REVISITING THE DIMENSION DEPENDENCE OF ZEROTH-ORDER OPTIMIZATION UNDER HESSIAN TRACE-NORM CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order (ZO) optimization is broadly applied to high-dimensional problems where gradients are unavailable or costly, yet classical theory predicts rates that scale with the ambient dimension. Recent work replaces the dimension with a spectral quantity of the Hessian but assumes a fixed Hessian or a curvature envelope on a neighborhood or on the whole space. We show that, for the standard two-point Gaussian estimator, a pointwise Hessian trace-norm condition suffices: the sum of the absolute Hessian eigenvalues is bounded by a constant H at every point. Under this condition, we prove convergence to first-order stationary points at rate $\mathcal{O}(H/T)$ for general non-convex objectives and, under the Polyak–Łojasiewicz (PL) condition with constant μ, linear convergence with contraction factor $1 - \Theta(\mu/H)$. A "Twisted Valley" function whose principal curvature direction sweeps through every coordinate direction satisfies the condition with constants independent of the dimension, whereas the envelope conditions of prior analyses hold on it only with a trace that grows linearly in the dimension. Numerically, on this function and on the forward-only adaptation of a CNN and a Vision Transformer, the loss reached by the Gaussian estimator within a fixed query budget is insensitive to the dimension, while uniform coordinate probing degrades with it.

1 INTRODUCTION

Zeroth-order (ZO) optimization, which relies solely on function values to estimate descent directions, is widely used in machine learning. Its applications include black-box adversarial attacks on deep neural networks (Chen et al., 2017; Ye et al., 2025), reinforcement learning (Salimans et al., 2017), and hyperparameter tuning (Snoek et al., 2012). More recently, ZO methods such as MeZO (Malladi et al., 2023) have been used to fine-tune large language models with forward passes only, which reduces the memory footprint of optimizing billion-parameter models.

Despite this practical success, the theoretical understanding of ZO optimization in high dimensions remains incomplete. Classical bounds scale linearly with the ambient dimension d, giving an $\mathcal{O}(d)$ query complexity (Nesterov & Spokoiny, 2017; Duchi et al., 2015). Taken literally, this dependence would make ZO optimization of modern models prohibitively slow, a prediction that the efficiency observed in practice suggests is too pessimistic.

To resolve this discrepancy, recent analyses have introduced notions of effective dimension or weak dimension dependence. Yue et al. (2023) showed that for convex quadratics the convergence rate is governed by the eigenvalues of the Hessian rather than by d, and the analyses of MeZO (Malladi et al., 2023) and DPZero (Zhang et al., 2024) replace d by the effective rank of a Hessian envelope. These analyses, however, rely on convexity or on a single matrix that dominates the Hessian everywhere or on a neighborhood of each iterate, a curvature envelope. This leaves open the non-convex case in which the curvature is concentrated on few directions at every point but these directions change along the optimization trajectory.

A natural question is whether this reliance on an envelope stems from a property of the algorithm or from its analysis. The two-point Gaussian estimator draws isotropic directions and refers to no basis, so nothing in its update favors aligned curvature directions. An envelope, in contrast, cannot

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

follow a change of the curvature directions: when the principal curvature direction changes as the parameters move, it must cover every direction met in the neighborhood, so its trace grows with their number. This suggests that the envelope is needed by the analysis rather than by the algorithm, and we ask whether the Gaussian estimator can be analyzed without an envelope.

We answer this question affirmatively for the standard two-point Gaussian estimator (ZO-Gauss): with exact or noisy function values, its descent is governed by a pointwise bound H on the sum of the absolute Hessian eigenvalues, that is, on the trace norm of the Hessian, rather than by the ambient dimension. For general non-convex objectives, we establish convergence to first-order stationary points at rate $\mathcal{O}(H/T)$, and when the objective additionally satisfies the Polyak–Łojasiewicz (PL) condition, the same descent inequality yields linear convergence with contraction factor $1-\Theta(\mu/H)$. Since the trace-norm bound H is imposed at each point separately, it requires no envelope and lets the dominant curvature directions change freely along the trajectory.

Our contributions are as follows.

- We prove that, under a pointwise bound H on the sum of the absolute Hessian eigenvalues, the standard two-point Gaussian estimator with exact or noisy function values converges to stationary points and, under the PL condition, linearly, with an iteration complexity that depends on the dimension only through H.
- We introduce the Twisted Valley, a spatially varying and indefinite PL objective that satisfies our condition with constants independent of the dimension while its principal curvature direction rotates through all coordinate directions, and on which the envelope conditions of prior analyses hold only with a trace that grows linearly in d.
- We illustrate the theory numerically: on the Twisted Valley, ZO-Gauss tracks the rotating valley at a rate independent of d up to $d = 50000$, and in the forward-only adaptation of a CNN and a small Vision Transformer, the Hessian has a trace norm two to three orders of magnitude below d times its operator norm.

2 Related Work

Random subspace and projection methods. Random subspace and projection methods (Berglund et al., 2022; Cartis & Roberts, 2023; Kozak et al., 2023; Roberts & Royer, 2023; Nozawa et al., 2025) restrict each step to a low-dimensional subspace and require choosing a subspace distribution. Rates in the subspace dimension typically rely on convexity, local assumptions, or model-based conditions. Ma & Huang (2025) instead optimize the perturbation distribution for minimum asymptotic variance. The isotropic Gaussian estimator studied here selects no subspace, and our question is how the curvature of the objective function enters its rate, including when the active subspace rotates along the optimization path.

Curvature conditions in zeroth-order analyses. The spectral quantity behind our condition is not new: the effective dimension ED_1 of Yue et al. (2023), the supremum of the trace norm of the Hessian over x, is the constant H of our condition. They showed for convex quadratics that the two-point Gaussian method is governed by ED_1 rather than by d, and accelerated it with a heavy-ball scheme. For non-quadratic functions, they additionally assumed a Lipschitz continuous Hessian and used methods based on cubic regularization, whose complexity retains a term linear in d. The analyses of MeZO (Malladi et al., 2023) and DPZero (Zhang et al., 2024) bound the Hessian by a low-trace envelope, on a ball around the iterate in the former and on the whole space in the latter (Section 5). HiZOO (Zhao et al., 2025) and ZO-SAH (Kim et al., 2025) use diagonal or subspace Hessian estimates. Another line of work exploits sparsity of the gradient or of the objective in high dimensions (Wang et al., 2018; Cai et al., 2022; Balasubramanian & Ghadimi, 2022), or of the fine-tuned parameters of language models (Guo et al., 2025). Our condition involves no sparsity and is invariant under orthogonal changes of coordinates. We analyze the standard, unpreconditioned two-point method beyond quadratics under the pointwise condition alone, which requires no envelope, admits indefinite Hessians, and yields both non-convex stationarity and PL convergence.

PL geometry, implicit bias, and low-rank structure. The PL condition is a standard relaxation of strong convexity (Karimi et al., 2016). Zeroth-order rates under PL (Gasnikov et al., 2024) retain

Algorithm 1 Zeroth-Order Gradient Descent with Two-Point Gaussian Estimator (ZO-Gauss)

Require: Initial iterate $x_0 \in \mathbb{R}^d$, step sizes $\{h_t\}_{t=0}^{T-1}$, smoothing parameter $\nu > 0$, iteration limit T.
Ensure: Final iterate x_T.
1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Sample $u_t \sim \mathcal{N}(0, I_d)$
3: Query the values $y_t^{\pm}$ of f at $x_t \pm \nu u_t$
4: $x_{t+1} = x_t - h_t \frac{y_t^+ - y_t^-}{2\nu} u_t$
5: **end for**

explicit factors of d, whereas our contraction factor depends on H/μ alone. This is in line with rates that depend on a latent dimension (Golovin et al., 2020) and with evidence that ZO methods find flat minima (Zhang et al., 2025).

3 PRELIMINARIES

This section fixes the notation, introduces the algorithm, and states the curvature condition used in Section 4: a pointwise bound on the sum of the absolute eigenvalues of the Hessian.

Notation. For $x \in \mathbb{R}^d$, $\|x\|$ denotes the Euclidean norm, I_d denotes the d-dimensional identity matrix, and $e_1, \ldots, e_d$ denote the standard basis vectors of $\mathbb{R}^d$. For a symmetric matrix $B \in \mathbb{R}^{d \times d}$ with eigenvalues $\lambda_1(B), \ldots, \lambda_d(B)$, we write $\|B\|_{\mathrm{op}} \coloneqq \max_i |\lambda_i(B)|$ for the operator norm and $\|B\|_{\mathrm{F}}$ for the Frobenius norm. For symmetric A and B, $A \preceq B$ means that $B - A$ is positive semidefinite.

3.1 PROBLEM SETTING AND ALGORITHM

We consider the unconstrained minimization problem

$$\min_{x \in \mathbb{R}^d} f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, in the zeroth-order setting: the gradient $\nabla f(x)$ is inaccessible, and the optimizer can only query function values of f, possibly corrupted by noise.

Two-point Gaussian estimator. At iteration t, given the current iterate x_t, we sample a direction $u_t \sim \mathcal{N}(0, I_d)$, query the values $y_t^{\pm}$ of f at $x_t \pm \nu u_t$, and update

$$x_{t+1} = x_t - h_t \tilde{D}_t u_t, \qquad \tilde{D}_t \coloneqq \frac{y_t^+ - y_t^-}{2\nu}, \tag{1}$$

where $\nu > 0$ is the smoothing parameter (or finite-difference radius) and $h_t > 0$ is the step size. The resulting procedure, ZO-Gauss, is summarized in Algorithm 1. With exact function values, $y_t^{\pm} = f(x_t \pm \nu u_t)$ and $\tilde{D}_t$ is the central finite difference

$$D_\nu f(x; u) \coloneqq \frac{f(x + \nu u) - f(x - \nu u)}{2\nu} \tag{2}$$

at (x_t, u_t). Then $D_\nu f(x_t; u_t) u_t$ is the Gaussian-smoothing estimator of Nesterov & Spokoiny (2017), whose expectation over u_t is the gradient of the smoothed objective $x \mapsto \mathbb{E}_u[f(x + \nu u)]$.

3.2 ASSUMPTIONS

We now state the standing assumptions: the usual requirement that the problem be well posed, the curvature condition announced at the beginning of this section, and a condition on the noise in the function values.

Assumption 3.1 (Lower-boundedness)**.** f is bounded from below, that is, $f^* \coloneqq \inf_x f(x) > -\infty$.

Our main curvature condition bounds the spectrum of the Hessian one point at a time.

