Hessian-Dependent Sample Complexity in Zeroth-Order Stochastic Optimization: Suboptimality of Convex-Support Sampling and Optimal Sample Complexity

Mengtian Hong
University of Glasgow
2960081H@student.gla.ac.uk

Jason D. Lee
University of California, Berkeley
jasondlee@berkeley.edu

Qian Yu
University of California, Santa Barbara
qianyu02@ucsb.edu

Abstract

Zeroth-order stochastic optimization is a 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 exploit local Hessian geometry to achieve optimal sample efficiency. We introduce the Spectrally Grouped Estimator (SGE), a novel gradient estimator that samples over a non-convex 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 conventional convex-sets sampling baseline methods. We complement these results with the first tight analyses of the baseline schemes, revealing a shared performance bottleneck and thus emphasizing the necessity of non-convex sampling for optimality. We further establish matching converse bounds that tightly characterize the optimal sample complexities within general subclasses of functions sharing the same Hessian at the global minimum, thereby proving the universal optimality of SGE with respect to the Hessian-dependent rates. This fully resolves an open conjecture from the preliminary version of this work.

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. Research under this topic has largely been directed towards clarifying the behavior of minimax sample complexities under a variety of regularity conditions. Foundational studies have investigated the optimal rates for both general and strongly convex functions [Shamir, 2017, Duchi et al., 2015], as well as the comprehension of optimal rates for functions characterized by Lipschitz continuity [Agarwal et al., 2011, Lattimore and Gyorgy, 2021, Bubeck et al., 2021] or Lipschitz gradients [Besbes et al., 2015, Agarwal et al., 2010, Shamir, 2013, Hazan and Levy, 2014]. While optimal rates are well-studied for such function classes, recent attention has shifted to higher-order smoothness [Agarwal et al., 2011, Lattimore and Gyorgy, 2021, Bubeck et al., 2021, Besbes et al., 2015, Agarwal et al., 2010, Hazan and Levy, 2014]. However, most of these studies target minimax rates that characterize worst-case performance over an entire function class, leaving open the question of how

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

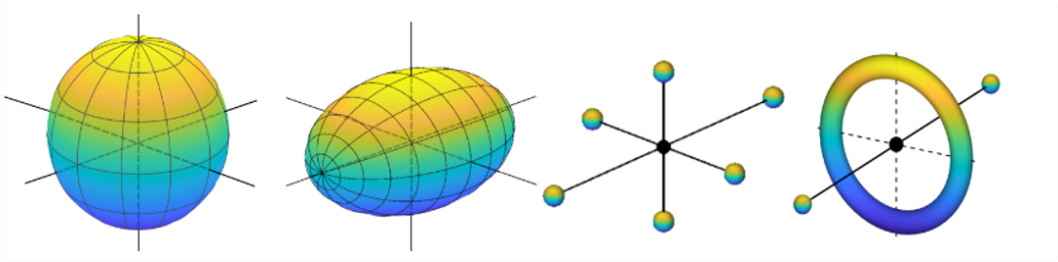

Figure 1: Four gradient estimation geometries: spherical sampling (isotropic sampling on a hypersphere), ellipsoidal sampling (anisotropic sampling on a hyperellipsoid), entry-wise sampling (sampling along coordinate axes), and the proposed spectrally grouped estimator (sampling on a union of sphere sections).

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.

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

instance-specific geometry, namely the spectral structure of the Hessian at the global minimizer, determines the optimal sample complexity for a given problem instance.

This work focuses on the scenario within the second-order smoothness conditions, building on the minimax framework of Yu et al. [2024] while targeting a strictly finer characterization. In particular, the minimax rate reflects the hardest instance within the function class and does not distinguish between instances whose Hessian spectra differ substantially, whereas the difficulty of a specific instance is in fact governed by the full eigenvalue distribution of its Hessian. While Yu et al. [2024] establish the minimax rate $O(dT^{-2/3})$ over the full smooth strongly convex class, this worst-case rate does not capture how the Hessian spectrum at the global optimizer affects sample complexity. To address this challenge, we formulate and study the minimax regret over subclasses of functions whose Hessian at the global minimizer equals a fixed matrix H, targeting optimal instance-dependent rates that reflect the full spectral geometry of H.

The principal challenge in obtaining these instance-dependent rates is designing a 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 [Agarwal et al., 2010, Shamir, 2017, Akhavan et al., 2020, Novitskii and Gasnikov, 2021] and more adaptive, non-isotropic sampling on a hyperellipsoid [Yu et al., 2024, Abernethy et al., 2008].

However, our analysis reveals a shared limitation of the spherical and generalized single-ellipsoid estimators studied here: both 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. More broadly, our lower-bound construction implies that any estimator based on uniform sampling from a single convex body remains orderwise suboptimal relative to SGE, even when the sampling body is chosen with complete knowledge of the local Hessian geometry. A fully general characterization of convex-body-based sampling strategies, including schemes that alternate or randomize among multiple convex bodies, is beyond the scope of this comparison; such schemes may induce an aggregate non-convex support similar in spirit to SGE. An alternative, the entry-wise estimator, decomposes the problem along the standard basis, but its prohibitive sample complexity renders it highly suboptimal [Blum, 1954, Dai, 2016]. These limitations suggest rethinking the sampling geometry and going beyond a single convex support.

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 SGE achieves a strictly lower instance-dependent simple regret than the baseline estimators analyzed in this work. A detailed comparison of the regret bounds is provided in Table 1. These gains are particularly pronounced for heterogeneous spectra spanning multiple spectral groups with substantial multiplicities, where SGE separates estimation across groups while retaining joint spherical estimation within each group. In particular, for polynomial spectra, the estimator demonstrates orderwise improved dependence on the dimension d, as illustrated in Figure 2. **(ii)** We establish the tight worst-case regret bounds for three classes of baseline estimators, characterizing the $H_{\min}$ bottleneck of spherical and generalized single-ellipsoid sampling and the sample inefficiency of entry-wise estimation. We further establish a corresponding lower bound for the optimal Hessian-dependent instance-dependent rates, demonstrating that our proposed achievability guarantees are tight and confirming the universal optimality of our estimator (in the instance-dependent sense of $\mathcal{R}(T; H)$, as defined in Section 2 rather than minimax optimality in d and T). **(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. Importantly, our established matching converse bounds completely characterize the optimal Hessian-dependent sample complexities. We summarize our main results in the following theorem.

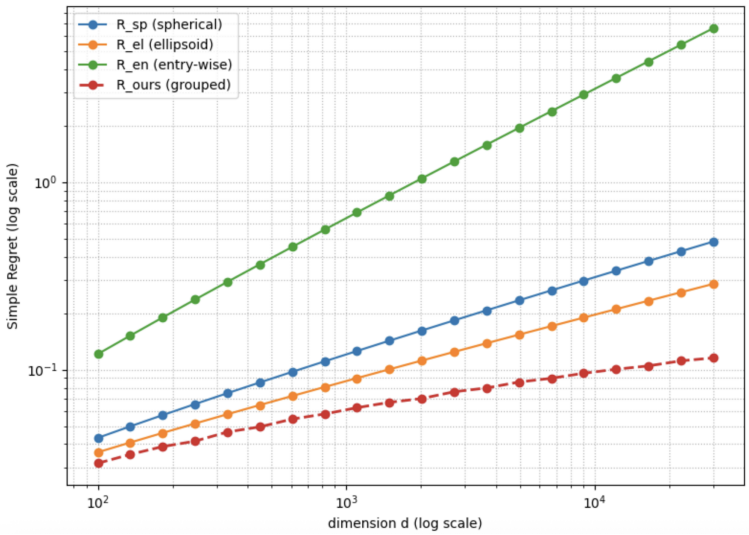

Figure 2: An illustration of the order-wise gain of the proposed scheme over baseline designs for a representative instance where the Hessian eigenvalues at the global minimizer are equally spaced ($H_i = iM$ for $i = 1, \ldots, d$). The simple regrets are normalized by a common factor $\rho^{2/3}T^{-2/3}M^{-1}$.

Informal Theorem (Main Result)

For second-order smooth, strongly convex functions with Hessian H at the minimizer, the optimal simple regret satisfies, asymptotically up to universal constants,

$$\mathcal{R}(T;\, H) = \Theta\left(\rho^{2/3}T^{-2/3}\left[\sum_k \left(\frac{d_k}{\lambda_k}\right)^{3/5}\right]^{5/3}\right),$$

where λ_k and d_k denote the representative eigenvalue and the number of eigenvalues of the k-th dyadic spectral group of H. This rate is universally achieved by the proposed Spectrally Grouped Estimator (SGE), without prior knowledge of H.

Related works. **(i) Minimax rates.** Classical results establish minimax sample complexities for zeroth-order and bandit convex optimization under various regularity conditions [Shamir, 2017, Duchi et al., 2015, Besbes et al., 2015, Agarwal et al., 2010, 2011, Hazan and Levy, 2014]. More recent works exploit higher-order smoothness to improve sample complexity [Akhavan et al., 2020, Novitskii and Gasnikov, 2021, Yu et al., 2024], but all characterize worst-case performance over a function class rather than rates depending on the full Hessian spectrum of individual instances. In comparison, the present work extends beyond the minimax formulation and establishes universal optimality over general local Hessian instances.**(ii) Instance-dependent rates.** Instance-dependent rates are

a well-studied concept in bandit and online learning, where problem difficulty is characterized by suboptimality gaps or other instance-specific quantities. Classical work established gap-dependent logarithmic regret lower bounds and finite-time algorithms matching these bounds [Lai and Robbins, 1985, Auer et al., 2002]; see also Bubeck and Cesa-Bianchi [2012] for a comprehensive treatment. Beyond standard multi-armed bandits, instance-dependent analyses have also been developed for structured and combinatorial action spaces [Combes et al., 2015, 2017, Zimmert et al., 2019], and cascading bandits [Kveton et al., 2015, Zhong et al., 2021]. Yu et al. [2023] introduced an information-theoretic framework for Hessian-dependent instance-dependent rates in the continuous zeroth-order setting for quadratic loss functions. The present work extends significantly beyond the quadratic case to general second-order smooth strongly convex functions and tightly characterizes the optimal Hessian-dependent sample complexity. **(iii) Sampling geometries.** Isotropic sampling on hyperspheres [Agarwal et al., 2010, Shamir, 2017] and non-isotropic or convex-body sampling geometries [Yu et al., 2024, Abernethy et al., 2008] both operate within a single convex support and therefore cannot decouple the estimation difficulty across eigendirections; entry-wise sampling [Blum, 1954, Dai, 2016] avoids this coupling by operating along coordinate axes, but incurs prohibitive sample complexity. Our analysis formalizes this tradeoff, showing that single-convex-support methods share a fundamental $H_{\min}$ bottleneck while fully entry-wise sampling is overly sample-inefficient.

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 $H \triangleq \nabla^2 f(x^\star)$ represent the Hessian at the global minimizer $x^\star$. Finally, we employ conventional notations (i.e., O, Ω) to describe regret bounds concerning the total number of samples, denoted by T.

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 $\mathrm{Var}[w_t|\{(x_\tau, y_\tau)\}_{\tau<t}, x_t] \le 1$. 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 in Yu et al. [2024].

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 standard class of second-order smooth, strongly convex functions with well-behaved local geometry. These assumptions are standard in smooth strongly convex optimization [Boyd and Vandenberghe, 2004] and have been adopted in prior zeroth-order works under higher-order smoothness [Akhavan et al., 2020, Yu et al., 2024]. Among the assumptions, the Frobenius-norm Lipschitz Hessian condition (Assumption 2.1) is particularly natural for our setting, as it is the regime where the minimax simple regret has been tightly characterized [Yu et al., 2024], making it a natural baseline for our instance-dependent analysis.

Assumption 2.1 (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 with respect to the Frobenius norm:

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

Assumption 2.2 (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, $\nabla^2 f(x) \succeq MI$.

Assumption 2.3 (Bounded distance to optimum)**.** We assume the unique global minimizer $x^\star$ of the function f is located within a ball of a known radius $R > 0$ from the origin, i.e., $\|x^\star\|_2 \le R$.

Our goal is to go beyond the worst-case minimax perspective and exploit instance-dependent rates that reflect the full spectral geometry of the Hessian at the global minimizer. While the classical minimax rate over the full class $\mathcal{F}$ characterizes the hardest instance [Yu et al., 2024], it does not distinguish between instances whose Hessian spectra differ substantially. To capture this finer structure, we introduce the following. Let H be a positive definite matrix, and let $\mathcal{F}_H$ denote the subclass of all

$f \in \mathcal{F}$ satisfying $\nabla^2 f(x^\star) = H$. The instance-dependent difficulty is then captured by the minimax simple regret over the subclass $\mathcal{F}_H$:

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

By fixing $\nabla^2 f(x^\star) = H$, the subclass $\mathcal{F}_H$ isolates the effect of spectral heterogeneity on sample complexity. Here, optimality is understood in the instance-dependent sense of $\mathcal{R}(T; H)$, rather than solely in terms of the dependence on d and T. Under this formulation, single-convex-support schemes remain limited by $H_{\min}$, whereas we aim for designs that exploit the full spectrum of H. We further aim for a universal algorithm that matches the rate of $\mathcal{R}(T; H)$ simultaneously across all instances without prior knowledge of H.

