r_k e_k$, respectively, using $\lfloor n_k/2 \rfloor$ evaluations at each point;

A single-coordinate estimate is formed as follows, $\hat{g}_k = \dfrac{y_{+,k} - y_{-,k}}{2r_k}$;

end

return $\hat{g} = \sum_{k=1}^{d} \hat{g}_k e_k$

Proposition 9 (Entrywise Estimation Error of EntrywiseGradEst) *For any fixed x, n_k, and basis direction e_k, the entrywise estimator $\hat{g}_k$ obtained in EntrywiseGradEst satisfies the following inequality:*

$$\mathbb{E}\big[(\hat{g}_k - e_k \cdot \nabla f(x))^2\big] \;\le\; \frac{\rho^2 r_k^4}{36} \;+\; \frac{1}{2\lfloor n_k/2 \rfloor r_k^2}.$$

Proof We start the variance characterization. Recall that in our earlier analysis for SphericalGradEst (Proposition 5), the variance bound contains three terms, a direction-randomness term, a direction-nonlinearity interaction term, and a pure noise term. Since for EntrywiseGradEst, the estimation direction is fixed to the coordinate basis vector e_k, and therefore the first two terms vanish. The remaining variance comes solely from averaging n_k independent noisy oracle evaluations. Hence,

$$\mathrm{Var}[\hat{g}_k] \le \frac{1}{2\lfloor n_k/2 \rfloor r_k^2}.$$

On the other hand, the bias characterization for each coordinate-wise estimation is simply obtained by specializing the spherical-estimation analysis to the 1D case.

$$||\mathbb{E}[\hat{g}_k - e_k \cdot \nabla f(x)]||_2 \;\le\; \frac{\rho r_k^2}{6}.$$

Then, the proposition is proved by combining the two components. ■

We present the following theorem, which is obtained by choosing the optimal allocation of samples for each direction.

Theorem 10 ($\hat{H}^{-1}$-Risk Characterization for the Entry-wise Estimator) *For any positive-definite matrix $\hat{H}$ and any point x, there exists a choice of input parameters for the EntrywiseGradEst algorithm (Algorithm 5) that uses n samples and returns a gradient estimate $\hat{g}$ such that for all sufficiently large n, the following holds uniformly over all functions f that satisfy the Lipschitz Hessian*

condition:

$$\mathbb{E}\Big[(\hat{g}-\nabla f(x))^\top \hat{H}^{-1}(\hat{g}-\nabla f(x))\Big] \le C\cdot\frac{\rho^{2/3}}{n^{2/3}}\left(\sum_{i=1}^{d}\hat{H}_i^{-3/5}\right)^{5/3}, \tag{10}$$

where C is a universal constant, and $\hat{H}_i$ is the i-th eigenvalue of $\hat{H}$.

Appendix B. Hessian Estimator

We employ a standard algorithm, presented in [17] and included as in Algorithm 6, which performs an entrywise estimation of the Hessian, followed by a projection operation. The following lemma provides a guarantee on its accuracy under the Lipschitz Hessian condition.

Algorithm 6: HessianEst [17]

Input: x, r, n

`// A coordinate-wise estimator for` $\nabla^2 f(x)$ `with sample complexity scaling as` $O(nd^2)$

An orthonormal basis $\{e_1,\dots,e_d\}$ for $\mathbb{R}^d$ is selected;
The value of the function y is initially estimated by averaging over n samples of f;
for $k \leftarrow 1$ **to** d **do**
  The averaged values $y_{+,k}$ and $y_{-,k}$ are obtained by querying the oracle n times at $x+re_k$ and $x-re_k$, respectively;
  The k-th diagonal entry of the Hessian is then estimated as $H_{kk}=(y_{+,k}+y_{-,k}-2y)/r^2$;
  for $l \leftarrow k+1$ **to** d **do**
    Off-diagonal entries H_{kl} and H_{lk} are estimated by averaging n samples of the quantity: $\big(f(x+re_k+re_l)+f(x-re_k-re_l)-f(x+re_k-re_l)-f(x-re_k+re_l)\big)/(4r^2)$
  end
end
Let $\hat{H}_0=\{H_{jk}\}_{(j,k)\in[d]^2}$ be the initial estimate;
$\hat{H}$ is formed by preserving the eigenvectors of $\hat{H}_0$ while lifting each eigenvalue λ to $\max\{\lambda, M\}$. `// projected via its eigenspectrum`
return $\hat{H}$

Lemma 11 *Let f be a function satisfying the Lipschitz Hessian (A1) conditions with parameterρ. Let $\hat{H}$ be the output of the* `HessianEst` *subroutine (Algorithm 6) with inputs x, r, n. Then, for any sufficiently large n, the estimation error in the Frobenius norm is bounded as follows:*

$$\mathbb{E}\left[\|\hat{H}-\nabla^2 f(x)\|_F^2\right] \le C\left(d^2\rho^2 r^2 + d^2/(nr^4)\right). \tag{11}$$

where $C>0$ is a universal constant.