Assumption 3.2 (Pointwise spectral condition)**.** There exists a constant $H > 0$ such that the sum of the absolute eigenvalues of the Hessian satisfies

$$\|\nabla^2 f(x)\|_* := \sum_{i=1}^{d} \left|\lambda_i(\nabla^2 f(x))\right| \le H, \qquad \forall x \in \mathbb{R}^d. \tag{3}$$

The left-hand side of (3) is the trace norm of $\nabla^2 f(x)$, which satisfies $\|B\|_{\mathrm{op}} \le \|B\|_{\mathrm{F}} \le \|B\|_*$ and $|\mathrm{Tr}(B)| \le \|B\|_*$ for every symmetric B, with $\|B\|_* = \mathrm{Tr}(B)$ when $B \succeq 0$. Hence Assumption 3.2 implies global H-smoothness, $\|\nabla^2 f(x)\|_{\mathrm{op}} \le H$, whereas an L-smooth function satisfies (3) only with $H = dL$ in general. The constant H stays bounded independently of d whenever the curvature is concentrated on a few directions at every point. The condition is invariant under orthogonal changes of coordinates, so it singles out no basis. It is also imposed one point at a time, so it bounds the eigenvalues of $\nabla^2 f(x)$ without restricting how its eigenvectors, that is, the dominant curvature directions, rotate along the trajectory. The quantity $\sup_x \|\nabla^2 f(x)\|_*$ is the effective dimension ED_1 of Yue et al. (2023), so Assumption 3.2 states that $\mathrm{ED}_1 \le H$.

In most applications the values of f are available only up to noise, and we allow each of the two evaluations of the finite difference to be perturbed. The following condition describes this noise.

Assumption 3.3 (Noisy function values)**.** At iteration t, the two queried values are $y_t^{\pm} = f(x_t \pm \nu u_t) + \varepsilon_t^{\pm}$, so that $\tilde{D}_t = D_\nu f(x_t; u_t) + e_t/\nu$ in (1) with $e_t := \frac{1}{2}(\varepsilon_t^+ - \varepsilon_t^-)$. Let $\mathcal{F}_t$ denote the history before u_t is sampled. There is a constant $\sigma \ge 0$ such that, for every t,

$$\mathbb{E}[e_t \mid \mathcal{F}_t, u_t] = 0, \qquad \mathbb{E}\big[e_t^6 \mid \mathcal{F}_t\big] \le \sigma^6. \tag{4}$$

The first condition in (4) states that the error of the finite difference has mean zero. The second condition bounds its tail uniformly over the iterates. The proof needs a moment of order higher than four, and we fix the order six, the smallest even order above four, which keeps the constants explicit.

4 Convergence Analysis

In this section, we analyze Algorithm 1 under the setting of Section 3. Theorem 4.1 establishes convergence to first-order stationary points for general non-convex objectives, Theorem 4.3 establishes linear convergence under the PL condition, and Corollary 4.4 derives the resulting iteration complexities, in which the ambient dimension d enters only through the constant H of Assumption 3.2. Proofs are deferred to Appendix A.

Our first main result concerns convergence to first-order stationary points for general non-convex objectives.

Theorem 4.1 (Convergence to stationary points)**.** *Suppose Assumptions 3.1 to 3.3 hold, let $\nu > 0$ and $h_t = h \in (0, \frac{1}{100H}]$ for every t, and let $\{x_t\}_{t\ge 0}$ be generated by Algorithm 1. Then*

$$\min_{0\le t<T} \mathbb{E}\Big[\|\nabla f(x_t)\|^2\Big] \le \frac{2(f(x_0) - f^*)}{hT} + 86\nu^2 H^2 + 56Hh\frac{\sigma^2}{\nu^2}. \tag{5}$$

The first term is of order $\mathcal{O}(H/T)$ and equals $200H(f(x_0) - f^*)/T$ at the largest admissible step size. Classical analyses of the two-point estimator give $\mathcal{O}(dL/T)$ with $L := \sup_x \|\nabla^2 f(x)\|_{\mathrm{op}}$, that is, d times the rate $\mathcal{O}(L/T)$ of first-order gradient descent. Since $L \le H \le dL$, our bound replaces this factor d by the effective rank H/L of the Hessian. The second term is the finite-difference bias and the third the contribution of the noise; neither depends on d except through H. Since ν is an algorithmic parameter, the bias can be made smaller than any target accuracy ϵ with $\nu = \mathcal{O}(\sqrt{\epsilon}/H)$ without affecting the number of function evaluations, and the noise term is then made small by the step size.

We next show that the stationarity guarantee sharpens to linear convergence under the Polyak–Łojasiewicz condition, which we recall first.

Assumption 4.2 (Polyak–Łojasiewicz condition)**.** There exists $\mu > 0$ such that, for all $x \in \mathbb{R}^d$,

$$f(x) - f^* \le \frac{1}{2\mu}\|\nabla f(x)\|^2. \tag{6}$$

The PL condition does not require convexity and admits multiple global minimizers. In particular, it is compatible with singular and indefinite Hessians.

Theorem 4.3 (Linear convergence under the PL condition)**.** *Suppose Assumptions 3.1 to 3.3 and the PL condition (Assumption 4.2) hold, let $\nu > 0$ and $h_t = h \in (0, \frac{1}{100H}]$ for every t, and let $\{x_t\}_{t\geq 0}$ be generated by Algorithm 1. Then*

$$\mathbb{E}[f(x_t) - f^*] \leq (1-\mu h)^t (f(x_0) - f^*) + \frac{1}{\mu}\left(43\nu^2 H^2 + 28Hh\frac{\sigma^2}{\nu^2}\right). \tag{7}$$

Choosing the step size and the finite-difference radius in these bounds gives the following iteration complexities; each iteration uses two function evaluations.

Corollary 4.4 (Iteration complexity)**.** *Let $\Delta \coloneqq f(x_0) - f^*$ and $\epsilon > 0$.*

(i) Under the assumptions of Theorem 4.1, the choices $\nu^2 = \frac{\epsilon}{344H^2}$ and $h = \min\{\frac{1}{100H}, \frac{\epsilon^2}{8\cdot 10^4 H^3\sigma^2}\}$[1] give $\min_{0\leq t<T}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \epsilon$ for all $T \geq \frac{4\Delta}{h\epsilon}$, that is, after

$$T = \mathcal{O}\left(\frac{H\Delta}{\epsilon}\left(1 + \frac{H^2\sigma^2}{\epsilon^2}\right)\right)$$

iterations.

(ii) Under the assumptions of Theorem 4.3, the choices $\nu^2 = \frac{\mu\epsilon}{172H^2}$ and $h = \min\{\frac{1}{100H}, \frac{\mu^2\epsilon^2}{2\cdot 10^4 H^3\sigma^2}\}$ give $\mathbb{E}[f(x_T) - f^*] \leq \epsilon$ for all $T \geq \frac{1}{\mu h}\log\frac{2\Delta}{\epsilon}$, that is, after

$$T = \mathcal{O}\left(\frac{H}{\mu}\left(1 + \frac{H^2\sigma^2}{\mu^2\epsilon^2}\right)\log\frac{\Delta}{\epsilon}\right)$$

iterations.

With exact function values, $\sigma = 0$, the iteration counts reduce to $\mathcal{O}(H\Delta/\epsilon)$ and $\mathcal{O}(\frac{H}{\mu}\log\frac{\Delta}{\epsilon})$, those of gradient descent with the smoothness constant replaced by H. Noise adds the factor $1 + H^2\sigma^2/\epsilon^2$ or $1 + H^2\sigma^2/(\mu^2\epsilon^2)$, which comes from the smaller step size needed to control the noise term, while ν is chosen from ϵ and H alone.

Remark 4.5 (Dependence on the dimension)**.** The iteration counts in Corollary 4.4, and the choices of h and ν, depend on d only through H, whereas classical analyses of the two-point Gaussian estimator carry a factor of d (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). For a quadratic $\frac{1}{2}x^\top Ax$ with $A \succeq 0$, $H = \mathrm{Tr}(A)$ and $H/\mu \geq d$ whenever A has full rank, so no dimension-free behavior is claimed for full-dimensional curvature.

Remark 4.6 (Comparison with Yue et al. (2023))**.** The constant H of Assumption 3.2 was proposed by Yue et al. (2023) as the effective dimension ED_1. They analyze the two-point method only for quadratics, strongly or weakly convex, and state that the analysis extends to varying Hessians with a uniform upper bound but omit this extension. Our analysis requires no such bound. For non-quadratic convex objectives and for second-order stationary points of non-convex objectives, they use methods based on cubic regularization with a subroutine of d coordinate differences, whose complexities contain terms linear in d of lower order in ϵ.

In the rest of this section, we outline the proofs of Theorems 4.1 and 4.3. The full proofs are given in Appendix A.

Proof outline. Fix x_t, condition on $\mathcal{F}_t$, and write $g_t \coloneqq \nabla f(x_t)$, $q_t \coloneqq g_t^\top u_t$, $\delta_t \coloneqq D_\nu f(x_t; u_t) - q_t$, and $\Phi(z) \coloneqq f(x_t + z) - f(x_t) - g_t^\top z$, so that $\tilde{D}_t = q_t + \delta_t + e_t/\nu$ and the update gives

$$f(x_{t+1}) = f(x_t) - h\tilde{D}_t q_t + \Phi(-h\tilde{D}_t u_t).$$

Since $\mathbb{E}\left[\tilde{D}_t q_t\right] = \|g_t\|^2 + \mathbb{E}[\delta_t q_t]$ by the first condition in (4), it remains to bound the cross term $\mathbb{E}[\delta_t q_t]$ and the second-order term $\mathbb{E}\left[\left|\Phi(-h\tilde{D}_t u_t)\right|\right]$ by multiples of H. For a quadratic with Hessian

[1] Here and in (ii), if $\sigma = 0$, the second term in the minimum is read as $+\infty$, so that $h = \frac{1}{100H}$.

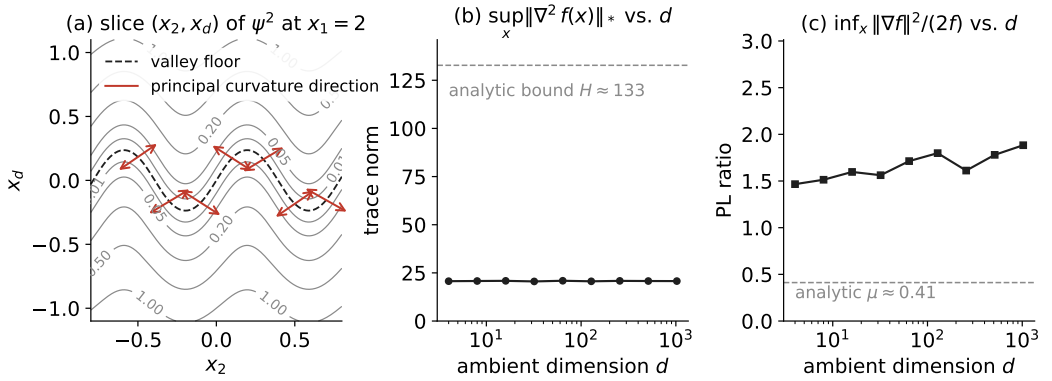

Figure 1: **The Twisted Valley** (8) **with** $(a, \omega, \kappa) = (1/4, 8, 2)$**.** (a) Slice (x_2, x_d) of the first term of (8) at $x_1 = 2$: the valley floor (dashed) and the principal curvature direction (red), which is transverse to the floor and changes along it. (b) Largest trace norm of $\nabla^2 f$ and (c) smallest PL ratio $\|\nabla f\|^2/(2f)$ over 3000 random points per dimension, together with the analytic constants of Proposition 5.1; neither depends on d.