3 Main Results

We first establish the following achievability guarantee for the Hessian-dependent simple regret.

Theorem 3.1 (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 $\mathcal{F}_H$*

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

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

The achievability result above relies on the estimator-level guarantee of SGE. The following theorem characterizes this guarantee through its $\widehat{H}^{-1}$-risk, which gives rise to the spectral dependence in Theorem 3.1.

Theorem 3.2 ($\hat{H}^{-1}$-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 $n_{\mathrm{th}} \leq 4d$, the following holds uniformly over all function f that satisfy the Lipschitz Hessian condition*

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

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 $[\hat{H}_{\min} \cdot 2^k, \hat{H}_{\min} \cdot 2^{k+1})$.

Finally, we establish a matching lower bound, showing that the rate in Theorem 3.1 is optimal.

Theorem 3.3 (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.*

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

Together, Theorems 3.1 and 3.3 characterize the asymptotic instance-dependent regret over each fixed-Hessian subclass $\mathcal{F}_H$ under Assumption 2.1–Assumption 2.3. The upper bound assumes known ρ, M, and R but not H, whereas the lower bound holds even when ρ, M, R, and H are all known to the algorithm; the former allows bounded-variance noise, while the latter assumes i.i.d. standard Gaussian noise.

Algorithm 1 Achievability Framework for Theorem 3.1

Input: T, ρ, M
Let $n_{\mathrm{B}} = n_{\mathrm{H}} = n_{\mathrm{G}} = \lfloor (T-1)/3 \rfloor$ // Assign a constant fraction of samples to each routine
Let $\boldsymbol{x}_{\mathrm{B}} = \text{Bootstrapping}(n_{\mathrm{B}}, \rho, M)$ // Bootstrapping stage
Let $\hat{H} = \text{HessianEst}(\boldsymbol{x}_{\mathrm{B}},\ n_{\mathrm{H}})$ // Hessian Estimation; see Algorithm 6 in Appendix B
Let $\hat{\boldsymbol{g}} = \text{GroupedEst}(\boldsymbol{x}_{\mathrm{B}}, n_{\mathrm{G}})$ // Gradient Estimation; see Algorithm 2 in Section 4.2
Let scaling factor $\alpha = \max\left\{1, \frac{\rho}{M^{3/2}}\sqrt{\hat{\boldsymbol{g}}^{\intercal}\hat{H}^{-1}\hat{\boldsymbol{g}}}\right\}$, and $\boldsymbol{r} = -\frac{1}{\alpha}\hat{H}^{-1}\hat{\boldsymbol{g}}$ // Clipped Newton
Return $\boldsymbol{x} = \boldsymbol{x}_{\mathrm{B}} + \boldsymbol{r}$

4 Proof of Main results

We present the proofs of the three main theorems. For Theorem 3.1, the analysis follows the two-stage structure of Algorithm 1. By Proposition 4.2, the bootstrapping stage first produces a sufficiently accurate preliminary estimate, after which Hessian estimation and GroupedEst provide the estimates used in the clipped Newton refinement, with the latter characterized by Theorem 3.2. Proposition 4.1 then combines these guarantees into a simple-regret bound. Finally, a stability argument relates the $\hat{H}$-dependent factor to its H-dependent counterpart, yielding the rate in Theorem 3.1.

4.1 Algorithm Overview

Our proposed algorithm follows a two-stage structure (see Algorithm 1). First, a bootstrapping function returns a preliminary estimate $\boldsymbol{x}_{\mathrm{B}}$ of the global minimizer. This stage is required to produce a point x_B satisfying $\mathbb{E}\big[\|\nabla f(x_B)\|_2^3\big] = o(T^{-2/3})$, so that the gradient-dependent remainder terms in the second stage are lower-order. Such a guarantee is provided by Proposition 4.2. Then, a refinement stage performs a gradient-clipped Newton step based on the gradient and Hessian estimators $\hat{\boldsymbol{g}}$ and $\hat{H}$. The resulting simple regret depends on the quality of these individual steps, as characterized in the following proposition and proved in Section C.

Proposition 4.1. *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*

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

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 $\frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{x}})^{\intercal}\hat{H}(\boldsymbol{x} - \hat{\boldsymbol{x}})$.

The above proposition characterizes the simple regret in terms of the gradient estimation error the Hessian estimation error, and the gradient norm at x_B. The following result provides the guarantee on the latter quantity.

Proposition 4.2 (Bootstrapping Guarantees on Gradient Norm [Yu et al., 2024])**.** *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 Assumption 2.1 - Assumption 2.3.*

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

Specifically, the bootstrapping stage is designed to produce a point x_B sufficiently close to x^*, so that the gradient-dependent remainder terms are lower-order. This is achieved via an algorithm of [Yu et al., 2024], which performs a non-isotropic normalization of the Newton update, enabling the gradient norm guarantee required by our analysis.

The Hessian estimation error is controlled separately by Corollary B.2. In particular, the Hessian estimate $\hat{H}$ in Algorithm 1 satisfies

$$\mathbb{E}\Big[\Big\|\hat{H} - \nabla^2 f(\boldsymbol{x}_{\mathrm{B}})\Big\|_F\Big] = O\Big(d^{4/3}\rho^{2/3}n_{\mathrm{H}}^{-1/6}\Big).$$

Algorithm 2 GroupedEst

Input: $\boldsymbol{x}_B, \hat{H}, n$ // $\boldsymbol{x}_B$: query point; $\hat{H}$: Hessian estimate; n: number of samples. Return $\hat{g}$ as an estimator of $\nabla f(\boldsymbol{x}_B)$

Let $\hat{H}_1, \ldots, \hat{H}_d$ be the eigenvalues of $\hat{H}$, and let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ 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 for $i = 0, 1, \ldots, \lfloor \log_2(\hat{H}_{\max}/\hat{H}_{\min}) \rfloor$:

$G_i = \{\, k \mid \hat{H}_k \in [\hat{H}_{\min} \cdot 2^i,\ \hat{H}_{\min} \cdot 2^{i+1}) \,\}$

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

for all $G_i \neq \varnothing$ **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 for

Return $\hat{g} \leftarrow \sum_{i=1}^{I} \hat{g}_i$

4.2 Proof of Theorem 3.2

To minimize the gradient estimation error, we proposed an algorithm that samples the objective function on a union of spherical sections (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 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.

The key step in the proof of Theorem 3.2 is to first decompose and derive the $\hat{H}^{-1}$-risk under a general allocation of samples across spectral groups. Optimizing this allocation yields the sample allocation of n_i in Algorithm 2, and substituting this choice gives the rate stated in Theorem 3.2. We refer readers to Appendix E for the detailed proof.

We remark that, independent of the proof of Theorem 3.1–Theorem 3.3, the $\hat{H}^{-1}$-risk bound in Theorem 3.2 is in fact order-optimal in the regime of $\|\nabla f(\boldsymbol{x}_{\mathrm{B}})\| \to 0$ and $n \to \infty$: no gradient estimator can achieve a strictly better $\hat{H}^{-1}$-risk in this regime. The formal statement and proof of this optimality are given in Appendix G. Crucially, we also establish a modified version of the estimator that maintains this optimality beyond the small-gradient regime.

4.3 Proof of Theorem 3.1

According to the earlier analysis, the gradient estimation guarantee provided by Theorem 3.2 implies the following achievability result for Algorithm 1 when T is sufficiently large.

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

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 3.1 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 the ϵ_H and $||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2$ dependent terms.

4.4 Proof of Theorem 3.3

We first review an elementary one-dimensional class of hard instances, then explain how to lift this scalar construction to higher dimensions and adapt it to the local Hessian geometry at the global minimizer, to yield converse rates that match the achievability guarantees of the proposed gradient estimator.

4.4.1 A One-Dimensional Hard Instance

As a basic building block, our hard instances are constructed with the following function.

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

The key properties we reply on of the above function are that it has zero total integral, is nonzero only on a bounded interval, and is sufficiently smooth in the sense that its second derivative is Lipschitz continuous with a universal constant.

For any fixed parameters $x_0, h > 0$, the above function can be used to construct a pair of univariate functions, which can be viewed as a choice of hard-instance functions in the 1D special case, whose global minimizers are (orderwise) separated by $\rho x_0^2/h$, while satisfying a ρ-Lipschitz Hessian condition.

$$f_s(x; x_0) = hx^2 + \tfrac{1}{48}\rho x_0^3 \int_{-\pi}^{sx/x_0} g(z)\, dz, \quad s \in \{\pm 1\}.$$

4.4.2 A Dyadic Hessian-Adapted Construction

To prove the needed lower bound, we generalize the construction of the above f_s to multivariate functions, which map the task of estimating the global minimizer to a sequence of binary decision problems with a tension over shared sampling resources.

To illustrate the proof idea, we first consider the special case of the function subclasses $\mathcal{F}_H$ defined by hessian parameters H with a unique minimum eigenvalue $H_0 > M$, and the rest of the eigenvalues given by $H_k = H_0 \cdot 2^k$, each with a multiplicity of d_k. In this special case, we further assume that $d_k | d_{k+1}$, where we define $d_0 = 1$. The proof details and the proof of the general case are in Appendix F.

To simplify the presentation of the construction, we introduce the following notation to represent the input x. Under the eigenbasis of H, we denote the entries of the input x that corresponds to eigenvalue H_k by $x_{k,0}, ..., x_{k,d_k-1}$. Due to the divisibility assumption of $d_k | d_{k+1}$, there exists a partition of the entries $\{0, 1, ..., d_j - 1\}$ into d_i subsets of equal size for any $0 \le i \le j \le \log_2(H_{\max}/H_0)$ (e.g., by congruence), and we denote them by

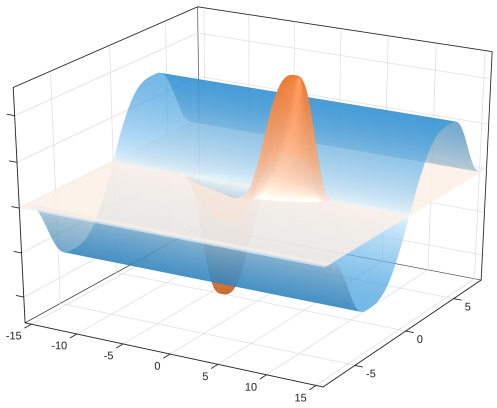

Figure 3: An illustration of the hard instance function f in Theorem 3.3, plotted along a representative direction as a projected view of the high-dimensional construction. The red and blue curves correspond to two candidate modulation functions that differ only in the location of the global minimizer; distinguishing between them is required to achieve estimation error below a given threshold.

$\{\mathcal{I}_{(i,j),\ell}\}_{\ell\in\{0,1,\ldots,d_i-1\}}$. Furthermore, by transitivity of inclusion, we can fix an arbitrary choice of such partitions such that $\mathcal{I}_{(j,k),m} \subseteq \mathcal{I}_{(i,k),\ell}$ for all $m \in \mathcal{I}_{(i,j),\ell}$.

Our constructed function class is parameterized with binary indicators $s_{k,\ell} \in \{\pm 1\}$, where the integer indices range over $0 \le k \le \log_2(H_{\max}/H_0)$, $0 \le \ell \le d_k - 1$. The function is in the following general form

$$f(x) = \frac{1}{2}x^\intercal \tilde{H} x + \sum_{j=0}^{\log_2\left(\frac{H_{\max}}{H_0}\right)} \sum_{\ell=0}^{d_j-1} \tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x)), \tag{4}$$

where the base function $\tilde{f}_s$ is given by

$$\tilde{f}_s(y; y_0) = \frac{\rho y_0^3}{C} \int_0^{sy/y_0} g(z)dz,$$

the modulation term $\phi_{j,\ell}(x)$ is defined as

$$\phi_{j,\ell}(x) := r_j \cos^2\left(\min\left\{\frac{1}{r_j}\cdot\sqrt{\sum_{k>j}\sum_{m\in\mathcal{I}_{(j,k),\ell}} \frac{x_{k,m}^2}{2^{\frac{k-j}{4}}\left(\frac{d_k}{d_j}\right)^{\frac{1}{2}}}}, \frac{\pi}{2}\right\}\right),$$

the scale parameters are chosen as and $r_k = \rho^{-1/3}t_k^{-1/6}$, $t_k = \frac{(d_k/H_k)^{\frac{3}{5}}T}{\sum_j (d_j/H_j)^{\frac{3}{5}}}$, $\tilde{H}$ is a fixed positive definite matrix to be designed to match the Hessian condition at the global minimum, and $C < 258000$ is a sufficiently large universal constant to ensure the required smoothness condition. The modulation term $\phi_{j,\ell}(x)$ couples the perturbation associated with group j to a weighted magnitude of the coordinates in higher spectral groups. As this quantity increases, $\phi_{j,\ell}(x)$ decreases and reduces the effective scale of the corresponding perturbation, creating the cross-group resource-allocation tradeoff formalized by the dual condition in Appendix F.4.