Appendix C. Proof of Proposition 4

Proof The first step of our proof is to show the following inequality, which holds universally for all $\boldsymbol{x}$ with $||\boldsymbol{x}-\boldsymbol{x}_{\mathrm{B}}||_2 \le \frac{M}{\rho}$.

$$f(\boldsymbol{x})-f^* \le 2\hat{f}(\boldsymbol{x})\left(1+\epsilon_H/M\right)+2\frac{\epsilon_H||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{M^2}+\frac{3\sqrt{2}\rho||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3}. \tag{12}$$

For brevity, let $\tilde{\boldsymbol{x}} \triangleq \boldsymbol{x}_{\mathrm{B}} - (\nabla^2 f(\boldsymbol{x}_{\mathrm{B}}))^{-1}\nabla f(\boldsymbol{x}_{\mathrm{B}})$, and $\tilde{f}(\boldsymbol{x})$ denote the corresponding quadratic approximation $\frac{1}{2}(\boldsymbol{x}-\tilde{\boldsymbol{x}})^{\intercal}\nabla^2 f(\boldsymbol{x}_{\mathrm{B}})(\boldsymbol{x}-\tilde{\boldsymbol{x}})$. We leverage Proposition 4.5 in [17] and obtain that

$$f(\boldsymbol{x}) - f^* \leq 2\tilde{f}(\boldsymbol{x}) + \frac{12\rho(f(\boldsymbol{x}_{\mathrm{B}}) - f^*)^{\frac{3}{2}}}{M^{\frac{3}{2}}}. \tag{13}$$

By applying the PL condition, which is implied by strong convexity, the second term in the above inequality is upper bounded by $\frac{3\sqrt{2}\rho||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3}$. Therefore, it remains to upper bound $\tilde{f}(\boldsymbol{x})$.

Recall the definitions of $\tilde{f}$ and $\hat{f}$. We have the following identity.

$$\begin{aligned}2(\hat{f}(\boldsymbol{x}) - \tilde{f}(\boldsymbol{x})) = &(\boldsymbol{x}-\hat{\boldsymbol{x}})^{\intercal}(\hat{H} - \nabla^2 f(\boldsymbol{x}_{\mathrm{B}}))(\boldsymbol{x}-\hat{\boldsymbol{x}}) \\ &+ \nabla f(\boldsymbol{x}_{\mathrm{B}})^{\intercal}(\nabla^2 f(\boldsymbol{x}_{\mathrm{B}}))^{-1}(\hat{H} - \nabla^2 f(\boldsymbol{x}_{\mathrm{B}}))(2(\boldsymbol{x}-\hat{\boldsymbol{x}}) + (\hat{\boldsymbol{x}}-\tilde{\boldsymbol{x}})).\end{aligned}$$

Given that the eigenvalues of $\hat{H}$ and $\nabla^2 f(\boldsymbol{x}_{\mathrm{B}})$ are lower bounded by M, we have $||\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}||_2 \leq ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2/M$. Therefore, by AM-GM inequality, the above equality implies

$$\tilde{f}(\boldsymbol{x})) \leq \hat{f}(\boldsymbol{x}) + \epsilon_H \left(\hat{f}(\boldsymbol{x})/M + ||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2/M^2\right),$$

which implies inequality (12).

Note that when all eigenvalues of $\hat{H}$ are lower bounded by M, the vector $\boldsymbol{x}$ returned by the final stage of our algorithm always satisfies $||\boldsymbol{x} - \boldsymbol{x}_{\mathrm{B}}||_2 \leq \frac{M}{\rho}$. Thus, inequality (12) can be applied. The proposition is proved through the following bound, which is by elementary geometry,

$$\hat{f}(\boldsymbol{x}) \leq \frac{1}{2}(\hat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}}))^{\intercal}\hat{H}^{-1}(\hat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}})) + \mathbb{1}\left(\hat{f}(\boldsymbol{x}_{\mathrm{B}}) \geq \frac{M^3}{2\rho^2}\right)\hat{f}(\boldsymbol{x}_{\mathrm{B}}), \tag{14}$$

and the second term on the RHS is further bounded by $\frac{\rho||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{2M^3}$. ■

Appendix D. Proof Details of Theorem 1

As noted in Section 4.3, we first characterize the stability of the $\hat{H}$-dependent factor in inequality (3). We summarize the first desired property in the following proposition.