A, both follow from the identity $\mathbb{E}\big[(g^\top u)^2 u^\top A u\big] = \|g\|^2 \operatorname{Tr}(A) + 2g^\top A g$. For general f, the Hessian enters at points that depend on u_t, and bounding it there by its operator norm would cost a factor of d. We avoid this in two ways.

For the cross term, the weight q_t is unchanged by the rotations that fix $\hat{g}_t = g_t/\|g_t\|$, so Φ may be replaced by its average over these rotations. The averaged function has $d-2$ equal Hessian eigenvalues, each at most $H/(d-2)$ by the trace-norm bound, which makes the bound dimension-free (Lemmas A.2 and A.3).

For the second-order term, at radius τ along the ray through u_t, the derivative of Φ is of order $H\tau$ in root mean square, although pointwise it can be of order $H\tau d$. This follows from an integrated Bochner identity for the Ornstein–Uhlenbeck operator (Bakry et al., 2014) (Lemma A.5). Integrating along the ray gives $\mathbb{E}[|\Phi(\theta u_t)|] = \mathcal{O}(H\|\theta\|_6^2)$ for any random radius θ (Lemma A.6), which we apply with $\theta = -h\tilde{D}_t$; the noise enters only through $\|e_t/\nu\|_6 \le \sigma/\nu$ (Proposition A.8).

5 Twisted Valley: A Canonical Rotating Landscape

This section gives an explicit objective in the class of Assumptions 3.2 and 4.2 whose principal curvature direction rotates through every coordinate direction as one moves along its valley, while the constants H and μ do not depend on d (Section 5.1). It then uses this objective to show that the envelope conditions of prior analyses pay for this rotation, whereas ours does not (Section 5.2).

5.1 Construction

The objective consists of a valley whose floor is bent by a sum of localized bumps, one for each coordinate direction. For $a, \omega, \kappa > 0$, define

$$f(x) := \left(x_d - \frac{\Lambda(x)}{1+x_d^2}\right)^2 + \frac{\kappa}{2}x_1^2, \qquad \Lambda(x) := a\sum_{k=2}^{d-1} \sin(\omega x_k)\left(1-(x_1-k)^2\right)_+^3, \tag{8}$$

where $(\cdot)_+ := \max\{\cdot, 0\}$. The factor $(1-(x_1-k)^2)_+^3$ is C^2 in x_1 and vanishes unless $|x_1 - k| < 1$, that is, outside the k-th *window*. Hence, for each value of x_1 at most two terms of Λ are nonzero, and f depends only on x_1, x_d, and the one or two coordinates x_k with $|x_1 - k| < 1$. The first term vanishes on the valley floor, the set where $x_d = \Lambda(x)/(1+x_d^2)$, and the second term decreases along the floor toward the minimizers, at which $x_1 = 0$ and $x_d = 0$. As the *clock coordinate* x_1 moves from $d-1$ to 2 along the floor, the active coordinate moves from x_{d-1} to x_2, and the direction transverse to the floor tilts from e_d toward the active coordinate direction (Figure 1).

Proposition 5.1. *Let $d \geq 4$, $\omega, \kappa > 0$, and $0 < a < \frac{4}{8+3\sqrt{3}}$, and let $H = H(a, \omega, \kappa)$ be the explicit constant* (22) *in Appendix B, which does not depend on d. The function* (8) *has the following properties.*

(i) $f \in C^2(\mathbb{R}^d)$, $f^* = 0$, the minimizers are the points with $x_1 = x_d = 0$, and the valley floor is the graph of a continuous function of $(x_1, \ldots, x_{d-1})$ with values in $[-2a, 2a]$.

(ii) Assumption 3.2 holds with H and Assumption 4.2 with $\mu = \min\{2\big(1 - \frac{3\sqrt{3}}{4}a\big)^2 - 8a^2,\ \frac{\kappa}{4}\}$.

In particular, for $(a, \omega, \kappa) = (\frac{1}{4}, 8, 2)$ we have $H < 133$, $\mu > 0.41$, and $H/\mu < 325$ for every $d \geq 4$.

Proposition 5.1 bounds the curvature transverse to the floor by constants independent of d, but the direction of this curvature is not fixed. At the floor points with $x_d = 0$, $x_1 = k \in \{2, \ldots, d-1\}$, and $x_j \in \frac{\pi}{\omega}\mathbb{Z}$ for all $j \geq 2$, the Hessian has rank two, with the eigenvector e_1 of eigenvalue κ and a principal eigenvector proportional to $e_d \mp a\omega\, e_k$ with eigenvalue $2(1 + a^2\omega^2)$, as a direct computation from (8) shows. The principal curvature direction is thus tilted from e_d by the angle $\arctan(a\omega)$ toward $e_2, e_3, \ldots, e_{d-1}$ in turn as x_1 increases, and no subspace of dimension smaller than $d-2$ contains all of these directions, whereas H and μ remain fixed. Moreover, $\operatorname{rank} \nabla^2 f \leq 4$ everywhere and $\nabla^2 f$ is indefinite at some points, so f is not convex. By Corollary 4.4(ii) with exact function values, $h = \frac{1}{100H}$, and $\nu^2 = \frac{\mu\epsilon}{172H^2}$, Algorithm 1 reaches $\mathbb{E}[f(x_T)] \leq \epsilon$ after $T = \mathcal{O}\Big(\frac{H}{\mu} \log \frac{f(x_0)}{\epsilon}\Big)$ iterations, and neither T nor ν depends on d.

5.2 Comparison with Envelope Conditions

The guarantee above uses only the pointwise constant H, so the rotation of the curvature costs nothing in our analysis. Analyses that go beyond quadratics and still replace d by a spectral quantity instead bound the Hessian on a region by a single matrix, a curvature envelope. Such a matrix cannot follow the rotation: it has to cover at once every curvature direction met in the region. We now make this precise.

DPZero (Zhang et al., 2024) and MeZO (Malladi et al., 2023) require a matrix $A \succeq 0$ that dominates the Hessian, $\nabla^2 f(y) \preceq A$, at every point y of a neighborhood, and measure the curvature by the trace or the effective rank of A in place of H. The neighborhood is the whole space for the global envelope of DPZero, and a ball of radius $\eta d G(x_t)$ around the iterate for the local effective-rank condition of MeZO, where $G(x_t) \geq \|\nabla f(x_t)\|$ is the largest per-example gradient norm. A two-sided envelope $-A \preceq \nabla^2 f(y) \preceq A$ implies $\|\nabla^2 f(y)\|_* \leq \operatorname{Tr}(A)$ (Lemma B.4), so Assumption 3.2 is implied by such conditions. The converse fails on the Twisted Valley: each window carries curvature $2a^2\omega^2$ along its own axis e_k, and an envelope must carry this amount on the diagonal entry of every window its neighborhood meets. The following proposition quantifies this cost for a ball of radius r.

Proposition 5.2. *Let $d \geq 4$, let f be the function* (8) *with the constant H of Proposition 5.1, let $x \in \mathbb{R}^d$ with $2 \leq x_1 \leq d-1$, and let $r > 0$.*

(i) If $r^2 > x_d^2 + \pi^2/(4\omega^2)$ and $A \succeq 0$ satisfies $\nabla^2 f(y) \preceq A$ for all y with $\|y - x\| \leq r$, then

$$\operatorname{Tr}(A) \geq 2a^2\omega^2 \min\{\lfloor \rho \rfloor,\, d-2\}, \qquad \rho := \sqrt{r^2 - x_d^2 - \pi^2/(4\omega^2)}. \tag{9}$$

(ii) There is $A \succeq 0$ with $-A \preceq \nabla^2 f(y) \preceq A$ for all y with $\|y - x\| \leq r$ and $\operatorname{Tr}(A) \leq H \min\{2r + 5,\, d\}$.

By Proposition 5.2, an envelope pays between $2a^2\omega^2$ and $2H$ for every window met by its neighborhood, so the size of the neighborhood decides what the rotation costs. A single matrix dominating the Hessian everywhere meets every window and has trace at least $2a^2\omega^2(d-2)$. The ball of MeZO has radius $\eta d G(x_t) \geq \eta d \|\nabla f(x_t)\|$, which grows linearly in d on the valley floor, so its envelope has trace of order d as well. The Twisted Valley thus satisfies Assumption 3.2 with a constant independent of d, but none of these envelope conditions with a trace independent of d.

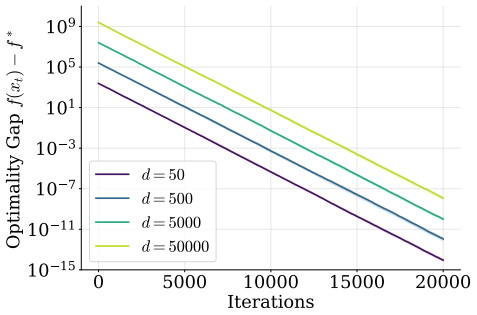

Figure 2: **ZO-Gauss on the Twisted Valley** (8) from $x_1 = d - 1$, crossing all windows, for $d \in \{50, 500, 5000, 50000\}$ ($h = 5 \times 10^{-4}$; lines are pointwise medians over 20 runs, bands the 10th to 90th percentiles).

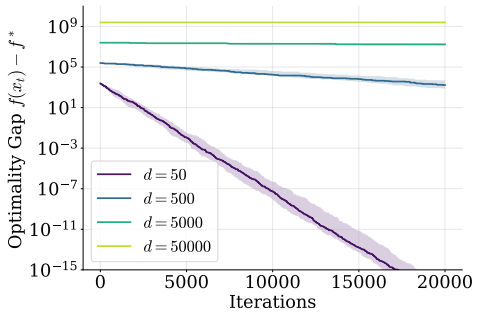

Figure 3: **ZO-Coord on the Twisted Valley** from $x_1 = d - 1$ with the step size $h = 0.03/d$ (medians and percentile bands as in Figure 2, same axes).

6 Numerical Experiments

In this section, we numerically illustrate the prediction of Theorems 4.1 and 4.3 that the query complexity of ZO-Gauss (Algorithm 1) is governed by the trace norm H rather than by the ambient dimension d. We do so on the Twisted Valley (8) for d up to 50000 (Section 6.1) and on the forward-only adaptation of a CNN and a small Vision Transformer whose trainable dimension is varied by a factor of thirteen (Section 6.2). As a baseline we use ZO-Coord, the zeroth-order counterpart of randomized coordinate descent (Nesterov, 2012; Richtárik & Takáč, 2014), which probes a single uniformly random coordinate axis e_i, $i \sim \mathrm{Unif}([d])$, per iteration and uses the estimator

$$g_{\mathrm{coord}}(x) = d\,\frac{f(x + \nu e_i) - f(x - \nu e_i)}{2\nu} e_i, \tag{10}$$

whose factor d makes its expectation over i the central-difference approximation of $\nabla f(x)$. The averaged Gaussian and sparse perturbation methods match ZO-Gauss on both problems (Appendices D and F), so the differences reported below stem from probing one axis at a time, not from the rotation of the curvature.