The existence of matrix $\tilde{H}$ for $f \in \mathcal{F}_H$ for all choices of s parameters is proved by proving the contractiveness of a function $U(\tilde{H}) := \tilde{H} + H - \nabla^2 f(x^*_{\tilde{H}})$, which is a self-map within a neighborhood of H, where $x^*_{\tilde{H}}$ is a global minimizer of f. By the unique fixed point principle, the existence of the unique fixed point $\tilde{H}$ such that $U(\tilde{H}) = \tilde{H}$ implies the required Hessian condition. We refer readers to appendix F.2 for proof details.

To provide a rigorous lower bound of the simple regret over the above hard-instance class, we present a quadratic-approximation-based lower bound on the simple regret to show the cost is fundamentally characterized by the estimation of the binary $s_{j,\ell}$ indicators. Specifically, to lower bound the sample complexities worst case, we consider the Bayes setting where all entries of $\boldsymbol{s}$ are i.i.d. uniformly random. Note that Bayes-averaged regret provide a lower bound to the achieved regret in the worst case over the choice of function f, we only need to show that under this uniform prior, the expected simple regret achieve by any algorithm is lower bounded by the rate stated in Theorem 3.3. Let $p_{j,\ell}$ denote the posterior distribution of $s_{j,\ell} = 0$ conditioned on the history from first $T-1$ rounds. A quadratic characterization of the objective function f implies that the following inequality holds for sufficiently large T (see Appendix F.3 for details).

$$E[f(x_T)] - f^* \ge \sum_{j,\ell} \frac{\rho^2 r_j^4}{C^2 H_j} \mathbb{E}\left[2p_{j,\ell}(1-p_{j,\ell}) - \frac{1}{16}\right]. \tag{5}$$

Note that individually, we have the following bound under the Gaussian noise condition, which can be derived from a standard KL-divergence-based approach (e.g., see proofs in Yu et al. [2023]),

$$\mathbb{E}[p_{j,\ell}(1-p_{j,\ell})] \ge \frac{1}{4}\mathbb{E}\left[1 - \sum_{t=1}^{T-1}\left(\tilde{f}_{s_{j,\ell}}(x_{t;j,\ell};\, \phi_{j,\ell}(x_t))\right)^2\right],$$

where $x_{t;j,\ell}$ denotes the entry of the action x_t at round t with index (j,ℓ). This immediately implies a tension of resource allocation among sampling these individual entries emerges, similar to the quadratic case in Yu et al. [2023], where to obtain a better estimate of any $s_{j,\ell}$, we want to sample at

points where the absolute values of $\tilde{f}_{s_{j,\ell}}(x_{t;j,\ell};\, \phi_{j,\ell}(x_t))$ is larger, while these amplitudes are not maximized at the same points across different j indices.

However, the following dual condition can be found, which holds for sufficiently large C, to imply a needed uniform bound. We provide a proof in Appendix F.4

$$T \sum_{j,\ell} \frac{\rho^2 r_j^4}{H_j} \left(\tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x)) \right)^2 \leq \sum_{j,\ell} \frac{\rho^2 r_j^4}{2H_j}.$$

This condition essentially states that on average there is no allocation that can learn this hard instance class beyond reducing the regret more than a constant factor compared to a random guess.

Formally, by applying this dual condition to the earlier derived inequalities, we have that,

$$E[f(x_T)] - f^* \geq \sum_{j,\ell} \frac{\rho^2 r_j^4}{8C^2 H_j} \geq \frac{\rho^{\frac{2}{3}}}{8C^2 T^{\frac{2}{3}}} \left(\sum_j \left(\frac{d_j}{H_j} \right)^{\frac{3}{5}} \right)^{\frac{5}{3}},$$

which proves the needed lower bounded for this special class of H with integer division properties. The proof details and extensions to the general Hessian conditions are provided in Appendix F.

5 Conclusion and Future 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. We also established a matching lower bound, thereby proving the universal optimality of our proposed estimator and conclusively addressing the open conjecture regarding its efficiency. 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}) , together with a systematic finite-sample evaluation in controlled nonquadratic settings, remains future work. Additionally, the present analysis relies on strong convexity in essential steps, including initialization and bootstrapping into the local regime where the second-order analysis applies; extending the framework beyond strongly convex objectives would require substantial new techniques and remains an important direction for future work.

Acknowledgements and Disclosures

JDL acknowledges support of NSF IIS 2107304, NSF CCF 2212262, NSF CAREER Award 2540142, NSF 2546544, NSF CCF 2019844 and ONR N00014-24-1-2639. This material is based upon work supported by the U.S. National Science Foundation under Cooperative Agreement No. 2433450. Q. Yu acknowledges the support of a UC faculty research grant.

References

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 263–274, 2008.

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

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.

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.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

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

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

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.

Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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.

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.

Michael Edelstein. On fixed and periodic points under contractive mappings. *Journal of the London Mathematical society*, 1(1):74–79, 1962.

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

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

Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 767–776. PMLR, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

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

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

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

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

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.

Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason D Lee. Sample complexity for quadratic bandits: Hessian dependent bounds and optimal algorithms. *Advances in Neural Information Processing Systems*, 36:35121–35138, 2023.

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.

Zixin Zhong, Wang Chi Cheung, and Vincent Y. F. Tan. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):9915–9980, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692, 2019.

Algorithm 3 SphericalGradEst [Shamir, 2017]

// 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 $k = 1$ **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 for
Return $\hat{g} = \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} g_k$

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 worst-case simple-regret guarantees following the steps in Section 4.3. This analysis serves to precisely characterize their 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 [Ghadimi and Lan, 2013, Akhavan et al., 2020], relying on isotropic sampling along random directions drawn from the unit hypersphere. To characterize its performance in the $\hat{H}^{-1}$ norm, we first prove the following preliminary result.

Proposition A.1. *For any fixed x, r, n, the estimator $\hat{g}$ returned by SphericalGradEst satisfies the following inequality for any unit vector $\boldsymbol{e}$.*

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

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

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

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.

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

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 (6). On the other hand, the contribution due to the bias of the estimation can be characterized as follows.

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

where the second inequality is due to the sharp characterization of the estimation bias in [Yu et al., 2024]. By choosing the optimal sampling radius that balances the two components, we obtain the following characterization.

Theorem A.2 ($\hat{H}^{-1}$-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 $n \geq 2$, there exists a choice of r such that the following holds uniformly over all functions f satisfying the Lipschitz Hessian condition*

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

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.

Proof. The $\hat{H}^{-1}$-risk decomposes into a variance term and a bias term. By Proposition A.1, the variance contribution to the $\hat{H}^{-1}$-risk is bounded by

$$\sum_{i=1}^{d} \hat{H}_i^{-1} \mathrm{Var}[\hat{g} \cdot e_i] \leq \frac{C}{n} \cdot \left(\sum_{i=1}^{d} \hat{H}_i^{-1}\right) \left(\|\nabla f(x)\|_2^2 + d\rho^2 r^4 + \frac{d}{r^2}\right),$$

and the bias contribution satisfies

$$\mathbb{E}[\hat{g} - \nabla f(x)]^\top \hat{H}^{-1} \mathbb{E}[\hat{g} - \nabla f(x)] \leq \frac{r^4 \rho^2 d}{4\hat{H}_{\min}(d+2)^2}.$$

where C is a universal constant absorbing the numerical factors in Proposition A.1. The gradient-dependent term $\|\nabla f(x)\|_2^2$ is absorbed into the multiplicative factor $\left(1 + \|\nabla f(x)\|_2^2 \left(\frac{1}{n\rho^2}\right)^{1/3}\right)$ in the theorem, while the nonlinearity-variance term $\frac{d \cdot \rho^2 r^4}{n}$ is lower-order for sufficiently large n. Choosing $r \asymp \left(\frac{d^2 \widehat{H}_{\min} \sum_{i=1}^{d} \widehat{H}_i^{-1}}{n\rho^2}\right)^{1/6}$ to balance the leading terms of the variance and bias, yields the bound stated in the theorem. □

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 [Yu et al., 2024, Abernethy et al., 2008]. 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.

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 A.3. *For any fixed x, Z, n, the estimator $\hat{g}$ returned by EllipsoidGradEst satisfies the following inequality for any unit vector $\boldsymbol{e}$.*

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

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 A.1) to g. □

Algorithm 4 EllipsoidGradEst [Yu et al., 2024]

// Non-isotropic Ellipsoidal gradient estimator
Input: x, Z, n //x: query point; Z: shaping matrix $\in \mathbb{R}^{d\times d}$; n: number of samples. The resulting $\hat{g}$ provides a stochastic estimate of $Z^\intercal \nabla f(x)$
for $k = 1$ **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 for
Return $\hat{g} = \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} g_k$

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 Yu et al. [2024]. Specifically, as shown by Yu et al. [2024], 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

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

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

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

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

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

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

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

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

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

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

Theorem A.4 ($\hat{H}^{-1}$-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 $n \ge 2$, the following holds uniformly over all functions f that satisfy the Lipschitz Hessian condition:*

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

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.

Algorithm 5 EntrywiseGradEst [Blum, 1954, Dai, 2016]

Input: $x, \{r_k, e_k, n_k\}_{k=1}^d$ //x: query point; r_k: step sizes for coordinate k; n_k: number of samples for coordinate k; $\{e_k\}_{k\in\{1,\dots,d\}}$: an orthonormal basis of $\mathbb{R}^d$. Return $\hat{g}$ as an estimator of $\nabla f(x)$

for $k = 1$ **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 for

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

Proof. The $\hat{H}^{-1}$-risk decomposes into a variance term and a bias term. Taking $Z = \mathrm{diag}(z_i)$ under the eigenbasis of $\hat{H}$, the variance contribution to the $\hat{H}^{-1}$-risk is bounded by Proposition A.3 as

$$\frac{C}{n}\left(\sum_i \hat{H}_i^{-1} z_i^{-2}\right)\left(\|Z^\top \nabla f(x)\|_2^2 + d\rho^2 \|Z\|^4 + d\right),$$

and the bias contribution satisfies

$$\frac{\|ZZ^\top\|_F^2 \rho^2}{4\hat{H}_{\min}(d+2)^2},$$

where C is a universal constant. Since $Z = \mathrm{diag}(z_i)$, we have $\|ZZ^\top\|_F^2 = \sum_i z_i^4$. The $\|Z^\top \nabla f(x)\|_2^2$ term corresponds to the gradient-dependent multiplicative factor in the theorem, while the nonlinearity-variance term $d\rho^2\|Z\|^4$ is lower-order for sufficiently large n. Since the pure-noise variance term $\frac{Cd}{n}\sum_i \hat{H}_i^{-1} z_i^{-2}$ and the bias term $\frac{\rho^2 \sum_i z_i^4}{\hat{H}_{\min} d^2}$ are per-i separable, choosing $z_i \asymp \left(\frac{d^3 \hat{H}_{\min} \hat{H}_i^{-1}}{n\rho^2}\right)^{1/6}$ to balance each i's contribution and absorbing constants into C yields the bound stated in the theorem. □

A.3 Entry-wise sampling

The entry-wise (or coordinate-wise) estimator presents a clear and effective strategy for gradient estimation [Blum, 1954, Mokkadem and Pelletier, 2007]. 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.

Proposition A.5 (Entrywise Estimation Error of EntrywiseGradEst)**.** *For any fixed x, n_i, and basis direction e_i, the entrywise estimator $\hat{g}_i$ obtained in EntrywiseGradEst satisfies the following inequality:*

$$\mathbb{E}\left[(\hat{g}_i - e_i \cdot \nabla f(x))^2\right] \;\le\; \frac{\rho^2 r_i^4}{36} \;+\; \frac{1}{2\lfloor n_i/2 \rfloor r_i^2}.$$

Proof. We start the variance characterization. Recall that in our earlier analysis for SphericalGradEst (Proposition A.1), 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_i, and therefore the first two terms vanish. The remaining variance comes solely from averaging n_i independent noisy oracle evaluations. Hence,

$$\mathrm{Var}[\hat{g}_i] \le \frac{1}{2\lfloor n_i/2 \rfloor r_i^2}.$$

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.