Proposition 12 *There exists a universal constant C such that the following inequality hold for any symmetric $\hat{H}$ with all eigenvalues greater than or equal to M.*

$$\left(\sum_{k=0}^{\lceil \log_2(\hat{H}_{\max}/\hat{H}_{\min})\rceil}\left(\frac{\hat{d}_k}{\hat{H}_{\min}\cdot 2^k}\right)^{\frac{3}{5}}\right)^{\frac{5}{3}} \leq CR(\nabla^2 f(\boldsymbol{x}_B))\left(1+\frac{\epsilon_H}{M}\right). \tag{15}$$

Proof We first show that the LHS of inequality (15) is within a universal constant factor of $R(\hat{H})$. Let d_k denote the number of eigenvalues of $\hat{H}$ within $[M\cdot 2^k, M\cdot 2^{k+1})$, and k_0 be the integer that

satisfies $\hat{H}_{\min} \in [M \cdot 2^{k_0}, M \cdot 2^{k_0+1})$, we have

$$\begin{aligned}\sum_k \left(\frac{\hat{d}_k}{\hat{H}_{\min} \cdot 2^k}\right)^{\frac{3}{5}} &\leq \sum_k \left(\frac{d_{k+k_0} + d_{k+k_0+1}}{M \cdot 2^{k+k_0}}\right)^{\frac{3}{5}} \\ &\leq \sum_k \left(\frac{d_{k+k_0}}{M \cdot 2^{k+k_0}}\right)^{\frac{3}{5}} + \left(\frac{d_{k+k_0+1}}{M \cdot 2^{k+k_0}}\right)^{\frac{3}{5}} \\ &\leq \left(1 + 2^{\frac{3}{5}}\right) \sum_k \left(\frac{d_k}{M \cdot 2^k}\right)^{\frac{3}{5}},\end{aligned}$$

which is equivalent to this needed condition.

Now let $\hat{H}_1, ..., \hat{H}_d$ and $\lambda_1, ..., \lambda_d$ denote the eigenvalues of $\hat{H}$ and $\nabla^2 f(\boldsymbol{x}_B)$, respectively, in an non-decreasing order. By Von Neumann's trace inequality, we have $\epsilon_H^2 \geq \sum_i (\hat{H}_i - \lambda_i)^2$. Therefore, we can apply a similar analysis to control the growth of the R function. ■

By choosing the Hessian estimation algorithm in Appendix B, we have the following inequality for sufficiently large T and a universal constant C.

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}) - f^* | \boldsymbol{x}_B] \leq C \cdot \frac{\rho^{\frac{2}{3}}}{T^{\frac{2}{3}}} R(\nabla^2 f(\boldsymbol{x}_B)) \left(1 + \frac{||\nabla f(\boldsymbol{x}_B)||_2^2}{(n\rho^2)^{\frac{1}{3}}}\right) \\ + 2\frac{d^2 \rho^{\frac{4}{3}} ||\nabla f(\boldsymbol{x}_B)||_2^2}{T^{\frac{1}{3}} M^2} + \frac{8\rho ||\nabla f(\boldsymbol{x}_B)||_2^3}{M^3}.\end{aligned} \tag{16}$$

Then, the theorem can be proved based on the non-sensitivity of R and a bootstrapping guarantee provided in [17], which we cite below.

Proposition 13 (Bootstrapping Guarantees on Gradient Norm [17]) *There exists a bootstrapping algorithm, such that the point x_B it returns after T samples satisfies the following condition uniformly for all function f that satisfies assumptions (A1)-(A3).*

$$\mathbb{E}\left[\left\|\nabla f(x_B)\right\|^3\right] = o(T^{-2/3}). \tag{17}$$

Appendix E. Proof Details of Theorem 2

Note that in the special isotropic case ($\hat{H} = \lambda I_d$), the $\hat{H}^{-1}$-error reduces to the squared norm, which has been analyzed in the minimax case in [17]. We show that our proposed Spectrally Grouped Estimator allows a natural adaptation of this this methodology.

Specifically, note that the restriction of a function with Lipschitz Hessian to an affine subspace still satisfies the same Lipschitz Hessian condition, the contribution of the error from entries in group G_i can be directly analyzed via the arguments we provided for Spherical Sampling in Appendix A.1. Therefore, by Theorem 6, the $\hat{H}^{-1}$-error of the Spectrally Grouped Estimator is characterized as follows.

$$\mathbb{E}\left[(\hat{g} - \nabla f(x))^\top \hat{H}^{-1} (\hat{g} - \nabla f(x))\right] \leq \sum_i C \cdot \frac{\rho^{2/3} d_i}{n_i^{2/3}} \cdot \frac{1}{\hat{H}_{\min} \cdot 2^i} \cdot \left(1 + \|g_i\|_2^2 \left(\frac{1}{n\rho^2}\right)^{1/3}\right),$$

where g_i is the component of $\nabla f(x)$ in group G_i. Our theorem is proved by applying the choice of n_i in Algorithm 2, which optimizes the achieved $\hat{H}^{-1}$-error.