6.1 Experiments on the Twisted Valley

We minimize the Twisted Valley (8) with $(a, \omega, \kappa) = (\frac{1}{4}, 8, 2)$, the parameters certified in Proposition 5.1, $\nu = 10^{-2}$, and 20 runs per dimension. All runs start from $x_1 = d - 1$, $x_k = 0.1$ for $2 \le k \le d-1$, and $x_d = 1$, so that the quadratic term pulls the clock coordinate x_1 through all $d-2$ windows $k = d-1, \ldots, 2$, and the principal curvature direction changes at each of them. ZO-Gauss uses $h_t = 5 \times 10^{-4}$. ZO-Coord uses $h = 0.03/d$, so that its update along the sampled axis is 0.03 times the finite difference. Figures 2 and 3 report the optimality gap $f(x_t) - f^*$, and Appendix C the distance from the valley floor.

For ZO-Gauss the curves are parallel: the optimality gap decreases at a rate that does not depend on d. They are shifted upward only because the initial gap $f(x_0) \approx \frac{\kappa}{2}(d-1)^2$ grows with d, which enters through the factor $\log(f(x_0)/\epsilon)$ of Theorem 4.3. An envelope over the visited region, in contrast, meets every window, and its trace grows a thousandfold (Section 5.2). ZO-Coord moves x_1 only when the first axis is sampled, which happens with probability $1/d$, and it degrades as d grows, so that for $d \ge 5000$ the optimality gap stays close to its initial value.

6.2 Forward-Only Adaptation of a CNN and a Vision Transformer

We split CIFAR-10 into a source task (classes 0–4) and a target task (classes 5–9), pre-train a three-layer CNN and a small Vision Transformer (ViT) on the source task with first-order optimization, and freeze them. On the target task, we train a residual adapter of width $r \in \{2, 8, 16, 32, 64\}$ ($64 \to r \to 64$ with $\tanh$ activation, initialized as the identity) together with a new five-way head

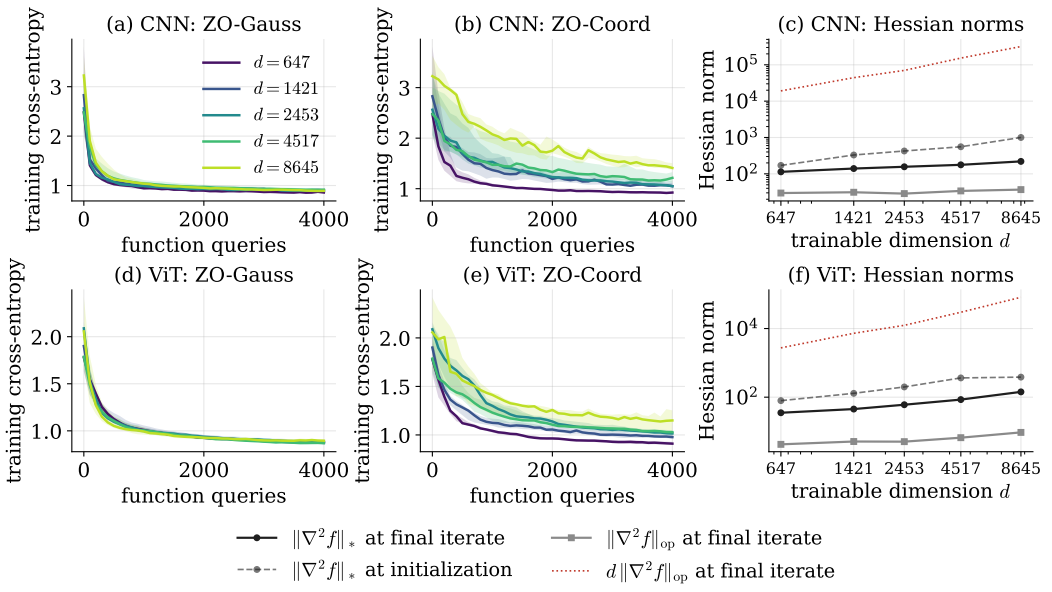

Figure 4: **Forward-only adaptation on CIFAR-10 for trainable dimensions d between 647 and 8645:** (a, b, d, e) training cross-entropy of ZO-Gauss and ZO-Coord over 4000 function queries (medians over five seeds with 10th to 90th percentile bands); (c, f) trace norm of the Hessian at initialization and at the final ZO-Gauss iterate, its operator norm, and the reference $d\,\|\nabla^2 f\|_{\mathrm{op}}$.

with function values only, so that the trainable dimension $d = 129r + 389$ ranges from 647 to 8645. A query is the cross-entropy on a minibatch of 128 target images with $\nu = 10^{-3}$. Each configuration is run for 4000 queries and five seeds with a step size chosen by a pilot run, and the exact Hessian of the training loss is computed by automatic differentiation at the initial and final ZO-Gauss iterates (Appendix F).

As d grows from 647 to 8645 in Figure 4, ZO-Gauss keeps a final training loss of about 0.9 and a target test accuracy of about 65% on both models, whereas the loss of ZO-Coord grows by about a half on the CNN and a quarter on the ViT (Table 1). The trace norm grows with d but stays two to three orders of magnitude below $d\,\|\nabla^2 f\|_{\mathrm{op}}$ (panels (c) and (f)). The added adapter directions, however, are not exploited by ZO-Gauss within the budget: without the adapter ($r = 0$, $d = 389$) it reaches the same loss, whereas Adam lowers the training loss monotonically in r (Appendix F). They thus neither help nor hurt ZO-Gauss, while they dilute the sampling of ZO-Coord, as in Section 6.1, where ZO-Coord samples the first axis only with probability $1/d$.

7 Conclusion

We analyzed the standard two-point Gaussian estimator with exact or noisy function values under a pointwise bound H on the trace norm of the Hessian. The stationarity and PL rates require no curvature envelope, and the dimension enters them and the finite-difference radius only through H. Numerically, ZO-Gauss crosses every window of the Twisted Valley at a rate independent of d and, within a fixed budget, is insensitive to the trainable dimension of a CNN and a Vision Transformer adapter, whereas coordinate probing degrades with d. Sharper constants, which would enlarge the admissible step size, and a probe-distribution theory covering the random subspace and sparse estimators are left for future work.

AI use statement

In this work, we used generative AI tools for writing assistance (drafting and polishing the text), for assistance with the proofs (checking derivations and suggesting intermediate steps), for assis-

tance with the experiments (writing and debugging experiment code), and for partial assistance with surveying the related work (finding and summarizing candidate references). We have not used generative AI tools for proposing the research concept or the main ideas of this paper or for designing the overall structure of the paper; these were carried out by the authors. Other tasks with required disclosure, such as the generation of data or the use of AI as an evaluator, are not applicable to this work. We have reviewed all AI-assisted work. All AI-assisted proof steps were re-derived and checked by hand by the authors, all AI-assisted experiment code was verified and tested for correctness by the authors, and every cited reference was checked against its original source by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work is theoretical and methodological in nature. It studies the convergence of zeroth-order optimization algorithms and evaluates them on a synthetic test function (the Twisted Valley) and on the forward-only adaptation of a CNN and a Vision Transformer pre-trained on the standard CIFAR-10 benchmark, together with Hessian diagnostics of these models. It does not involve human subjects, the release of new datasets, or personally identifiable information. We are not aware of direct negative societal impacts specific to this work beyond those generally associated with improved black-box optimization methods.

Reproducibility statement

All assumptions are stated explicitly in Sections 3.2 and 4. Complete proofs of the results in Section 4 and Section 5 are given in Appendix A and Appendix B, respectively. The test function, the dimensions, the smoothing parameter, and the step size used in the numerical experiments are specified in Section 6.1; the setup of the CNN and Vision Transformer experiments and Hessian diagnostics are described in Section 6.2 and Appendices E and F.

References

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer, Cham, 2014. doi: 10.1007/978-3-319-00227-9.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

Erik Berglund, Sarit Khirirat, and Xiaoyu Wang. Zeroth-order randomized subspace Newton methods. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6002–6006. IEEE, 2022.

HanQin Cai, Daniel McKenzie, Wotao Yin, and Zhenliang Zhang. Zeroth-order regularized optimization (ZORO): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022.

Coralia Cartis and Lindon Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization. *Mathematical Programming*, 199(1–2):461–524, 2023.

Aochuan Chen, Yimeng Zhang, Jinghan Jia, James Diffenderfer, Konstantinos Parasyris, Jiancheng Liu, Yihua Zhang, Zheng Zhang, Bhavya Kailkhura, and Sijia Liu. DeepZero: Scaling up zeroth-order optimization for deep model training. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order finetuning for language models with low-rank structures. In *The Thirteenth International Conference on Learning Representations*, 2025.

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.

Alexander Vladimirovich Gasnikov, Aleksandr Vladimirovich Lobanov, and Fedor Sergeevich Stonyakin. Highly smooth zeroth-order methods for solving optimization problems under the PL condition. *Computational Mathematics and Mathematical Physics*, 64(4):739–770, 2024.

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

Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang. Gradientless descent: High-dimensional zeroth-order optimization. In *International Conference on Learning Representations*, 2020.

Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R. Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, Beidi Chen, and Zhaozhuo Xu. Zeroth-order finetuning of LLMs with transferable static sparsity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Dongyoon Kim, Sungjae Lee, Wonjin Lee, and Kwang In Kim. Subspace-based approximate Hessian method for zeroth-order optimization. *arXiv preprint arXiv:2507.06125*, 2025.

David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 199(1–2):1179–1219, 2023.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse MeZO: Less parameters for better performance in zeroth-order LLM fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shaocong Ma and Heng Huang. Revisiting zeroth-order optimization: Minimum-variance two-point estimators and directionally aligned perturbations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Ryota Nozawa, Pierre-Louis Poirion, and Akiko Takeda. Zeroth-order random subspace algorithm for non-smooth convex optimization. *Journal of Optimization Theory and Applications*, 204(3): 53, 2025.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1–2):1–38, 2014.

Lindon Roberts and Clément W Royer. Direct search based on probabilistic descent in reduced spaces. *SIAM Journal on Optimization*, 33(4):3057–3082, 2023.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, volume 25, pp. 2951–2959, 2012.

Yining Wang, Simon Du, Sivaraman Balakrishnan, and Aarti Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1356–1365. PMLR, 2018.

Haishan Ye, Zhichao Huang, Cong Fang, Chris Junchi Li, and Tong Zhang. Hessian-aware zeroth-order optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6): 4869–4877, 2025.

Pengyun Yue, Long Yang, Cong Fang, and Zhouchen Lin. Zeroth-order optimization with weak dimension dependency. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4429–4472. PMLR, 2023.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, and Niao He. DPZero: Private fine-tuning of language models without backpropagation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59210–59246. PMLR, 2024.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, Michael Muehlebach, and Niao He. Zeroth-order optimization finds flat minima. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor W. Tsang. Second-order fine-tuning without pain for LLMs: A Hessian informed zeroth-order optimizer. In *The Thirteenth International Conference on Learning Representations*, 2025.