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

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 A.6 ($\hat{H}^{-1}$-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:*

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

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

Proof. Let $\{e_i\}$ be the eigenbasis of $\hat{H}$ with corresponding eigenvalues $\hat{H}_i$. Since the entry-wise estimator estimates each coordinate independently along this basis, we have

$$(\hat{g}-\nabla f(x))^\top \hat{H}^{-1}(\hat{g}-\nabla f(x)) = \sum_i \hat{H}_i^{-1}(\hat{g}_i-\nabla_i f(x))^2.$$

Taking expectations and applying Proposition A.5 to each coordinate gives

$$\mathbb{E}\left[(\hat{g}-\nabla f(x))^\top \hat{H}^{-1}(\hat{g}-\nabla f(x))\right] \le C\sum_i \hat{H}_i^{-1}\left(\rho^2 r_i^4+\frac{1}{n_i r_i^2}\right),$$

where C is a universal constant. Since the bound is per-i separable, choosing $r_i \asymp (n_i\rho^2)^{-1/6}$ to balance the bias and variance terms for each coordinate gives

$$\mathbb{E}\left[(\hat{g}-\nabla f(x))^\top \hat{H}^{-1}(\hat{g}-\nabla f(x))\right] \le C\rho^{2/3}\sum_i \hat{H}_i^{-1}n_i^{-2/3}.$$

Choosing $n_i \asymp n\cdot\frac{\hat{H}_i^{-3/5}}{\sum_j \hat{H}_j^{-3/5}}$ to optimize the allocation across coordinates and substituting yields the bound stated in the theorem. □

A.4 Tightness of Baseline Bounds and Intrinsic Suboptimality

The $\hat{H}^{-1}$-risk bounds established in Theorems A.2, A.4, and A.6 are essentially sharp. The rates cannot generally be improved by tuning the sampling radius r (or the shaping matrix Z in the ellipsoid or the per-coordinate radii $\{r_i\}$ in the entry-wise case), since for each estimator, there exist hard instance functions that are locally polynomial, for which the corresponding $\hat{H}^{-1}$-risk admit a simple closed-form expression, and the optimal choice of estimation parameters leads to $\hat{H}^{-1}$-risk that matches with our provided upper bounds.

These hard instances show that the suboptimality of conventional baseline estimators is intrinsic. On the other hand, as illustrated in Figure 2, there are Hessian spectra for which these intrinsic limitations lead to order-wise gaps between the baseline estimators and the proposed SGE. In such instances, SGE achieves a strictly better rate by adapting to the spectral structure of the Hessian.

B Hessian Estimator

We employ a standard Hessian estimation algorithm, adapted from Yu et al. [2024] and included as 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.

Lemma B.1. *Let f be a function satisfying the Lipschitz Hessian (Assumption 2.1) conditions with parameter ρ. Let $\hat{H}$ be the output of the `HessianEst` subroutine (Algorithm 6) with inputs x, n_H. Then, for any sufficiently large n_{H}, the estimation error in the Frobenius norm is bounded as follows:*

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

where $C>0$ is a universal constant, $n=\lfloor n_{\mathrm{H}}/(2d^2+1)\rfloor$ is the per-entry sample count and $r=n^{-1/6}\rho^{-1/3}$ is the sampling radius as defined in Algorithm 6.

Algorithm 6 HessianEst (Modified from Yu et al. [2024])

Input: x, n_{H} // x: query point; n_{H}: total sample budget for Hessian estimation.

Set $n = \left\lfloor \dfrac{n_{\mathrm{H}}}{2d^2+1} \right\rfloor$ as the per-entry sample count

Set $r = n^{-1/6}\rho^{-1/3}$

The value of the function y is initially estimated by averaging over n samples of f

for $k = 1$ **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 $l = k+1$ **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 for

end for

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 $\hat{H}$

Proof. We first analyze the unprojected estimator $\widehat{H}_0$. Let $H = \nabla^2 f(x)$. We decompose the proof into the analysis of diagonal and off-diagonal entries.

For each diagonal entry, the estimator is

$$\widehat{H}_{kk} = \frac{\bar{y}_{+,k} + \bar{y}_{-,k} - 2\bar{y}}{r^2},$$

where $\bar{y}_{+,k}$, $\bar{y}_{-,k}$, and $\bar{y}$ denote the empirical averages of noisy oracle evaluations at $x + re_k$, $x - re_k$, and x, respectively.

We first bound the bias. By the second-order Taylor expansion with integral remainder,

$$f(x + re_k) + f(x - re_k) - 2f(x) = r^2 \partial^2_{kk} f(x) + R_{+,k} + R_{-,k}.$$

where $R_{+,k}$ and $R_{-,k}$ denote the integral remainder terms given by

$$R_{+,k} = \int_0^r (r-t)\, e_k^\top \left(\nabla^2 f(x + te_k) - \nabla^2 f(x)\right) e_k \, dt,$$

and

$$R_{-,k} = \int_0^r (r-t) e_k^\top \left(\nabla^2 f(x - te_k) - \nabla^2 f(x)\right) e_k \, dt.$$

By the Lipschitz Hessian condition,

$$|R_{+,k}| \le \int_0^r (r-t)\rho t \, dt = \frac{\rho r^3}{6}, \qquad |R_{-,k}| \le \frac{\rho r^3}{6}.$$

Therefore,

$$\left|\mathbb{E}[\widehat{H}_{kk}] - \partial^2_{kk} f(x)\right| \le \frac{1}{3}\rho r.$$

On the other hand, since each empirical average is computed from n independent noisy oracle evaluations with variance bounded by a universal constant,

$$\mathrm{Var}(\widehat{H}_{kk}) = \mathrm{Var}\left(\frac{\bar{y}_{+,k} + \bar{y}_{-,k} - 2\bar{y}}{r^2}\right) \le \frac{6}{nr^4}.$$

Combining the bias and variance bounds gives

$$\mathbb{E}\left[\left(\widehat{H}_{kk} - \partial^2_{kk} f(x)\right)^2\right] \le \frac{1}{9}\rho^2 r^2 + \frac{6}{nr^4}.$$

Next, consider an off-diagonal entry with $k < l$. The estimator is

$$\widehat{H}_{kl} = \frac{\bar{y}_{++} + \bar{y}_{--} - \bar{y}_{+-} - \bar{y}_{-+}}{4r^2}.$$

By symmetry of the four-point finite difference,

$$f(x + re_k + re_l) + f(x - re_k - re_l) - f(x + re_k - re_l) - f(x - re_k + re_l) = 4r^2 \partial^2_{kl} f(x) + R_{kl},$$

where R_{kl} denotes the remainder term satisfying

$$|R_{kl}| \leq \frac{4\sqrt{2}}{3} \rho r^3,$$

by the Lipschitz Hessian condition. Hence,

$$\left| \mathbb{E}[\widehat{H}_{kl}] - \partial^2_{kl} f(x) \right| \leq \frac{\sqrt{2}}{3} \rho r.$$

The variance is bounded similarly. Since the estimator is computed from a constant number of empirical averages and has denominator of $4r^2$, we have

$$\mathrm{Var}(\widehat{H}_{kl}) \leq \frac{1}{4nr^4}.$$

Therefore,

$$\mathbb{E}\left[\left(\widehat{H}_{kl} - \partial^2_{kl} f(x) \right)^2 \right] \leq \frac{2}{9} \rho^2 r^2 + \frac{1}{4nr^4}.$$

Finally, we sum the entrywise bounds. Since the Hessian is symmetric, the Frobenius norm can be written as

$$\|\widehat{H}_0 - \nabla^2 f(x)\|_F^2 = \sum_{k=1}^{d} \left(\widehat{H}_{kk} - \partial^2_{kk} f(x) \right)^2 + 2 \sum_{1 \leq k < l \leq d} \left(\widehat{H}_{kl} - \partial^2_{kl} f(x) \right)^2.$$

The number of entries counted above is

$$d + 2 \cdot \frac{d(d-1)}{2} = d^2.$$

Therefore,

$$\mathbb{E}\left[\left\| \widehat{H}_0 - \nabla^2 f(x) \right\|_{\mathrm{F}}^2 \right] \leq C \left(d^2 \rho^2 r^2 + \frac{d^2}{nr^4} \right),$$

where $C > 0$ is a universal constant that absorbs the explicit constants from the diagonal and off-diagonal bounds above.

It remains to account for the eigenspectrum projection. Since $\nabla^2 f(x) \succeq MI$ and $\widehat{H}$ is obtained by projecting $\widehat{H}_0$ onto the cone of symmetric matrices with eigenvalues at least M, this projection does not increase the Frobenius distance to $\nabla^2 f(x)$. Hence,

$$\|\widehat{H} - \nabla^2 f(x)\|_{\mathrm{F}} \leq \|\widehat{H}_0 - \nabla^2 f(x)\|_{\mathrm{F}}.$$

This completes the proof. □

Corollary B.2. *Under the conditions of Lemma B.1, choosing*

$$r = n^{-1/6} \rho^{-1/3}$$

yields

$$\mathbb{E}\left[\|\widehat{H} - \nabla^2 f(x)\|_F^2 \right] \leq C d^{8/3} \rho^{4/3} n_{\mathrm{H}}^{-1/3}$$

for a universal constant C.

Corollary B.2 follows directly from Lemma B.1 by substituting $r = n^{-1/6} \rho^{-1/3}$ and $n = \lfloor n_{\mathrm{H}} / (2d^2 + 1) \rfloor$, which balances the bias and variance terms.

C Proof of Proposition 4.1

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

$$f(\boldsymbol{x}) - f^* \leq 2\hat{f}(\boldsymbol{x})\,(1 + \epsilon_H/M) + 2\frac{\epsilon_H||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{M^2} + \frac{3\sqrt{2}\rho||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3}. \tag{7}$$

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 Yu et al. [2024] and obtain that

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

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 (7).

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 (7) can be applied. The proposition is proved through the following bound, which is by elementary geometry,

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

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

D Proof details of Theorem 3.1

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 D.1. *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.*

$$\rho^{\frac{2}{3}}\left(\sum_{k=0}^{\lceil \log_2(\hat{H}_{\max}/\hat{H}_{\min})\rceil}\left(\frac{\hat{d}_k}{\hat{H}_{\min}\cdot 2^k}\right)^{\frac{3}{5}}\right)^{\frac{5}{3}} \leq CR(\nabla^2 f(\boldsymbol{x}_B))\left(1 + \frac{\epsilon_H}{M}\right). \tag{8}$$

Proof. We first show that the LHS of inequality (8) 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 $\hat{H}_{\min} \in [M\cdot 2^{k_0}, M\cdot 2^{k_0+1})$. Recall that $\hat{d}_k$ is defined as the number of eigenvalues of $\hat{H}$ in $[\hat{H}_{\min}\cdot 2^k, \hat{H}_{\min}\cdot 2^{k+1})$, which belongs to $[M\cdot 2^{k+k_0}, M\cdot 2^{k+k_0+2})$. We have $\hat{d}_k \leq d_{k+k_0} + d_{k+k_0+1}$.

Hence,

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

which is equivalent to this needed condition.

It remains to prove the stability estimate that $R(\hat{H})$ is within a universal constant factor of $R(\nabla^2 f(\boldsymbol{x}_B))\left(1+\frac{\epsilon_H}{M}\right)$. For convenience, 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 a non-decreasing order. By Von Neumann's trace inequality, we have $\epsilon_H^2 \ge \sum_i (\hat{H}_i - \lambda_i)^2$. Therefore, we can apply a similar analysis to control the growth of the R function.

Specifically, for $j, \ell \ge 0$, we define the migration counts $m_{\ell j}$ to be the number of indices i such that $\lambda_i \in [M\cdot 2^\ell, M\cdot 2^{\ell+1})$ and $\hat{h}_i \in [M\cdot 2^j, M\cdot 2^{j+1})$. Note that the number of eigenvalues of $\hat{H}$ in $[M\cdot 2^j, M\cdot 2^{j+1})$ equals $\sum_{\ell\ge 0} m_{\ell j}$. We have that the $\hat{H}$-dependent factor in $R(\hat{H})$ can be expressed and characterized as

$$\sum_{j\ge 0}\left(\frac{\sum_{\ell\ge 0} m_{\ell j}}{M\cdot 2^j}\right)^{\frac{3}{5}} \le \sum_{\ell\ge 0}\sum_{j\ge 0}\left(\frac{m_{\ell j}}{M\cdot 2^j}\right)^{\frac{3}{5}}.$$

Fix ℓ and let $d_\ell^{(H)}$ denotes the number of eigenvalues of H in $[M\cdot 2^\ell, M\cdot 2^{\ell+1})$. We will show that

$$\sum_{j\ge 0}\left(\frac{m_{\ell j}}{M\cdot 2^j}\right)^{\frac{3}{5}} \le C\left(1+\frac{\epsilon_H}{M}\right)\left(\frac{d_\ell^{(H)}}{M\cdot 2^\ell}\right)^{\frac{3}{5}} \tag{9}$$

for a universal constant C. Summing this estimate over ℓ will give the desired comparison.

We split the estimated bins into two regions $j \ge \ell - 1$ and $j \le \ell - 2$. For the first region, no perturbation estimate is needed. Since $\sum_{j\ge \ell-1} m_{\ell j} \le d_\ell^{(H)}$, Hölder's inequality gives

$$\begin{aligned}\sum_{j\ge \ell-1}\left(\frac{m_{\ell j}}{M\cdot 2^j}\right)^{\frac{3}{5}} &= M^{-\frac{3}{5}}\sum_{j\ge\ell-1} m_{\ell j}^{\frac{3}{5}}\cdot 2^{-\frac{3}{5}j} \\ &\le M^{-\frac{3}{5}}\left(\sum_{j\ge\ell-1} m_{\ell j}\right)^{\frac{3}{5}}\left(\sum_{j\ge\ell-1} 2^{-\frac{3}{2}j}\right)^{\frac{2}{5}} \\ &\le 2\left(\frac{d_\ell^{(H)}}{M\cdot 2^\ell}\right)^{\frac{3}{5}},\end{aligned}$$

where the last step uses the fact that the geometric series is bounded by $2^{-\frac{3}{2}(\ell-1)}/(1-2^{-3/2})$.

It remains to control the lower tail $j \le \ell - 2$, where eigenvalues of H have moved down by at least one full dyadic scale. For convenience, define

$$P_\ell := \sum_{j\le \ell-2} m_{\ell j}.$$

Using Hölder's inequality once more,

$$\begin{aligned}\sum_{j\le\ell-2}\left(\frac{m_{\ell j}}{M\cdot 2^j}\right)^{\frac{3}{5}} &= M^{-\frac{3}{5}}\sum_{j\le\ell-2} m_{\ell j}^{\frac{3}{5}}\cdot 2^{-\frac{3}{5}j}\\ &\le M^{-\frac{3}{5}}\left(\sum_{j\le\ell-2} m_{\ell j}\right)^{\frac{3}{5}}\\ &= M^{-\frac{3}{5}}P_\ell^{\frac{3}{5}}\end{aligned}$$

On the other hand, note that when $\lambda_i \in [M\cdot 2^\ell, M\cdot 2^{\ell+1})$ and $\hat{h}_i \in [M\cdot 2^j, M\cdot 2^{j+1})$ with $j \le \ell-2$, then $\hat{h}_i < M\cdot 2^{j+1} \le M\cdot 2^{\ell-1}$ while $\lambda_i \ge M\cdot 2^\ell$. Hence

$$|\lambda_i - \hat{h}_i| \ge M\cdot 2^{\ell-1}.$$

Therefore,

$$P_\ell\cdot(M\cdot 2^{\ell-1})^2 \le \sum_{i=1}^{d}|\lambda_i-\hat{h}_i|^2 \le \epsilon_H^2,$$

and so

$$P_\ell \le 4\left(\frac{\epsilon_H}{M}\right)^2\cdot 2^{-2\ell}.$$

Combining the above results and using the fact that $P_\ell \le d_\ell^{(H)}$, we have

$$\begin{aligned}\sum_{j\le\ell-2}\left(\frac{m_{\ell j}}{M\cdot 2^j}\right)^{\frac{3}{5}} &\le 2M^{-\frac{3}{5}}\cdot\frac{\epsilon_H}{M}2^{-\ell}\left(d_\ell^{(H)}\right)^{\frac{1}{10}}\\ &\le 2\frac{\epsilon_H}{M}\left(\frac{d_\ell^{(H)}}{M\cdot 2^\ell}\right)^{\frac{3}{5}}\end{aligned}$$

Combining the upper-tail and lower-tail bounds yields inequality (9). Summing over ℓ, we conclude that

$$R(\hat{H}) \le CR(\nabla^2 f(\boldsymbol{x}_B))\left(1+\frac{\epsilon_H}{M}\right).$$

Finally, combining this with the first comparison between the $\hat{H}_{\min}$-based dyadic expression and $R(\hat{H})$ gives the inequality in the proposition statement. □

By choosing the Hessian estimation algorithm in Appendix B and applying Proposition D.1, we have the following inequality for sufficiently large T and a universal constant C.

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x})-f^*|\boldsymbol{x}_B] \le C\cdot\frac{\rho^{\frac{2}{3}}}{T^{\frac{2}{3}}}R(\nabla^2 f(\boldsymbol{x}_B))\left(1+\frac{||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{(T\rho^2)^{\frac{1}{3}}}\right)\\ + C\frac{d^{\frac{4}{3}}\rho^{\frac{2}{3}}||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{T^{\frac{1}{6}}M^2}+\frac{8\rho||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3}.\end{aligned}$$

where $\boldsymbol{x}$ denotes the output of the final stage of Algorithm 1, i.e., $\boldsymbol{x} = \boldsymbol{x}_B + \boldsymbol{r}$ with $\boldsymbol{r} = -\frac{1}{\alpha}\hat{H}^{-1}\hat{g}$.

Then, the theorem can be proved based on the non-sensitivity of R and a bootstrapping guarantee in Proposition 4.2. We first apply the non-sensitivity of R established in the proof of Proposition D.1, i.e.

$$\begin{aligned}R(\nabla^2 f(\boldsymbol{x}_B)) &\le 2R(H)(1+||H-\nabla^2 f(\boldsymbol{x}_B)||_{\mathrm{F}}/M)\\ &\le 2R(H)(1+\rho||\nabla f(\boldsymbol{x}_B)||_2/M^2),\end{aligned}$$

where the second inequality is due to Lipschitz Hessian and strong convexity. This yields

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x})-f^*|\boldsymbol{x}_B] \le 2C\cdot\frac{\rho^{\frac{2}{3}}}{T^{\frac{2}{3}}}R(H)\left(1+\frac{||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{(T\rho^2)^{\frac{1}{3}}}\right)\left(1+\frac{\rho||\nabla f(\boldsymbol{x}_B)||_2}{M^2}\right)\\ + C\frac{d^{\frac{4}{3}}\rho^{\frac{2}{3}}||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^2}{T^{\frac{1}{6}}M^2}+\frac{8\rho||\nabla f(\boldsymbol{x}_{\mathrm{B}})||_2^3}{M^3}.\end{aligned} \tag{10}$$

Taking the expectation over x_B in (10) and applying Proposition 4.2, the gradient-dependent remainder terms are lower-order as $T \to \infty$. Hence,

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

which completes the proof of Theorem 3.1.

E Proof details of Theorem 3.2

Before proving the theorem, we compare the general bound with the isotropic case $\hat{H} = \lambda I_d$ as a consistency check. In this setting, the $\hat{H}^{-1}$-error reduces to the squared Euclidean error:

$$(\hat{g} - \nabla f(x))^\top \hat{H}^{-1} (\hat{g} - \nabla f(x)) = \lambda^{-1} \|\hat{g} - \nabla f(x)\|_2^2.$$

Thus, in this isotropic setting there is only a single spectral group, and the achievability bound in Theorem 3.2 reduces to $O\left(\frac{\rho^{\frac{2}{3}} d}{n^{\frac{2}{3}} \lambda}\right)$, which coincides with the squared-error bound for the spherical estimator analyzed in Yu et al. [2024] and formally implied by Theorem A.2. We prove Theorem 3.2 through a refined analysis of the proposed Spectrally Grouped Estimator, which extends the analysis beyond the isotropic setting and yields strictly improved order-wise bounds in non-isotropic regimes.

Specifically, note that the restriction of a function with Lipschitz Hessian to an affine subspace still satisfies the same Lipschitz Hessian condition. Therefore, the contribution of the error from entries in group G_i can be analyzed directly via the arguments we provided for Spherical Sampling in Appendix A.1. Therefore, by Theorem A.2, the $\hat{H}^{-1}$-error of the Spectrally Grouped Estimator is characterized as follows.

Let P_i denote the orthogonal projection onto the subspace spanned by the eigenvectors of $\hat{H}$ whose eigenvalues lie in group G_i. Let

$$g_i := P_i \nabla f(x), \qquad \hat{g}_i := \text{the estimator returned for group } G_i.$$

Then

$$\hat{g} = \sum_i \hat{g}_i, \qquad \nabla f(x) = \sum_i g_i.$$

Since the groups correspond to orthogonal eigenspaces of $\hat{H}$, and $\hat{H}^{-1}$ is diagonal under the same eigenbasis, we have

$$(\hat{g} - \nabla f(x))^\top \hat{H}^{-1} (\hat{g} - \nabla f(x)) = \sum_i (\hat{g}_i - g_i)^\top \hat{H}^{-1} (\hat{g}_i - g_i).$$

Moreover, restricting f to the eigensubspace associated with G_i preserves the Lipschitz Hessian condition. For each group G_i, every eigenvalue of $\hat{H}$ indexed by G_i lies in

$$[\hat{H}_{\min} 2^i, \hat{H}_{\min} 2^{i+1}),$$

and therefore

$$\sum_{k\in G_i} \hat{H}_k^{-1} \le \frac{d_i}{\hat{H}_{\min} 2^{i-1}}.$$

Applying Theorem A.2 to the restriction of f onto the eigensubspace corresponding to G_i, we obtain

$$\mathbb{E}\Big[(\hat{g}_i - g_i)^\top \hat{H}^{-1} (\hat{g}_i - g_i)\Big] \le C\rho^{2/3} \frac{d_i}{n_i^{2/3} \hat{H}_{\min} 2^{i-1}} \left(1 + \|g_i\|_2^2 (n\rho^2)^{-1/3}\right).$$

Summing over all groups yields

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

where

$$c_i := \frac{d_i}{\hat{H}_{\min} 2^i}.$$

Since g_i is the orthogonal projection of $\nabla f(x)$ onto the subspace corresponding to G_i, we have

$$\|g_i\|_2^2 \le \|\nabla f(x)\|_2^2.$$

Therefore,

$$\sum_i \frac{c_i}{n_i^{2/3}} \left(1 + \|g_i\|_2^2 (n\rho^2)^{-1/3}\right) \le \left(\sum_i \frac{c_i}{n_i^{2/3}}\right) \left(1 + \|\nabla f(x)\|_2^2 (n\rho^2)^{-1/3}\right).$$

Our theorem is proved by applying the choice of n_i in Algorithm 2, which optimizes the achieved $\hat{H}^{-1}$-error. Specifically, this is motivated by the observation that the contribution to the $\hat{H}^{-1}$-risk in the above inequality from each group G_i is characterized by the factor $\frac{c_i}{n_i^{2/3}}$. The choice of n_i in our algorithm is obtained by solving

$$\text{minimize} \qquad \sum_i \frac{c_i}{n_i^{2/3}} \qquad \text{subject to} \qquad \sum_i n_i \le \frac{n}{2},$$

where the linear constraint is due to the sample-budget constraint.

Using a Lagrange multiplier, the optimizer approximately satisfies

$$n_i \propto c_i^{3/5}.$$

Hence, choosing

$$n_i = \left\lceil \frac{(n-2d) c_i^{3/5}}{2 \sum_j c_j^{3/5}} \right\rceil$$

which ensures that the total number of samples used is no greater than n, gives

$$\sum_i \frac{c_i}{n_i^{2/3}} \le \frac{4^{2/3} \left(\sum_i c_i^{3/5}\right)^{5/3}}{n^{2/3}}$$

for $n \ge 4d$. Substituting this bound into the previous inequality proves the theorem.

F Proof details of Theorem 3.3

F.1 Lipschitz Hessian Condition for the Hard-Instance Construction

We first verify that every function in the hard-instance family constructed in Section 4.4.2 satisfies the ρ-Lipschitz Hessian condition. For convenience, define

$$G(u) := \int_0^u g(z)\, dz, \qquad \phi_A(\boldsymbol{a}) := \cos^2\left(\min\left\{\sqrt{\boldsymbol{a}^\intercal A \boldsymbol{a}}, \frac{\pi}{2}\right\}\right),$$

where A is positive semidefinite and $\|A\|_{\mathrm{F}}$ is upper bounded by a universal constant. For the purpose of our proof, we let $\|A\|_{\mathrm{F}} \le \sqrt{\frac{1}{2^{\frac{1}{4}}-1}}$. A key property we use is that the map

$$f_A(y, \boldsymbol{a}) := \phi_A(\boldsymbol{a})^3 G(y/\phi_A(\boldsymbol{a})),$$

with its value at $\phi_A(\boldsymbol{a}) = 0$ defined by zero, satisfies a Lipschitz Hessian condition with respect to vector $(y, \boldsymbol{a})$. More specifically, there exists a universal constant C_0 such that

$$\|\nabla^2 f_A(y, \boldsymbol{a}) - \nabla^2 f_A(y', \boldsymbol{a}')\|_{\mathrm{F}} \le C_0 \|(y, \boldsymbol{a}) - (y', \boldsymbol{a}')\|_2.$$

This follows from the fact that $\nabla^2 f_A$ is globally continuous and fails to be differentiable only at the interfaces induced by the piecewise definition of g. It therefore suffices to bound the third derivatives of the entries of $\nabla^2 f_A$ individually on each region where f is three times differentiable.