A PROOFS OF THEOREMS 4.1 AND 4.3

This appendix proves Theorems 4.1 and 4.3 and Corollary 4.4 by establishing the one-step descent inequality of Proposition A.8 and summing it over the iterations. The difficulty is that the update point $x_t - h\tilde{D}_t u_t$ depends on the direction u_t through the random radius $h\tilde{D}_t$, and bounding the Hessian there by its operator norm would introduce a factor of d. We avoid this factor by Gaussian estimates that use only the trace-norm bound. Expectations that are linear in the Taylor remainder Φ defined in (11) below are bounded by a symmetrization over the rotations that fix the gradient (Lemmas A.2 and A.3). The derivative of Φ along rays through the origin is bounded in mean square by the Gaussian Poincaré inequality and an integrated Bochner identity (Lemmas A.4 and A.5), and its integrated form bounds Φ at a point whose radius is an arbitrary random variable (Lemma A.6).

Fix an iterate x_t and define a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ by

$$\Phi(z) \coloneqq f(x_t + z) - f(x_t) - \nabla f(x_t)^\top z. \tag{11}$$

Lemma A.1. *Suppose Assumption 3.2 holds. Then $\Phi(\mathbf{0}) = 0$, $\nabla\Phi(\mathbf{0}) = \mathbf{0}$, and*

$$\|\nabla^2\Phi(z)\|_* \le H \qquad \text{for all } z \in \mathbb{R}^d.$$

Consequently, it holds that

$$|\Phi(z)| \le \frac{H}{2}\|z\|^2, \qquad \|\nabla\Phi(z)\| \le H\|z\|. \tag{12}$$

Proof. The first two identities follow from (11), and $\nabla^2\Phi(z) = \nabla^2 f(x_t + z)$ has trace norm at most H by Assumption 3.2. Since $\|\nabla^2\Phi\|_{\mathrm{op}} \le \|\nabla^2\Phi\|_*$, Taylor's theorem around $z = \mathbf{0}$ gives (12). □

We write $\mathcal{C}(H)$ for the class of functions $\phi \in C^2(\mathbb{R}^d)$ with $\phi(\mathbf{0}) = 0$, $\nabla\phi(\mathbf{0}) = \mathbf{0}$, and $\|\nabla^2\phi(z)\|_* \le H$ for all $z \in \mathbb{R}^d$. Lemma A.1 states that $\Phi \in \mathcal{C}(H)$, and (12) holds for every $\phi \in \mathcal{C}(H)$ with the same proof.

Let $u \sim \mathcal{N}(0, I_d)$, which represents a random direction in (2). For a unit vector $v \in \mathbb{R}^d$, write $P_\perp \coloneqq I_d - vv^\top$, let $G_v \coloneqq \{Q \in O(d) : Qv = v\} \cong O(d-1)$ with Haar probability measure dQ, and define the G_v-average $\bar{\Phi}_v(z) \coloneqq \int_{G_v} \Phi(Qz)dQ$. The key idea is to trade the expectation over u for an average over Q: since $Qu \overset{d}{=} u$ for every $Q \in G_v$,

$$\mathbb{E}[\Phi(u)] = \int_{G_v} \mathbb{E}[\Phi(Qu)]dQ = \mathbb{E}\big[\bar{\Phi}_v(u)\big],$$

and $\bar{\Phi}_v$ is easier to bound than Φ itself.

Lemma A.2. *Let $d \ge 3$. If Assumption 3.2 holds, then for all $z \in \mathbb{R}^d$,*

$$\big|\bar{\Phi}_v(z)\big| \le \frac{H}{2}(v^\top z)^2 + \frac{H}{2(d-2)}\|P_\perp z\|^2. \tag{13}$$

Proof. Step 1: Properties of $\bar{\Phi}_v$. Since G_v is compact and $\Phi \in C^2$, differentiation under the integral gives

$$\nabla\bar{\Phi}_v(z) = \int Q^\top \nabla\Phi(Qz)dQ$$

and

$$\nabla^2\bar{\Phi}_v(z) = \int Q^\top \nabla^2\Phi(Qz)QdQ.$$

Hence $\bar{\Phi}_v(\mathbf{0}) = 0$, $\nabla\bar{\Phi}_v(\mathbf{0}) = \mathbf{0}$, and, by the unitary invariance of the trace norm and Jensen's inequality, we obtain

$$\|\nabla^2\bar{\Phi}_v(z)\|_* \le \int \|\nabla^2\Phi(Qz)\|_* dQ \le H.$$

By right-invariance of the Haar measure, $\bar{\Phi}_v(Q_0 z) = \bar{\Phi}_v(z)$ for every $Q_0 \in G_v$. If $z, z' \in \mathbb{R}^d$ satisfy $v^\top z = v^\top z'$ and $\|P_\perp z\| = \|P_\perp z'\|$, then some orthogonal matrix Q fixes v and maps $P_\perp z$ to $P_\perp z'$, so $Qz = z'$ with $Q \in G_v$ and hence $\bar{\Phi}_v(z) = \bar{\Phi}_v(z')$. Thus $\bar{\Phi}_v(z)$ depends on z only through $v^\top z$ and $\|P_\perp z\|$: there is a function $\varphi : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ with

$$\bar{\Phi}_v(z) = \varphi(s, \rho), \qquad s := v^\top z, \quad \rho := \|P_\perp z\|.$$

Concretely, $\varphi(s, \rho) = \bar{\Phi}_v(sv + \rho e)$ for any fixed unit vector $e \perp v$; thus $\varphi \in C^2(\mathbb{R}^2)$ and φ is even in ρ.

Step 2: The multiplicity-$(d-2)$ eigenvalue. Fix z with $\rho > 0$ and choose an orthonormal basis $e_1, \ldots, e_d$ of $\mathbb{R}^d$ with $e_1 = v$ and $e_2 = P_\perp z/\rho$. For $i \geq 3$ and $t \in \mathbb{R}$ we have $v^\top(z + te_i) = s$ and $\|P_\perp(z + te_i)\| = \sqrt{\rho^2 + t^2}$, so

$$\bar{\Phi}_v(z + te_i) = \varphi\Big(s, \sqrt{\rho^2 + t^2}\Big), \qquad e_i^\top \nabla^2 \bar{\Phi}_v(z) e_i = \frac{d^2}{dt^2} \bar{\Phi}_v(z + te_i)\Big|_{t=0} = \frac{\varphi_\rho(s, \rho)}{\rho}.$$

Similarly, for $i \geq 3$ and $j \neq i$,

$$\bar{\Phi}_v(z + te_i + t'e_j) = \begin{cases} \varphi\Big(s + t', \sqrt{\rho^2 + t^2}\Big), & j = 1, \\ \varphi\Big(s, \sqrt{(\rho + t')^2 + t^2}\Big), & j = 2, \\ \varphi\Big(s, \sqrt{\rho^2 + t^2 + t'^2}\Big), & j \geq 3, \end{cases}$$

and each right-hand side depends on t only through t^2, so its mixed derivative $\partial_t \partial_{t'}$ vanishes at $(t, t') = (0, 0)$, i.e., $e_j^\top \nabla^2 \bar{\Phi}_v(z) e_i = 0$. Hence $\nabla^2 \bar{\Phi}_v(z) e_i = \frac{\varphi_\rho(s,\rho)}{\rho} e_i$ for $i = 3, \ldots, d$, that is, $\frac{\varphi_\rho(s,\rho)}{\rho}$ is an eigenvalue of $\nabla^2 \bar{\Phi}_v(z)$ of multiplicity at least $d - 2$. Since the trace norm is the sum of the absolute values of the eigenvalues,

$$(d-2)\frac{|\varphi_\rho(s, \rho)|}{\rho} \leq \|\nabla^2 \bar{\Phi}_v(z)\|_* \leq H,$$

which gives

$$|\varphi_\rho(s, \rho)| \leq \frac{H\rho}{d-2}.$$

Also $\varphi_{ss}(s, \rho) = v^\top \nabla^2 \bar{\Phi}_v(z) v$, so $|\varphi_{ss}| \leq \|\nabla^2 \bar{\Phi}_v\|_{\mathrm{op}} \leq H$; this holds for $\rho = 0$ as well, since $\varphi(s, 0) = \bar{\Phi}_v(sv)$.

Step 3: Integration. Since $\bar{\Phi}_v(z) = \varphi(s, \rho)$ with $s = v^\top z$ and $\rho = \|P_\perp z\|$, the triangle inequality gives

$$\big|\bar{\Phi}_v(z)\big| = |\varphi(s, \rho)| \leq |\varphi(s, 0)| + |\varphi(s, \rho) - \varphi(s, 0)|.$$

By Step 1, $\varphi(0, 0) = \bar{\Phi}_v(\mathbf{0}) = 0$ and $\varphi_s(0, 0) = v^\top \nabla \bar{\Phi}_v(\mathbf{0}) = 0$. Hence Taylor's theorem with integral remainder in s, together with $|\varphi_{ss}| \leq H$ from Step 2, gives

$$|\varphi(s, 0)| = \Big|\int_0^s (s - r)\varphi_{ss}(r, 0) dr\Big| \leq H\Big|\int_0^s (s - r) dr\Big| = \frac{Hs^2}{2}.$$

Next, the fundamental theorem of calculus in ρ, together with $|\varphi_\rho(s, r)| \leq \frac{Hr}{d-2}$ for $r > 0$ from Step 2, gives

$$|\varphi(s, \rho) - \varphi(s, 0)| = \left|\int_0^\rho \varphi_\rho(s, r) dr\right| \leq \int_0^\rho \frac{Hr}{d-2} dr = \frac{H\rho^2}{2(d-2)},$$

which proves (13). □

The following consequence of Lemma A.2 is used repeatedly below.

Lemma A.3 (Symmetrized expectations)**.** *Let $d \geq 3$ and suppose Assumption 3.2 holds. If $\theta, \eta : \mathbb{R}^d \to \mathbb{R}$ are measurable and G_v-invariant, i.e., $\theta(Qu) = \theta(u)$ and $\eta(Qu) = \eta(u)$ for $Q \in G_v$, and $\mathbb{E}_u\Big[\theta(u)^2 \|u\|^2 |\eta(u)|\Big] < \infty$, then*

$$|\mathbb{E}_u[\Phi(\theta(u)u)\eta(u)]| \leq \frac{H}{2}\mathbb{E}_u\left[\theta(u)^2\left((v^\top u)^2 + \frac{\|P_\perp u\|^2}{d-2}\right)|\eta(u)|\right]. \tag{14}$$

Proof. For $Q \in G_v$ we have $Q(\theta(u)u) = \theta(Qu)Qu$ and $\eta(u) = \eta(Qu)$, and $Qu \overset{d}{=} u$. Hence $\mathbb{E}[\Phi(\theta(u)u)\eta(u)] = \mathbb{E}[\Phi(Q\theta(u)u)\eta(u)]$ for every $Q \in G_v$; integrating over Q and using Fubini's theorem, justified by (12) and the moment assumption, gives $\mathbb{E}[\Phi(\theta(u)u)\eta(u)] = \mathbb{E}\big[\bar{\Phi}_v(\theta(u)u)\eta(u)\big]$. Now apply (13) with $z = \theta(u)u$. □

The next lemma bounds the second moment of $\nabla\Phi(u)$ by a constant depending only on H, whereas the crude bound (12) gives only $H^2\mathbb{E}\Big[\|u\|^2\Big] = dH^2$.