We apply the above fact to by observing that each individual base function $\tilde{f}_s$ can be written in a form of f_A. Particularly, for any parameters (j, ℓ), we have

$$\tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x)) = \frac{\rho r_j^3}{C} f_A\left(\frac{s_{j,\ell} x_{j,\ell}}{r_j}, \frac{\boldsymbol{u}}{r_j}\right),$$

where $\boldsymbol{u}$ is a vector with indices k, m satisfying $k > j$, $m \in \mathcal{I}_{(j,k),\ell}$ and entries given by $u_{k,m} = 2^{-\frac{k-j}{16}} x_{k,m}$, and A is the diagonal matrix with entries given by $2^{-\frac{k-j}{8}} \left(\frac{d_k}{d_j}\right)^{-\frac{1}{2}}$. Given this identity, the Lipschitz Hessian condition of f_A essentially stats that[1]

$$\begin{aligned}
&\sum_{k \geq j} \sum_{m \in \mathcal{I}_{(j,k),\ell}} \sum_{k' \geq j} \sum_{m' \in \mathcal{I}_{(j,k'),\ell}} 2^{\frac{k+k'-2j}{8}} \left(\partial_{k,m}\partial_{k',m'} \widetilde{f}_{s_{j,\ell}}(x) - \partial_{k,m}\partial_{k',m'} \widetilde{f}_{s_{j,\ell}}(x')\right)^2 \\
&\leq \frac{\rho^2 C_0^2}{C^2} \left[\sum_{k \geq j} 2^{-\frac{k-j}{8}} \sum_{m \in \mathcal{I}_{(j,k),\ell}} (x_{k,m} - x'_{k,m})^2\right]. \qquad (11)
\end{aligned}$$

We prove the needed smoothness condition by taking summation over the above inequality. Specifically, observe that

$$\begin{aligned}
\|\nabla^2 f(x) - \nabla^2 f(x')\|_{\mathrm{F}}^2 &= \left\| \sum_{j,\ell} \nabla^2 \tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x)) - \nabla^2 \tilde{f}_{s_{j,\ell}}\big(x'_{j,\ell};\, \phi_{j,\ell}(x')\big) \right\|_{\mathrm{F}}^2 \\
&= \sum_{k,m} \sum_{k',m'} \left(\sum_{j,\ell} \partial_{k,m}\partial_{k',m'} \tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x)) - \partial_{k,m}\partial_{k',m'} \tilde{f}_{s_{j,\ell}}\big(x'_{j,\ell};\, \phi_{j,\ell}(x')\big) \right)^2,
\end{aligned}$$

where the terms inside the summation are non-zero only when $j \leq k, k'$, $m \in \mathcal{I}_{(j,k),\ell}$, and $m' \in \mathcal{I}_{(j,k'),\ell}$. Consequently, for any fixed k, k', m, m' and each j, at most one value of ℓ contributes to a non-zero term, and this only occurs when $j \leq k, k'$. Applying the Cauchy-Schwarz inequality, each squared inner sum is therefore bounded by

$$\begin{aligned}
&\left(\sum_{j,\ell} 2^{\frac{k+k'-2j}{8}} \left(\partial_{k,m}\partial_{k',m'} \tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x)) - \partial_{k,m}\partial_{k',m'} \tilde{f}_{s_{j,\ell}}\big(x'_{j,\ell};\, \phi_{j,\ell}(x')\big)\right)^2 \right) \\
&\cdot \left(\sum_{j \leq \min\{k,k'\}} 2^{-\frac{k+k'-2j}{8}} \right),
\end{aligned}$$

where the scalar factor is upper bounded by $\frac{1}{1-2^{-\frac{1}{4}}}$. Therefore, by applying inequality (11) and Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\nabla^2 f(x) - \nabla^2 f(x')\|_{\mathrm{F}}^2 &\leq \frac{\rho^2 C_0^2}{\left(1 - 2^{-\frac{1}{4}}\right) C^2} \sum_{j,\ell} \left[\sum_{k \geq j} 2^{-\frac{k-j}{8}} \sum_{m \in \mathcal{I}_{(j,k),\ell}} (x_{k,m} - x'_{k,m})^2 \right] \\
&\leq \frac{\rho^2 C_0^2}{\left(1 - 2^{-\frac{1}{4}}\right) C^2} \sum_{k,m} \left[\left(\sum_{j \leq k} 2^{-\frac{k-j}{8}} \right) (x_{k,m} - x'_{k,m})^2 \right] \\
&\leq \frac{\rho^2 C_0^2}{\left(1 - 2^{-\frac{1}{4}}\right)\left(1 - 2^{-\frac{1}{8}}\right) C^2} \|x - x'\|^2.
\end{aligned}$$

Hence, we have proved that the ρ-Lipshitz Hessian condition holds uniformly for our hard instance function f defined for all $\tilde{H}$, T, and $s_{k,\ell}$, by choosing any universal constant $C \geq C_0 / \sqrt{\left(1 - 2^{-\frac{1}{4}}\right)\left(1 - 2^{-\frac{1}{8}}\right)}$.

[1] Note that by definition, $\mathcal{I}_{(j,j),\ell}$ is uniquely determined as $\{\ell\}$.

F.2 Existence of $\tilde{H}$ for Matching the Hessian Condition

We first prove the existence of matrix $\tilde{H}$ for the hard instance function f specified in (4) to be included in $\mathcal{F}_H$ for all choices of s parameters. Note that due to the boundedness of the second order-differential of each $\tilde{f}_s$, there exists a neighborhood $\mathcal{H}$ of H, such that for any sufficiently large T, the function f becomes strongly convex for any choice of $\tilde{H} \in \mathcal{H}$. Concretely, through a direct evaluation, the operator norm of the Hessian of $\tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x))$ at any $x \in \mathbb{R}^d$ is upper bounded by $A\rho r_j$ for a universal constant A. Therefore, as a rough estimate through triangle inequality, the Hessian eigenvalues of f are lower bounded by $\tilde{H}_{\min} - A\rho \sum_j r_j$, where $\tilde{H}_{\min}$ denotes the minimum eigenvalue of $\tilde{H}$. Recall that we assumed $H_0 > M$, we have that $\tilde{H}_{\min} - M \geq (H_0 - M)/2$ for any $\tilde{H}$ belonging to the hyperball (under the Frobenius norm) centered at H with radius $(H_0 - M)/2$, and we denote this neighbourhood as $\mathcal{H}$. Recall the construction of r_j, all functions f for all $\tilde{H}$ within this neighbourhood satisfies strong convexity when $T \geq 2^6 A^6 \rho^4 \left(\sum_j d_j^{-\frac{1}{10}} H_j^{\frac{1}{10}}\right)^6 \left(\sum_j (d_j/H_j)^{\frac{3}{5}}\right)/(H_0 - M)^6$. Under this condition, the global minimizer of f is unique, and we denote it by $x^*_{\tilde{H}}$.

We consider the function $\Delta(\tilde{H}) = H - \nabla^2 f(x^*_{\tilde{H}})$, which captures the Hessian error towards the targeted condition at the global minimizer. We show that the map $U(\tilde{H}) := \tilde{H} + \Delta(\tilde{H})$ is a contractive self-map on the compact closure of the above neighborhood $\mathcal{H}$, therefore the unique fixed point principle can be applied.

First, note that the gradient norm satisfies $||\nabla f(0)||_2 = \frac{\rho}{C}\sqrt{\sum_j d_j r_j^4}$. Hence, by strong convexity, the norm of the global minimizer $||x^*_{\tilde{H}}||_2$ is upper bounded by $\frac{\rho}{CM}\sqrt{\sum_j d_j r_j^4}$. Recall the definition of r_j and the Lipschitz Hessian condition. This implies that $\tilde{H} + \Delta(\tilde{H})$ is uniformly bounded within a hypersphere (under the Frobenius norm) centered at H with radius $\frac{\rho^2}{CM}\sqrt{\sum_j d_j r_j^4}$, which converges to 0 as $T \to \infty$. Therefore, by the definition of neighborhood $\mathcal{H}$, the constructed map U is a selfmap on its closure when we further have $T \geq 8\rho^4 \left(\sum_j d_j^{\frac{3}{5}} H_j^{\frac{2}{5}}\right)^{\frac{3}{2}} \left(\sum_j (d_j/H_j)^{\frac{3}{5}}\right) / \left(C^3 M^3 (H_0 - M)^3\right)$.

Next, consider any pair of matrices $\tilde{H}_1, \tilde{H}_2 \in \mathcal{H}$. Through a direct evaluation, the gradient norm of the function f defined by $\tilde{H} = \tilde{H}_2$ at the global minimizer $x^*_{\tilde{H}_1}$ is bounded by $\left\|\left(\tilde{H}_1 - \tilde{H}_2\right) x^*_{\tilde{H}_1}\right\|_2 \leq \left\|\tilde{H}_1 - \tilde{H}_2\right\|_{\mathrm{F}} \cdot \left\|x^*_{\tilde{H}_1}\right\|_2$. Therefore, by strong convexity, the distance between the corresponding minimizers satisfies $\left\|x^*_{\tilde{H}_1} - x^*_{\tilde{H}_2}\right\|_2 \leq \frac{1}{M}\left\|\tilde{H}_1 - \tilde{H}_2\right\|_{\mathrm{F}} \cdot \left\|x^*_{\tilde{H}_1}\right\|_2$. Then, by Lipshitz Hessian, we have that $\left\|U(\tilde{H}_1) - U(\tilde{H}_2)\right\|_{\mathrm{F}} \leq \frac{\rho}{M}\left\|\tilde{H}_1 - \tilde{H}_2\right\|_{\mathrm{F}} \cdot \left\|x^*_{\tilde{H}_1}\right\|_2$. Since the multiplicative factor $\frac{\rho}{M}\left\|x^*_{\tilde{H}_1}\right\|_2$ converges uniformly to zero as $T \to \infty$ as we have proved earlier, the constructed map U is contractive for sufficiently large T. Specifically, U is a contractive selfmap over the compact $\mathcal{H}$ when T further satisfies $T \geq 2\rho^4 \left(\sum_j d_j^{\frac{3}{5}} H_j^{\frac{2}{5}}\right)^{\frac{3}{2}} \left(\sum_j (d_j/H_j)^{\frac{3}{5}}\right) / \left(C^3 M^6\right)$. As a side consequence, the above analysis also implies the existence of a threshold $T_0(\rho, M, R, H)$ such that for all $T \geq T_0$, we have $\left\|x^*_{\tilde{H}}\right\|_2 \leq R$ for all $\tilde{H} \in \mathcal{H}$, which guarantees the bounded distance to optimum assumption.

To summarize, given the properties established above, when T is lower bounded by a threshold that only depends on ρ, M, R, and H, Edelstein's Fixed Point Theorem (Edelstein [1962]) implies the existence of a unique fixed point $\tilde{H}$ such that $U(\tilde{H}) = \tilde{H}$. This is equivalent to $\Delta(\tilde{H}) = 0$, which implies that the resulting function f belongs to $\mathcal{F}_H$.

F.3 Quadratic approximation

We first present a general quadratic approximation bound. For brevity, let $\boldsymbol{s}$ denote the collection of all such indicators. Then for any joint distribution over $\boldsymbol{s}$, we prove the following inequality for

sufficiently large T.

$$E_{\boldsymbol{s}}[f(x)] - f^* \geq \sum_{j,\ell} \frac{\rho^2 r_j^4}{C^2 H_j} \left(2\,\mathbb{P}[s_{j,\ell} = 0]\mathbb{P}[s_{j,\ell} = 1] - \frac{1}{16} \right), \tag{12}$$

where f^* denotes the common global minimum of f for any given $\boldsymbol{s}$. By viewing the above joint distribution of $\boldsymbol{s}$ as a posterior distribution of $\boldsymbol{s}$ in a Bayes setting after any given sampling points and results in the first $T-1$ iterations, the above equations states that the expected simple regret can be characterized as a linear combination of terms that each only depends on the individual estimation accuracies of each $s_{j,\ell}$ entry.

Proof of inequality (12). We prove the inequality for the regime where $f \in \mathcal{F}_H$, which holds for any T greater than the thresholds specified in Appendix F.2. First, note that $E_{\boldsymbol{s}}[f(x)] - f^*$ is a strongly convex function whose unique global minimum point lies within the convex hull of the global minimum points of individual instances of f. We denote its global minimum point by $x^*_{P_{\boldsymbol{s}}}$, and $\|x^*_{P_{\boldsymbol{s}}}\|$ upper bounded by the L2 norm of the global minimum point of any function f within this hard instance class, which is in turn upper bounded by $r(T) := \frac{\rho}{CM}\sqrt{\sum_j d_j r_j^4}$, as we proved in Appendix F.2.

Consider the Lipschitz Hessian Condition and the local Hessian condition induced by $\mathcal{F}_H$,

$$\left| f(x) - x^\intercal \nabla f(0) - \frac{1}{2} x^\intercal H x \right| \leq \frac{\rho}{6}\|x\|_2^3 + \frac{\rho}{2}\|x^*\|_2^3.$$

We have

$$f^* \leq f(-H^{-1}\nabla f(0)) \leq -\frac{1}{2}\sum_j \frac{\rho^2 d_j r_j^4}{C^2 H_j} + \frac{2\rho}{3}|r(T)|^3,$$

where we applied a direct evaluation of $\nabla f(0)$, and that $\|H^{-1}\nabla f(0)\|_2 \leq \|\nabla f(0)\|_2 / M$. Furthermore,

$$\begin{aligned}
E_{\boldsymbol{s}}[f(x)] &\geq x^\intercal E_{\boldsymbol{s}}[\nabla f(0)] - \frac{1}{2} x^\intercal H x - \frac{\rho}{6}\|x\|_2^3 - \frac{\rho}{2}\|x^*\|_2^3 \\
&\geq -\frac{1}{2} E_{\boldsymbol{s}}[\nabla f(0)]^\intercal H^{-1} E_{\boldsymbol{s}}[\nabla f(0)] - \frac{\rho}{6}\|x\|_2^3 - \frac{\rho}{2}\|x^*\|_2^3 \\
&\geq -\frac{1}{2}\sum_{j,\ell} \frac{\rho^2 r_j^4}{C^2 H_j} (\mathbb{P}[s_{j,\ell} = 0] - \mathbb{P}[s_{j,\ell} = 1])^2 - \frac{\rho}{6}\|x\|_2^3 - \frac{\rho}{2}|r(T)|^3.
\end{aligned}$$

Therefore, the LHS of inequality (12), which is lower bounded by its evaluation at $x^*_{P_{\boldsymbol{s}}}$, is characterized by the following inequality.

$$E_{\boldsymbol{s}}[f(x)] - f^* \geq 2\sum_{j,\ell} \frac{\rho^2 r_j^4}{C^2 H_j} \mathbb{P}[s_{j,\ell} = 0]\mathbb{P}[s_{j,\ell} = 1] - \frac{4\rho}{3}|r(T)|^3,$$

Note that when T is sufficiently large, explicitly, $T \geq \frac{64^3 \rho^4 \left(\sum_j d_j^{3/5} H_j^{2/5}\right)^{9/2}}{27 C^3 M^9 \left(\sum_j (d_j/H_j)^{3/5}\right)^2}$, the lower order term $\frac{4\rho}{3}|r(T)|^3$ is dominated by the high-order coefficient $\sum_{j,\ell} \frac{\rho^2 r_j^4}{C^2 H_j}$. Hence, the needed inequality is implied. □

Finally, the needed quadratic characterization provided in inequality (5) is obtained by taking expectation over inequality (12) over posterior distributions.

F.4 Proof of dual condition

We prove the inequality for a general class of grouped spectrum, where $H_0, H_1 ...$, are arbitrary positive values satisfying $H_j \geq 2H_{j-1}$. We show, by induction over the number of spectral groups, denoted by D, that the inequality holds for when $C \geq \pi$ when the dimensionality satisfies

$d_j/d_{j-1} \geq 2^{-5/6}(H_j/H_{j-1})$. The proof follows from a piecewise-linear upper bound on the base perturbation

$$\left|\widetilde{f}_s(y; y_0)\right| \leq \frac{\rho}{C} \min\left\{y_0^2|y|, \frac{\pi}{2}y_0^3\right\}.$$

When $D = 1$, the above inequality bounds all individual terms as follows.

$$\left|\widetilde{f}_{s_{0,\ell}}\left(x_{0,\ell}; \phi_{0,\ell}(x)\right)\right|^2 \leq \frac{\pi^2}{4C^2 t_0}.$$

Then the needed statement follows by direct evaluation. Now assume the inequality holds for $D = D_0$, we provide a proof for $D = D_0 + 1$.

First, note that both sides of the desired inequality can be split into partial sums indexed by ℓ_0, where each partial sum collects all pairs j, ℓ satisfying $\ell \in \mathcal{I}_{(0,j),\ell_0}$. Therefore, it suffices to prove the inequality separately for each corresponding pair of partial sums. Within each pair associated with a fixed ℓ_0, we prove the inequality by considering two possible regimes.

Suppose that $\phi_{0,\ell_0}(x) = 0$. The $(0, \ell_0)$-term vanishes, and the corresponding partial sum contains only indices with $j > 0$. Applying the induction hypothesis to these remaining D_0 groups with effective sample budget $T - t_0$, we directly obtain

$$T\sum_{j>0}\sum_{\ell\in\mathcal{I}_{(0,j),\ell_0}} \frac{\rho^2 r_j^4}{H_j}\left(\widetilde{f}_{s_{j,\ell}}(x_{j,\ell}; \phi_{j,\ell}(x))\right)^2 \leq \frac{T}{T - t_0}\sum_{j>0}\sum_{\ell\in\mathcal{I}_{(0,j),\ell_0}} \frac{\rho^2 r_j^4}{2H_j},$$

where the RHS exactly equals the corresponding partial sum required in the induction step, by the definitions of t_j and r_j.

Otherwise, we have $\phi_{0,\ell_0}(x) > 0$. In this regime,

$$\left|\widetilde{f}_{s_{0,\ell_0}}\left(x_{0,\ell_0}; \phi_{0,\ell_0}(x)\right)\right|^2 \leq \frac{\pi^2}{2C^2 t_0}\left(1 - \frac{4}{\pi^2 r_0^2}\sum_{j>0}\sum_{\ell\in\mathcal{I}_{(0,j),\ell_0}} \frac{x_{j,\ell}^2}{2^{j/4}(d_j/d_0)^{1/2}}\right),$$

and for any $j > 0$,

$$\left|\widetilde{f}_{s_{j,\ell}}\left(x_{j,\ell}; \phi_{j,\ell}(x)\right)\right|^2 \leq \frac{x_{j,\ell}^2}{C^2 t_j r_j^2}.$$

Therefore, for each $j > 0$, the corresponding terms on the left-hand side of the dual inequality are bounded by quadratic expressions with a positive coefficients. By our initial bounded assumption on d_j/d_{j-1}, this positive contribution is absorbed by the corresponding negative quadratic coefficient in the $(0, \ell_0)$-term. Consequently, the overall contribution to the left-hand side of the dual inequality from the fixed block indexed by ℓ_0 is upper bounded by $T\frac{\rho^2 r_0^4}{H_0}\frac{\pi^2}{2C^2 t_0}$. Similar to first case, the desired inequality is then proved by direct evaluation.

F.5 Proof of the Theorem

To provide the proof for the general case, we construct the hard instances in a similar approach, except for setting some of the global minimum entries be zero to effectively enforce a divisibility condition. Specifically, for any symmetric $H \succeq MI$, recall that d_k denote the number of its eigenvalue in the interval $[M \cdot 2^k, M \cdot 2^{k+1})$, counting multiplicities. We first prove the existence of a universal constant $c > 0$, such that

$$\liminf_{T\to\infty} \mathcal{R}(T; H) \cdot T^{\frac{2}{3}} \geq cR'(H) \triangleq c\rho^{\frac{2}{3}}\left(\sum_{k=1}^{\lfloor \log_2(H_{\max}/M)\rfloor}\left(\frac{d_k}{M \cdot 2^k}\right)^{\frac{3}{5}}\right)^{\frac{5}{3}}. \tag{13}$$

A simple analysis can be obtained by constructing hard instances leveraging the designs in the divisible case considered earlier. Specifically, let $\bar{H}$ denote the symmetric matrix sharing the same eigenbasis of H, obtained by replacing each eigenvalue each eigenvalue H_i with $M\,2^{\lfloor \log_2(H/M)\rfloor+1}$, and let the entries of x under this eigenbasis associated with the eigenvalue $M\,2^{k+1}$ of $\bar{H}$ be denoted

by $x_{k,0}, ..., x_{k,d_k-1}$. We consider the choice of $\tilde{H}$ such that the following function is in $\mathcal{F}_{\overline{H}}$, where $\tilde{f}_s$ and $\phi_{j,\ell}$ are defined in Section 4.4.2 bases on $\bar{H}$.

$$\bar{f}(x) = \frac{1}{2}x^\intercal \tilde{H} x + \sum_{j=1}^{\lfloor \log_2(\frac{H_{\max}}{M}) \rfloor} \sum_{\ell=0}^{d_j-1} \tilde{f}_{s_{j,\ell}}(x_{j,\ell};\, \phi_{j,\ell}(x))\,. \tag{14}$$

As proved in Appendix F.2, such choice of $\tilde{H}$ exists for sufficiently large T. Furthermore, when T is lower bounded by a threshold that only depends on ρ, M, R, and H, we can choose $\tilde{H}$ be a matrix sharing the same eigenbasis of H, such that Assumption 2.2 is strengthened to $2M$-strong convexity, and Assumption 2.3 is strengthened to $\|x^*\| \le R/2$. Upon such instances, an entry-wise scaling of each $x_{j,\ell}$ by a factor of $\left(\bar{H}_j/H_j\right)^{0.5}$ creates hard instances in $\mathcal{F}_H$ while maintaining the sample complexity lower bounds, which exactly recovers inequality (13).

Compared to the desired lower bound, a hard instance class can be constructed for the missing term in a similar approach, except that the 1D instance ensures simultaneous minimization of the function value and the second order derivative.

G Optimality of SGE as a Gradient Estimator

In this appendix, we formally establish the universal optimality of the proposed SGE.

Corollary G.1. *Fix any positive definite matrix $\hat{H}$ and any query point x_B. There exists a universal constant $c > 0$ such that, for any gradient estimator $\hat{g}$ constructed using n (possibly adaptively) noisy oracle queries with i.i.d standard Gaussian noise, the worst case $\hat{H}^{-1}$-risk over all objective functions f satisfying Assumption 2.1- Assumption 2.3 is lower bounded by:*

$$\sup_f \mathbb{E}\left[(\hat{g} - \nabla f(x_B))^\top \hat{H}^{-1}(\hat{g} - \nabla f(x_B))\right] \ge c \cdot \frac{\rho^{\frac{2}{3}}}{n^{\frac{2}{3}}} \left(\sum_{k=0}^{\lfloor \log_2(\hat{H}_{\max}/\hat{H}_{\min}) \rfloor} \left(\frac{\hat{d}_k}{\hat{H}_{\min} \cdot 2^k} \right)^{\frac{3}{5}} \right)^{\frac{5}{3}}$$

for sufficiently large n, where $\hat{H}_{\max}$, $\hat{H}_{\min}$ are the largest and smallest eigenvalues of $\hat{H}$, and $\hat{d}_k$ is the number of eigenvalues in the interval $[\hat{H}_{\min} \cdot 2^k, \hat{H}_{\min} \cdot 2^{k+1})$.

Proof. We first consider the special case $x_B = 0$ and $\widehat{H} \succ MI$. For any such $\widehat{H}$ and all sufficiently large n, the hard-instance construction used in the proof of Theorem 3.3 yields a subclass of $\mathcal{F}_{\widehat{H}}$ such that, for the output x of any adaptive algorithm using n noisy oracle queries,

$$\mathbb{E}[f(x) - f^\star] \ge c\, R(\widehat{H}) n^{-2/3},$$

where $c > 0$ is a universal constant and the expectation is taken over the stochastic oracle noise and the uniform prior over the hard instances.

Now fix any gradient estimator $\hat{g}$ constructed from these n oracle queries, and use it in the clipped Newton refinement step of Algorithm 1 with $x_B = 0$ and Hessian matrix $\hat{H}$. Let

$$S_{\hat{H}} := \mathbb{E}\Big[(\hat{g} - \nabla f(0))^\top \hat{H}^{-1}(\hat{g} - \nabla f(0))\Big]$$

denote its $\hat{H}^{-1}$-risk. Since $\hat{H} \succ MI$, Proposition 4.1 applies and, after taking expectation over the oracle randomness, gives

$$\begin{aligned}\mathbb{E}[f(x) - f^\star] \le S_{\hat{H}} \left(1 + \frac{\epsilon_H}{M}\right) + 2\frac{\epsilon_H \|\nabla f(0)\|_2^2}{M^2} \\ + \left(6 + \frac{\epsilon_H}{M}\right) \frac{\rho \|\nabla f(0)\|_2^3}{M^3},\end{aligned}$$

where

$$\epsilon_H := \left\| \hat{H} - \nabla^2 f(0) \right\|_{\mathrm{F}}.$$

Combining the preceding simple-regret lower bound with the upper bound from Proposition 4.1, we obtain

$$S_{\widehat{H}}\left(1+\frac{\epsilon_H}{M}\right)+2\frac{\epsilon_H\|\nabla f(0)\|_2^2}{M^2}+\left(6+\frac{\epsilon_H}{M}\right)\frac{\rho\|\nabla f(0)\|_2^3}{M^3} \geq c\,R(\widehat{H})n^{-2/3}. \tag{15}$$

For the hard-instance family above, Appendix F.2 gives

$$\|\nabla f(0)\|_2=\frac{\rho}{C}\left(\sum_j d_j r_j^4\right)^{1/2}, \qquad \|x^\star\|_2\leq\frac{\rho}{CM}\left(\sum_j d_j r_j^4\right)^{1/2}.$$

Together with the choice of radii in the hard-instance construction, these bounds imply, for fixed $\hat{H}$,

$$\|\nabla f(0)\|_2=O(n^{-1/3}), \qquad \|x^\star\|_2=O(n^{-1/3}).$$

Since $f\in\mathcal{F}_{\hat{H}}$, we have $\nabla^2 f(x^\star)=\hat{H}$. Therefore, by the Lipschitz Hessian condition,

$$\epsilon_H=\|\hat{H}-\nabla^2 f(0)\|_{\mathrm{F}}\leq\rho\|x^\star\|_2=O(n^{-1/3}).$$

Substituting these bounds into (15), the two remainder terms are $O(n^{-1})$, while $1+\epsilon_H/M=1+o(1)$. Hence,

$$S_{\hat{H}}(1+o(1))\geq cR(\hat{H})n^{-2/3}-O(n^{-1}).$$

Since $n^{-1}=o(n^{-2/3})$, for all sufficiently large n,

$$S_{\hat{H}}\geq c'R(\hat{H})n^{-2/3}$$

for a universal constant $c'>0$. This proves the desired lower bound for the special case $x_B=0$ and $\hat{H}\succ MI$.