Lemma A.4. *Suppose Assumption 3.2 holds. Then*

$$\mathbb{E}\Big[\|\nabla\Phi(u)\|^2\Big] \leq \frac{89}{25}H^2.$$

Proof. If $d \leq 2$, (12) gives $\mathbb{E}\Big[\|\nabla\Phi(u)\|^2\Big] \leq dH^2 \leq \frac{89}{25}H^2$, so let $d \geq 3$. We use the Gaussian Poincaré inequality (Bakry et al., 2014) $\mathrm{Var}(\phi(u)) \leq \mathbb{E}\Big[\|\nabla\phi(u)\|^2\Big]$, valid for every locally Lipschitz $\phi : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}\Big[\|\nabla\phi(u)\|^2\Big] < \infty$. Let $m \coloneqq \mathbb{E}[\nabla\Phi(u)]$. The Poincaré inequality for $\phi = \partial_i\Phi$, whose gradient is the i-th row of $\nabla^2\Phi$, summed over i gives

$$\mathbb{E}\Big[\|\nabla\Phi(u)\|^2\Big] = \sum_i \mathrm{Var}(\partial_i\Phi(u)) + \|m\|^2 \leq \mathbb{E}\Big[\big\|\nabla^2\Phi(u)\big\|_{\mathrm{F}}^2\Big] + \|m\|^2 \leq H^2 + \|m\|^2. \tag{15}$$

To bound $\|m\|$, we may assume $m \neq \mathbf{0}$ and put $v \coloneqq m/\|m\|$. Gaussian integration by parts gives $\|m\| = \mathbb{E}\big[v^\top\nabla\Phi(u)\big] = \mathbb{E}\big[(v^\top u)\Phi(u)\big]$. Apply (14) with this v, $\theta \equiv 1$, and $\eta(u) = v^\top u$; since $v^\top u \sim \mathcal{N}(0,1)$ and $\|P_\perp u\|^2 \sim \chi^2_{d-1}$ are independent,

$$\|m\| \leq \frac{H}{2}\mathbb{E}\left[\big|v^\top u\big|\left((v^\top u)^2 + \frac{\|P_\perp u\|^2}{d-2}\right)\right] = \frac{H}{2}\sqrt{\frac{2}{\pi}}\left(2 + \frac{d-1}{d-2}\right) \leq \frac{8}{5}H,$$

using $\mathbb{E}\Big[\big|v^\top u\big|^3\Big] = 2\sqrt{2/\pi}$, $\mathbb{E}\big[\big|v^\top u\big|\big] = \sqrt{2/\pi} \leq \frac{4}{5}$, and $\frac{d-1}{d-2} \leq 2$. Inserting this bound into (15) gives $\mathbb{E}\Big[\|\nabla\Phi(u)\|^2\Big] \leq H^2 + \frac{64}{25}H^2 = \frac{89}{25}H^2$. □

The proofs of Lemmas A.2 to A.4 use only the properties of Φ stated in Lemma A.1, so the three lemmas hold for every $\phi \in \mathcal{C}(H)$ in place of Φ. We apply them in this form to rescalings and finite differences of Φ. For a random variable X we write $\|X\|_p \coloneqq (\mathbb{E}[|X|^p])^{1/p}$.

The next lemma is the key estimate. At a point τu on the ray through u, the crude bound (12) gives $\big|u^\top\nabla\phi(\tau u)\big| \leq H|\tau|\|u\|^2$, whose second moment is of order $H^2\tau^2d^2$; in root mean square, the derivative along the ray is smaller by a factor of d.

Lemma A.5. *Let $\phi \in \mathcal{C}(H)$ and $\tau \in \mathbb{R}$. Then*

$$\mathbb{E}\Big[\big(u^\top\nabla\phi(\tau u)\big)^2\Big] \leq 10H^2\tau^2. \tag{16}$$

Proof. The case $\tau = 0$ is trivial, so let $\tau \neq 0$. Suppose first that $\phi \in C^3$ with bounded third derivatives, and let $\chi(z) \coloneqq \phi(\tau z)$ and $\mathcal{L}\chi(z) \coloneqq \Delta\chi(z) - z^\top\nabla\chi(z)$, the Ornstein–Uhlenbeck operator. Gaussian integration by parts gives $\mathbb{E}[\zeta(u)\mathcal{L}\xi(u)] = -\mathbb{E}\big[\nabla\zeta(u)^\top\nabla\xi(u)\big]$ for $\zeta \in C^1$ and

$\xi \in C^2$ that grow at most polynomially together with their derivatives, and χ has this property with its derivatives up to order three. Applying it with $\zeta = \mathcal{L}\chi$ and $\xi = \chi$, and then, using $\partial_i(\mathcal{L}\chi) = \mathcal{L}(\partial_i\chi) - \partial_i\chi$, with $\zeta = \xi = \partial_i\chi$, we obtain the integrated Bochner identity (Bakry et al., 2014)

$$\mathbb{E}\Big[\big(\mathcal{L}\chi(u)\big)^2\Big] = -\sum_{i=1}^{d} \mathbb{E}[\partial_i\chi(u)\mathcal{L}(\partial_i\chi)(u)] + \mathbb{E}\Big[\|\nabla\chi(u)\|^2\Big] = \mathbb{E}\Big[\big\|\nabla^2\chi(u)\big\|_{\mathrm{F}}^2\Big] + \mathbb{E}\Big[\|\nabla\chi(u)\|^2\Big].$$

Since $\nabla^2\chi(z) = \tau^2\nabla^2\phi(\tau z)$, we have $\big\|\nabla^2\chi\big\|_{\mathrm{F}} \le \|\nabla^2\chi\|_* \le \tau^2 H$ and $|\Delta\chi| = \big|\operatorname{Tr}\nabla^2\chi\big| \le \tau^2 H$. Moreover, $\chi/\tau^2 \in \mathcal{C}(H)$, so Lemma A.4 gives $\mathbb{E}\Big[\|\nabla\chi(u)\|^2\Big] \le \frac{89}{25}\tau^4H^2$, and hence $\mathbb{E}\big[(\mathcal{L}\chi(u))^2\big] \le \frac{114}{25}\tau^4H^2$. Since $\tau u^\top\nabla\phi(\tau u) = u^\top\nabla\chi(u) = \Delta\chi(u) - \mathcal{L}\chi(u)$, Minkowski's inequality gives

$$|\tau|\|u^\top\nabla\phi(\tau u)\|_2 \le \|\Delta\chi(u)\|_2 + \|\mathcal{L}\chi(u)\|_2 \le \left(1 + \sqrt{\frac{114}{25}}\right)\tau^2 H,$$

and (16) follows from $\left(1 + \sqrt{114/25}\right)^2 < 10$.

For general $\phi \in \mathcal{C}(H)$, let ρ_ε be a smooth probability density supported in the ball of radius ε, let $\phi_\varepsilon := \phi * \rho_\varepsilon$, and let $\hat\phi_\varepsilon(z) := \phi_\varepsilon(z) - \phi_\varepsilon(\mathbf{0}) - \nabla\phi_\varepsilon(\mathbf{0})^\top z$. Then $\nabla^2\hat\phi_\varepsilon = (\nabla^2\phi) * \rho_\varepsilon$ has trace norm at most H by the convexity of the trace norm, and the third derivatives $(\nabla^2\phi) * \nabla\rho_\varepsilon$ are bounded since $\|\nabla^2\phi\|_{\mathrm{op}} \le H$; hence $\hat\phi_\varepsilon \in \mathcal{C}(H)$ satisfies (16). As $\varepsilon \to 0$, $\nabla\phi_\varepsilon \to \nabla\phi$ pointwise and $\nabla\phi_\varepsilon(\mathbf{0}) \to \nabla\phi(\mathbf{0}) = \mathbf{0}$, so $u^\top\nabla\hat\phi_\varepsilon(\tau u) \to u^\top\nabla\phi(\tau u)$; since $\Big|u^\top\nabla\hat\phi_\varepsilon(\tau u)\Big| \le H|\tau|\|u\|^2$ by (12), dominated convergence yields (16) for ϕ. □

Integrating (16) along the ray bounds ϕ at a point whose radius is random.

Lemma A.6. *Let $\phi \in \mathcal{C}(H)$, and let θ be a real random variable, defined on the same probability space as u and possibly dependent on it, with $\mathbb{E}\big[\theta^6\big] < \infty$. Then*

$$\mathbb{E}[|\phi(\theta u)|] \le 7H\|\theta\|_6^2. \tag{17}$$

Proof. For fixed u, let $G(s) := u^\top\nabla\phi(su) = \frac{d}{ds}\phi(su)$. Since $\phi(\mathbf{0}) = 0$,

$$|\phi(\theta u)| = \Big|\int_0^\theta G(s)ds\Big| \le \int_0^\infty \big(\mathbf{1}_{\{\theta \ge s\}}|G(s)| + \mathbf{1}_{\{\theta \le -s\}}|G(-s)|\big)ds.$$

Taking expectations, Tonelli's theorem, the Cauchy–Schwarz inequality for each s, (16) with $\tau = \pm s$, and $\sqrt{p} + \sqrt{q} \le \sqrt{2(p+q)}$ give

$$\mathbb{E}[|\phi(\theta u)|] \le \int_0^\infty \Big(\sqrt{P(\theta \ge s)} + \sqrt{P(\theta \le -s)}\Big)\sqrt{10}Hs\,ds \le \sqrt{20}H\int_0^\infty s\sqrt{P(|\theta| \ge s)}ds.$$

By Markov's inequality, $P(|\theta| \ge s) \le \min\{1, M/s^6\}$ with $M := \mathbb{E}\big[\theta^6\big]$, so with $s_0 := M^{1/6}$,

$$\int_0^\infty s\sqrt{P(|\theta| \ge s)}ds \le \int_0^{s_0} sds + \int_{s_0}^\infty \frac{\sqrt{M}}{s^2}ds = \frac{s_0^2}{2} + \frac{\sqrt{M}}{s_0} = \frac{3}{2}\|\theta\|_6^2,$$

and (17) follows from $\frac{3}{2}\sqrt{20} < 7$. □

Lemma A.6 is applied with the radius $-h\tilde D_t$ of the update, which involves the finite-difference error; the last lemma bounds the sixth moment of this error.