We now remove the restrictions $x_B=0$ and $\hat{H}\succ MI$. For an arbitrary query point x_B, we apply the preceding hard-instance construction after translating its center from the origin to x_B. For a general positive-definite $\hat{H}$, we further add a fixed, known quadratic function to each translated hard instance so that the resulting objectives satisfy Assumption 2.1-Assumption 2.3.

Neither modification changes the underlying gradient-estimation difficulty. The translation only changes the coordinate system, while the added quadratic term is known and can therefore be subtracted from every oracle observation. Specifically, if $\bar{f}=f+q$ for a known quadratic function q, then

$$\hat{g}-\nabla\bar{f}(x_B)=\big(\hat{g}-\nabla q(x_B)\big)-\nabla f(x_B),$$

so an estimator of $\nabla\bar{f}(x_B)$ induces an estimator of $\nabla f(x_B)$ with the same $\hat{H}^{-1}$-estimation error. Hence the special-case lower bound above carries over to the translated and quadratically compensated hard-instance family. Therefore, for all sufficiently large n,

$$\sup_f\mathbb{E}\Big[(\hat{g}-\nabla f(x_B))^\top\hat{H}^{-1}(\hat{g}-\nabla f(x_B))\Big]\geq c'R(\hat{H})n^{-2/3},$$

which proves the claim for arbitrary x_B and $\hat{H}\succ 0$. □

The lower bound in Corollary G.1 establishes that the estimation guarantee in Theorem 3.2 is asymptotically optimal in the small-gradient regime, where $\|\nabla f(x_B)\|\to 0$. In this regime, the upper bound presented in the main text directly matches the sample-complexity lower bound derived in this appendix.

We further demonstrate that the lower bound in Corollary G.1 is attainable beyond the small-gradient regime. Our approach employs a rough entrywise gradient estimate to locally compensate for the objective's linear component. By applying the stochastic gradient estimator to this compensated function, which significantly reduces the effective gradient, we directly establish the tightness of the lower bound in the claimed order.

Algorithm 7 CompensatedGradEst

// Compensated two-stage gradient estimator;
Input: x_B, n_1, n_2, r_1, r_2 //x_B: query point; n_1, n_2: sample budgets; r_1, r_2: sampling radii.
Return $\hat{g}$ as an estimator of the target quantity at x_B
Obtain $\hat{g}^{(1)}$ by invoking Algorithm 5 at x using n_1 samples and radius r_1
Define the locally compensated function

$$\tilde{f}_x(u) := f(x+u) - f(x) - \langle \hat{g}^{(1)}, u \rangle$$

Obtain $\hat{g}^{(2)}$ by invoking Algorithm 2 at x on function $\tilde{f}$ using n_2 samples and radius r_2
// Note: Each query to $\tilde{f}(y)$ is implemented as $f(x+u) - f(x) - \langle \hat{g}^{(1)}, u \rangle$
Return $\hat{g} = \hat{g}^{(1)} + \hat{g}^{(2)}$

Theorem G.2 (Achievability via compensated gradient estimation)**.** *For any positive-definite matrix $\widehat{H}$, any query point x_{B}, and all sufficiently large n, there exists a choice of (n_1, n_2, r_1, r_2) with $n_1 + n_2 = n$ such that Algorithm 7 returns an estimator $\widehat{g}$ for which the following holds uniformly over all functions f satisfying the ρ-Lipschitz Hessian condition:*

$$\mathbb{E}\left[(\hat{g} - \nabla f(x_B))^\top \hat{H}^{-1} (\hat{g} - \nabla f(x_B))\right] \le C \cdot \frac{\rho^{\frac{2}{3}}}{n^{\frac{2}{3}}} \left(\sum_{k=0}^{\lfloor \log_2(\hat{H}_{\max}/\hat{H}_{\min}) \rfloor} \left(\frac{\hat{d}_k}{\hat{H}_{\min} \cdot 2^k} \right)^{\frac{3}{5}} \right)^{\frac{5}{3}}. \quad (16)$$

where $\hat{H}_{\max}$, $\hat{H}_{\min}$, and $\hat{d}_k$ are defined as in Corollary G.1. Consequently, the lower bound in Corollary G.1 is attainable up to a universal constant factor.

Proof. Fix $\widehat{H}$ and x_{B} as in the theorem. The proof follows the two-stage construction in Algorithm 7: an entry-wise estimate is first used to remove the linear component of the objective, after which Theorem 3.2 is applied to the compensated function.

Choose n_1 and n_2 as constant fractions of the total sample budget n, with $n_1 + n_2 = n$. For sufficiently large n, both sample budgets satisfy the requirements of Theorem A.6 and Theorem 3.2.

Let $\widehat{\boldsymbol{g}}^{(1)}$ denote the entry-wise gradient estimate at $\boldsymbol{x}_{\mathrm{B}}$ obtained using n_1 samples and the parameter choice in Theorem A.6. Applying Theorem A.6 with $\widehat{H} = I_d$ gives

$$\mathbb{E}\left[\left\| \widehat{\boldsymbol{g}}^{(1)} - \nabla f(\boldsymbol{x}_{\mathrm{B}}) \right\|_2^2 \right] \le C_1 \frac{\rho^{2/3} d^{5/3}}{n_1^{2/3}},$$

where $C_1 > 0$ is a universal constant.

Following Algorithm 7, define the compensated function

$$\widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\boldsymbol{u}) := f(\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{u}) - f(\boldsymbol{x}_{\mathrm{B}}) - \left\langle \widehat{\boldsymbol{g}}^{(1)}, \boldsymbol{u} \right\rangle, \qquad \boldsymbol{u} \in \mathbb{R}^d.$$

Then

$$\nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0}) = \nabla f(\boldsymbol{x}_{\mathrm{B}}) - \widehat{\boldsymbol{g}}^{(1)}.$$

Thus, the gradient-dependent term in Theorem 3.2 is controlled by the first-stage estimate:

$$\mathbb{E}\left[\left\| \nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0}) \right\|_2^2 \right] \le C_1 \frac{\rho^{2/3} d^{5/3}}{n_1^{2/3}}.$$

Moreover, for every realization of $\widehat{\boldsymbol{g}}^{(1)}$, subtracting a constant and a linear function does not change the Hessian. Hence, $\widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}$ satisfies the same ρ-Lipschitz Hessian condition as f.

The second-stage oracle evaluations for the compensated function can be implemented using the original noisy function-value oracle. Conditional on $\widehat{\boldsymbol{g}}^{(1)}$, define

$$\widetilde{y}(\boldsymbol{u}) := y(\boldsymbol{x}_{\mathrm{B}} + \boldsymbol{u}) - \left\langle \widehat{\boldsymbol{g}}^{(1)}, \boldsymbol{u} \right\rangle.$$

Then

$$\widetilde{y}(\boldsymbol{u}) = \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\boldsymbol{u}) + f(\boldsymbol{x}_{\mathrm{B}}) + w.$$

The additive constant $f(\boldsymbol{x}_{\mathrm{B}})$ cancels in the symmetric finite differences used by Algorithm 2. Moreover, using fresh second-stage observations and subtracting the known linear term leave the additive-noise assumptions unchanged.

Apply Algorithm 2 to $\widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}$ at $\boldsymbol{u} = \mathbf{0}$ using n_2 samples and the parameter choice guaranteed by Theorem 3.2. Let $\widehat{\boldsymbol{g}}^{(2)}$ denote the resulting estimate of $\nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0})$.

Conditioning on $\widehat{\boldsymbol{g}}^{(1)}$ and then taking expectation, Theorem 3.2 gives

$$\mathbb{E}\Big[\Big(\widehat{\boldsymbol{g}}^{(2)} - \nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0})\Big)^{\mathsf{T}} \widehat{H}^{-1} \Big(\widehat{\boldsymbol{g}}^{(2)} - \nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0})\Big)\Big]$$
$$\leq C_2 \frac{\rho^{2/3}}{n_2^{2/3}} \left(\sum_{k=0}^{\lfloor \log_2(\widehat{H}_{\max}/\widehat{H}_{\min}) \rfloor} \left(\frac{\widehat{d}_k}{\widehat{H}_{\min} 2^k} \right)^{3/5} \right)^{5/3} \left(1 + \mathbb{E}\left[\left\| \nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0}) \right\|_2^2 \right] \left(\frac{1}{n_2 \rho^2} \right)^{1/3} \right).$$

The final estimator returned by Algorithm 7 is

$$\widehat{\boldsymbol{g}} := \widehat{\boldsymbol{g}}^{(1)} + \widehat{\boldsymbol{g}}^{(2)}.$$

By the definition of $\widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}$,

$$\widehat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}}) = \widehat{\boldsymbol{g}}^{(2)} - \nabla \widetilde{f}_{\boldsymbol{x}_{\mathrm{B}}}(\mathbf{0}).$$

Combining the preceding bound with the first-stage estimate gives

$$\mathbb{E}\Big[(\widehat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}}))^{\mathsf{T}} \widehat{H}^{-1} (\widehat{\boldsymbol{g}} - \nabla f(\boldsymbol{x}_{\mathrm{B}}))\Big]$$
$$\leq C_2 \frac{\rho^{2/3}}{n_2^{2/3}} \left(\sum_{k=0}^{\lfloor \log_2(\widehat{H}_{\max}/\widehat{H}_{\min}) \rfloor} \left(\frac{\widehat{d}_k}{\widehat{H}_{\min} 2^k} \right)^{3/5} \right)^{5/3} \left(1 + C_1 \frac{d^{5/3}}{n_1^{2/3} n_2^{1/3}} \right).$$

Since n_1 and n_2 are constant fractions of n,

$$\frac{d^{5/3}}{n_1^{2/3} n_2^{1/3}} = O\left(\frac{d^{5/3}}{n} \right), \qquad \frac{1}{n_2^{2/3}} = O\left(\frac{1}{n^{2/3}} \right).$$

Thus, for sufficiently large n, the additional factor is bounded by a universal constant. Absorbing this factor and the numerical constants into a universal constant C yields the claimed bound.

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the three main contributions: the Spectrally Grouped Estimator (SGE), the tight analysis of convex-support baseline estimators establishing a shared performance bottleneck, and the matching converse bound proving the instance-optimality of SGE within the considered second-order smooth, strongly convex setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the current work in the conclusion, noting that the analysis is established within the strongly convex setting with a Lipschitz Hessian condition and focuses on asymptotic rates, and identifying extensions to broader function classes, sharper non-asymptotic guarantees, and systematic finite-sample evaluations as promising directions for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are stated in Section 2 and further specified in the theorem statements when needed. Section 4 provides proof sketches and the main proof ideas for the three main results, while complete proofs and supporting lemmas are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not include empirical experiments. The figures are used only to illustrate the sampling geometries and representative theoretical scalings, and they do not constitute experimental results in the empirical sense.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This paper is purely theoretical and does not include experiments requiring datasets or code. The main results are theoretical and are reproduced through the stated assumptions, algorithms, and proofs.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper does not include empirical experiments. The figures are used only to illustrate the sampling geometries and representative theoretical scalings, and do not involve experimental training or test settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This paper does not include empirical experiments and therefore does not report error bars or other measures of statistical significance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper does not include empirical experiments and therefore does not require reporting compute resources for experimental reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work conforms with the NeurIPS Code of Ethics. The research is purely theoretical and does not involve human subjects, personal data, sensitive datasets, or deployed systems; we are not aware of any foreseeable ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper presents purely theoretical results and is not tied to specific deployments or applications that would raise immediate societal concerns.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks, as it does not release any models or datasets that require safeguards for responsible release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: This paper does not use any existing assets such as datasets, code, or pretrained models.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper does not release any new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve human subjects and therefore does not require IRB approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing and editing assistance, and did not contribute to the core methodology, scientific content, or originality of this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.