Lemma A.7. *Let $\phi \in \mathcal{C}(H)$, $\nu > 0$, and $\Psi(z) := \frac{1}{2}(\phi(\nu z) - \phi(-\nu z))$. Then $\|\Psi(u)\|_6 \le \frac{23}{2}\nu^2 H$.*

Proof. The function Ψ lies in $\mathcal{C}(\nu^2 H)$: $\Psi(\mathbf{0}) = 0$, $\nabla\Psi(z) = \frac{\nu}{2}(\nabla\phi(\nu z) + \nabla\phi(-\nu z))$ vanishes at $\mathbf{0}$, and $\nabla^2\Psi(z) = \frac{\nu^2}{2}\big(\nabla^2\phi(\nu z) - \nabla^2\phi(-\nu z)\big)$ has trace norm at most $\nu^2 H$. Write $L := \nu^2 H$ and $G := \|\nabla\Psi\|$. Since $\|\nabla^2\Psi\|_{\mathrm{op}} \le L$, G is L-Lipschitz, and Lemma A.4 for Ψ gives $\mathbb{E}\big[G(u)^2\big] \le$

$\frac{89}{25}L^2$. We apply the Poincaré inequality to the powers of $\tilde{G} \coloneqq G - \mathbb{E}[G(u)]$, which are locally Lipschitz with $\|\nabla \tilde{G}^k\| \le k\left|\tilde{G}\right|^{k-1} L$ almost everywhere. It gives $\mathbb{E}\left[\tilde{G}^2\right] \le L^2$ and $\mathrm{Var}(\tilde{G}^2) \le 4L^2\mathbb{E}\left[\tilde{G}^2\right] \le 4L^4$, so $\mathbb{E}\left[\tilde{G}^4\right] \le 5L^4$. Then, we obtain

$$\mathrm{Var}(\tilde{G}^3) \le 9L^2\mathbb{E}\left[\tilde{G}^4\right] \le 45L^6 \quad \text{and} \quad \left(\mathbb{E}\left[\tilde{G}^3\right]\right)^2 \le \mathbb{E}\left[\tilde{G}^2\right]\mathbb{E}\left[\tilde{G}^4\right] \le 5L^6,$$

and therefore, $\mathbb{E}\left[\tilde{G}^6\right] \le 50L^6$.

Hence, by Minkowski's inequality,

$$\|G(u)\|_6 \le \mathbb{E}[G(u)] + \|\tilde{G}(u)\|_6 \le \left(\sqrt{\frac{89}{25}} + 50^{1/6}\right)L \le 3.81L.$$

Since Ψ is odd and $u \overset{d}{=} -u$, $\mathbb{E}\big[\Psi(u)^3\big] = 0$, so the Poincaré inequality for Ψ^3 and Hölder's inequality with exponents $\frac{3}{2}$ and 3 give

$$\mathbb{E}\big[\Psi(u)^6\big] = \mathrm{Var}\big(\Psi(u)^3\big) \le 9\mathbb{E}\big[\Psi(u)^4 G(u)^2\big] \le 9\big(\mathbb{E}\big[\Psi(u)^6\big]\big)^{2/3}\big(\mathbb{E}\big[G(u)^6\big]\big)^{1/3}.$$

The left-hand side is finite by (12), so $\|\Psi(u)\|_6 \le 3\|G(u)\|_6 \le 11.43L \le \frac{23}{2}\nu^2 H$. □

We now return to the algorithm. Condition on $\mathcal{F}_t$, fix the iterate x_t, and write

$$g_t \coloneqq \nabla f(x_t), \quad q_t \coloneqq g_t^\top u_t, \quad \delta_t \coloneqq D_\nu f(x_t; u_t) - q_t,$$

so that, under Assumption 3.3, the algorithm uses $\tilde{D}_t = q_t + \delta_t + e_t/\nu$ and $x_{t+1} = x_t - h\tilde{D}_t u_t$. By (11), $f(x_t \pm \nu u_t) = f(x_t) \pm \nu q_t + \Phi(\pm\nu u_t)$, hence

$$\delta_t = \frac{\Phi(\nu u_t) - \Phi(-\nu u_t)}{2\nu} = \frac{1}{\nu}\Psi(u_t), \qquad \Psi(z) \coloneqq \frac{1}{2}(\Phi(\nu z) - \Phi(-\nu z)). \tag{18}$$

We now combine these estimates into the one-step descent inequality. All expectations and norms $\|\cdot\|_6$ below are conditional on $\mathcal{F}_t$. For $g_t \neq \mathbf{0}$ we use the axis $\hat{g}_t \coloneqq g_t/\|g_t\|$ and write $Z \coloneqq \hat{g}_t^\top u_t$ and $W \coloneqq \|u_t - Z\hat{g}_t\|^2$, so that $q_t = \|g_t\| Z$, where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_{d-1}$ are independent.

Proposition A.8 (Descent lemma)**.** *Suppose Assumptions 3.2 and 3.3 hold. For every constant step size $0 < h \le \frac{1}{100H}$ and every $\nu > 0$,*

$$\mathbb{E}[f(x_{t+1}) \mid \mathcal{F}_t] \le f(x_t) - \frac{h}{2}\|\nabla f(x_t)\|^2 + 43h\nu^2H^2 + 28Hh^2\frac{\sigma^2}{\nu^2}. \tag{19}$$

Proof. By (11) with $z = -h\tilde{D}_t u_t$, for which $g_t^\top z = -h\tilde{D}_t q_t$,

$$f(x_{t+1}) = f(x_t) - h\tilde{D}_t q_t + \Phi(-h\tilde{D}_t u_t), \tag{20}$$

and we bound the conditional expectations of the last two terms.

First-order term. Since $\tilde{D}_t = q_t + \delta_t + e_t/\nu$, we have

$$\mathbb{E}\left[\tilde{D}_t q_t\right] = \mathbb{E}\big[q_t^2\big] + \mathbb{E}[\delta_t q_t] + \frac{1}{\nu}\mathbb{E}[e_t q_t].$$

Here $\mathbb{E}\big[q_t^2\big] = \|g_t\|^2$, and since q_t is a function of u_t, (4) gives $\mathbb{E}[e_t q_t] = \mathbb{E}[q_t\mathbb{E}[e_t \mid \mathcal{F}_t, u_t]] = 0$. For the finite-difference error, let $g_t \neq \mathbf{0}$. Since $\delta_t q_t = \frac{1}{2\nu}[\Phi(\nu u_t)q_t - \Phi(-\nu u_t)q_t]$, for $d \ge 3$ Lemma A.3 applies to each summand with axis $\hat{g}_t$, constant radius $\pm\nu$, and weight q_t, so that

$$|\mathbb{E}[\delta_t q_t]| \le \frac{1}{2\nu} \cdot 2 \cdot \frac{H\nu^2}{2}\|g_t\|\,\mathbb{E}\left[|Z|\left(Z^2 + \frac{W}{d-2}\right)\right] = \frac{H\nu\|g_t\|}{2}\sqrt{\frac{2}{\pi}}\left(2 + \frac{d-1}{d-2}\right) \le \frac{8}{5}H\nu\|g_t\|,$$

using $\mathbb{E}\Big[|Z|^3\Big] = 2\sqrt{2/\pi}$, $\mathbb{E}[|Z|] = \sqrt{2/\pi} \le \frac{4}{5}$, and $\frac{d-1}{d-2} \le 2$. For $d \le 2$, (12) gives $|\delta_t| \le \frac{\nu H}{2}\|u_t\|^2$, and $\mathbb{E}\Big[|Z|\|u_t\|^2\Big] = \sqrt{2/\pi}(d+1)$ gives the smaller bound $\frac{6}{5}H\nu\|g_t\|$. The bound $|\mathbb{E}[\delta_t q_t]| \le \frac{8}{5}\nu H\|g_t\|$ also holds when $g_t = \mathbf{0}$, since then $q_t = 0$. With $xy \le \frac{x^2}{8} + 2y^2$,

$$-h\mathbb{E}\Big[\tilde{D}_t q_t\Big] \le -h\|g_t\|^2 + \frac{h}{8}\|g_t\|^2 + \frac{128}{25}h\nu^2 H^2. \tag{21}$$

Second-order term. The radius $\theta = -h\tilde{D}_t$ is a function of u_t and e_t, so Lemma A.6, applied to $\Phi \in \mathcal{C}(H)$, gives $\mathbb{E}\Big[\Big|\Phi(-h\tilde{D}_t u_t)\Big|\Big] \le 7Hh^2\|\tilde{D}_t\|_6^2$. By Minkowski's inequality, $\|\tilde{D}_t\|_6 \le \|q_t\|_6 + \|\delta_t\|_6 + \|e_t\|_6/\nu$. Since $q_t \sim \mathcal{N}(0, \|g_t\|^2)$, $\|q_t\|_6 = 15^{1/6}\|g_t\|$; Lemma A.7 applied to Φ and (18) give $\|\delta_t\|_6 \le \frac{23}{2}\nu H$; and $\|e_t\|_6 \le \sigma$ by (4). Two applications of $(a+b)^2 \le 2a^2 + 2b^2$ give

$$\|\tilde{D}_t\|_6^2 \le 2 \cdot 15^{1/3}\|g_t\|^2 + 4\left(\frac{23}{2}\right)^2\nu^2 H^2 + \frac{4\sigma^2}{\nu^2} \le 4.94\|g_t\|^2 + 529\nu^2 H^2 + \frac{4\sigma^2}{\nu^2},$$

and multiplying by $7Hh^2$ and using $hH \le \frac{1}{100}$,

$$\mathbb{E}\Big[\Big|\Phi(-h\tilde{D}_t u_t)\Big|\Big] \le \frac{7}{20}h\|g_t\|^2 + 37.1h\nu^2 H^2 + 28Hh^2\frac{\sigma^2}{\nu^2}.$$

Adding this bound to (21) in (20) gives

$$\mathbb{E}[f(x_{t+1}) \mid \mathcal{F}_t] \le f(x_t) - h\|g_t\|^2\left(1 - \frac{1}{8} - \frac{7}{20}\right) + \left(\frac{128}{25} + 37.1\right)h\nu^2 H^2 + 28Hh^2\frac{\sigma^2}{\nu^2},$$

which implies (19), since $1 - \frac{1}{8} - \frac{7}{20} = \frac{21}{40} \ge \frac{1}{2}$ and $\frac{128}{25} + 37.1 \le 43$. □

For exact function values, $\sigma = 0$ and the residual $43h\nu^2 H^2$ in (19) does not depend on d. The constants are not optimized. In particular, the restriction $h \le \frac{1}{100H}$ comes from bounding $\Phi(-h\tilde{D}_t u_t)$ by Lemma A.6 as a whole, including the part with radius $-hq_t$, for which Lemma A.3 gives the constant $\frac{5}{2}$ in place of $\frac{7}{20} \cdot 100 = 35$; combining the two estimates would allow a larger step size.

The total expectations in the proofs below are finite. Indeed, f is H-smooth, so $|f(x)| \le |f(x_0)| + \|\nabla f(x_0)\|\,\|x - x_0\| + \frac{H}{2}\|x - x_0\|^2$ and $\|\nabla f(x_t)\| \le \|\nabla f(x_0)\| + H\|x_t - x_0\|$; together with $\mathbb{E}\Big[\|x_{t+1} - x_t\|^2 \mid \mathcal{F}_t\Big] = h^2\mathbb{E}\Big[\tilde{D}_t^2\|u_t\|^2\Big] \le h^2\|\tilde{D}_t\|_6^2\Big(\mathbb{E}\Big[\|u_t\|^3\Big]\Big)^{2/3}$ and the bound on $\|\tilde{D}_t\|_6^2$ above, an induction on t shows $\mathbb{E}\Big[\|x_t - x_0\|^2\Big] < \infty$ for every t.

Proof of Theorem 4.1. Take total expectations in (19), sum over $t = 0, \dots, T-1$, and use $\mathbb{E}[f(x_T)] \ge f^*$:

$$\frac{h}{2}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla f(x_t)\|^2\Big] \le f(x_0) - f^* + T\left(43h\nu^2 H^2 + 28Hh^2\frac{\sigma^2}{\nu^2}\right).$$

Dividing by $hT/2$ and bounding the minimum by the average gives

$$\min_{0 \le t < T}\mathbb{E}\Big[\|\nabla f(x_t)\|^2\Big] \le \frac{2(f(x_0) - f^*)}{hT} + 86\nu^2 H^2 + 56Hh\frac{\sigma^2}{\nu^2},$$

which is (5). □

Proof of Theorem 4.3. Inserting the PL inequality (6) into (19) and taking total expectations gives

$$\mathbb{E}[f(x_{t+1}) - f^*] \le (1 - \mu h)\mathbb{E}[f(x_t) - f^*] + 43h\nu^2 H^2 + 28Hh^2\frac{\sigma^2}{\nu^2}.$$

By Assumption 3.2, f is H-smooth, so $f^* \le f\big(x - \frac{1}{H}\nabla f(x)\big) \le f(x) - \frac{1}{2H}\|\nabla f(x)\|^2$ for every x; comparing this with (6) at a point with $f(x) > f^*$ gives $\mu \le H$, hence $0 < 1 - \mu h < 1$ for $h \le \frac{1}{100H}$. Iterating and summing the geometric series yields

$$\mathbb{E}[f(x_t) - f^*] \le (1-\mu h)^t (f(x_0) - f^*) + \frac{1}{\mu}\left(43\nu^2 H^2 + 28Hh\frac{\sigma^2}{\nu^2}\right),$$

which is (7). □

Proof of Corollary 4.4. (i) The choice $\nu^2 = \frac{\epsilon}{344H^2}$ gives $86\nu^2H^2 = \frac{\epsilon}{4}$, and $h \le \frac{\epsilon^2}{8\cdot 10^4 H^3\sigma^2}$ gives

$$56Hh\frac{\sigma^2}{\nu^2} = \frac{19264H^3h\sigma^2}{\epsilon} \le \frac{\epsilon}{4}.$$

Since $T \ge \frac{4\Delta}{h\epsilon}$ gives $\frac{2\Delta}{hT} \le \frac{\epsilon}{2}$, (5) yields the claim, and $\frac{4\Delta}{h\epsilon} = \max\big\{\frac{400H\Delta}{\epsilon}, \frac{3.2\cdot 10^5 H^3\sigma^2\Delta}{\epsilon^3}\big\}$.

(ii) The choice $\nu^2 = \frac{\mu\epsilon}{172H^2}$ gives $\frac{43}{\mu}\nu^2H^2 = \frac{\epsilon}{4}$, and $h \le \frac{\mu^2\epsilon^2}{2\cdot 10^4 H^3\sigma^2}$ gives

$$\frac{28}{\mu}Hh\frac{\sigma^2}{\nu^2} = \frac{4816H^3h\sigma^2}{\mu^2\epsilon} \le \frac{\epsilon}{4}.$$

Since $(1-\mu h)^T\Delta \le e^{-\mu hT}\Delta \le \frac{\epsilon}{2}$ for $T \ge \frac{1}{\mu h}\log\frac{2\Delta}{\epsilon}$, (7) yields the claim, and $\frac{1}{\mu h} = \max\big\{\frac{100H}{\mu}, \frac{2\cdot 10^4 H^3\sigma^2}{\mu^3\epsilon^2}\big\}$. □

B Proof of Proposition 5.1

This appendix proves Proposition 5.1. Throughout, we write

$$b(t) \coloneqq (1-t^2)_+^3, \qquad \beta(t) \coloneqq \frac{1}{1+t^2}, \qquad \psi(x) \coloneqq x_d - \beta(x_d)\Lambda(x),$$

so that $\Lambda(x) = a\sum_{k=2}^{d-1}\sin(\omega x_k)b(x_1 - k)$ and $f = \psi^2 + \frac{\kappa}{2}x_1^2$ in (8). We abbreviate $\mu_0 \coloneqq 2\big(1 - \frac{3\sqrt{3}}{4}a\big)^2$, so that the PL constant of Proposition 5.1 is $\mu = \min\{\mu_0 - 8a^2, \kappa/4\}$. Primes denote derivatives of functions of one variable, and ∂_k denotes the partial derivative with respect to x_k. Since Λ does not depend on x_d, the d-th component of $\nabla\Lambda$ and the d-th row and column of $\nabla^2\Lambda$ vanish. The constant H of Proposition 5.1 is

$$H \coloneqq 2\left[\left(1 + \tfrac{3\sqrt{3}}{4}a\right)^2 + G_{a,\omega}^2\right] + 2\left[\tfrac{K_{a,\omega}}{2} + G_{a,\omega} + 2a + 2a\left(K_{a,\omega} + \tfrac{3\sqrt{3}}{4}G_{a,\omega} + 4a\right)\right] + \kappa, \tag{22}$$

where $G_{a,\omega} \coloneqq a\sqrt{12 + 2\omega^2}$ and $K_{a,\omega} \coloneqq 2a\big(6 + \frac{7}{2}\omega + \omega^2\big)$.

The following lemmas describe the properties of the functions b and β.

Lemma B.1 (Properties of b)**.** *The function b belongs to $C^2(\mathbb{R})$, satisfies $0 \le b \le 1$, $b(0) = 1$, and $b(t) = 0$ for $|t| \ge 1$, and $\sup|b'| = \frac{96}{25\sqrt{5}} < 1.72$, $\sup|b''| = 6$.*

Proof. On $|t| < 1$, $b'(t) = -6t(1-t^2)^2$ and $b''(t) = (1-t^2)(30t^2 - 6)$; both vanish at $t = \pm 1$, so $b \in C^2(\mathbb{R})$. The derivative of b' is $6(1-t^2)(5t^2-1)$, so $\sup|b'| = \big|b'(1/\sqrt{5})\big| = \frac{96}{25\sqrt{5}}$; the derivative of b'' is $24t(3-5t^2)$, so we obtain

$$\sup|b''| = \max\{|b''(0)|, |b''(\sqrt{3/5})|\} = \max\{6, \tfrac{24}{5}\} = 6.$$

□

Lemma B.2 (Properties of β)**.** *The function β satisfies, for all $t \in \mathbb{R}$,*

$$\begin{aligned} 0 < \beta(t) \le 1, &\qquad |t\beta(t)| \le \tfrac{1}{2}, \\ |\beta'(t)| \le \tfrac{3\sqrt{3}}{8}, &\qquad |t\beta'(t)| \le \tfrac{1}{2}, \\ |\beta''(t)| \le 2, &\qquad |t\beta''(t)| \le 1. \end{aligned} \tag{23}$$

Proof. Write $w := t^2$, so that $\beta'(t) = -2t/(1+w)^2$ and $\beta''(t) = (6w-2)/(1+w)^3$. The function $t \mapsto 2t/(1+t^2)^2$ is maximized at $t = 1/\sqrt{3}$, where it equals $\frac{3\sqrt{3}}{8}$; and $|6w-2| \le 2+6w \le 2(1+w)^3$. Next, $|t\beta(t)| = \sqrt{w}/(1+w) \le \frac{1}{2}$ and $|t\beta'(t)| = 2w/(1+w)^2 \le \frac{1}{2}$, both by $2\sqrt{w} \le 1+w$. Finally, $|t\beta''(t)| = \sqrt{w}|6w-2|/(1+w)^3$. For $w \ge \frac{1}{3}$, $\sqrt{w}(6w-2) \le 6w^{3/2} \le 3w+3w^2 \le (1+w)^3$ by the AM–GM inequality; for $w \le \frac{1}{3}$, the function $w \mapsto \sqrt{w}(2-6w)$ is maximized at $w = \frac{1}{9}$, so $\sqrt{w}(2-6w) \le \frac{4}{9} \le (1+w)^3$. □

Lemma B.3. *For $x \in \mathbb{R}^d$ let $I(x_1) := \{k \in \{2,\dots,d-1\} : |x_1 - k| < 1\}$, a set with at most two elements. Then $\Lambda \in C^2(\mathbb{R}^d)$, $\partial_j \Lambda(x) = 0$ and $\partial_{ij}\Lambda(x) = 0$ for all i whenever $j \notin \{1\} \cup I(x_1)$, and*

$$|\Lambda(x)| \le 2a, \qquad \|\nabla\Lambda(x)\| \le G_{a,\omega}, \qquad \|\nabla^2\Lambda(x)\|_* \le K_{a,\omega}. \tag{24}$$

Proof. Since $b \in C^2$ and $b(x_1-k) = 0$ unless $k \in I(x_1)$, only the terms with $k \in I(x_1)$ contribute to Λ and to its derivatives at x, which gives the support statement. For $k \in I(x_1)$,

$$\begin{aligned}
\partial_1\Lambda &= a\sum_{k\in I(x_1)} \sin(\omega x_k) b'(x_1-k), & \partial_k\Lambda &= a\omega\cos(\omega x_k) b(x_1-k),\\
\partial_{11}\Lambda &= a\sum_{k\in I(x_1)} \sin(\omega x_k) b''(x_1-k), & \partial_{1k}\Lambda &= a\omega\cos(\omega x_k) b'(x_1-k),\\
\partial_{kk}\Lambda &= -a\omega^2\sin(\omega x_k) b(x_1-k), & \partial_{kl}\Lambda &= 0 \quad (k \ne l).
\end{aligned}$$

Using $|I(x_1)| \le 2$ and Lemma B.1, we get $|\Lambda| \le 2a$ and $\|\nabla\Lambda\|^2 \le (2a\cdot 1.72)^2 + 2(a\omega)^2 \le a^2(12+2\omega^2) = G_{a,\omega}^2$. Since the trace norm is a norm and $\|e_i e_j^\top\|_* = 1$, we obtain

$$\|\nabla^2\Lambda\|_* \le \sum_{i,j} |\partial_{ij}\Lambda| \le 2a\cdot 6 + 4a\omega\cdot 1.72 + 2a\omega^2 \le K_{a,\omega},$$

which gives the desired bound. □

We now prove Proposition 5.1.

Proof of Proposition 5.1. (i) By Lemma B.3 and the smoothness of β, $\psi \in C^2(\mathbb{R}^d)$ and hence $f \in C^2(\mathbb{R}^d)$. Fix $x_1,\dots,x_{d-1}$. Then $\partial_d\psi = 1 - \beta'(x_d)\Lambda(x) \ge 1 - \frac{3\sqrt{3}}{8}\cdot 2a = \sqrt{\mu_0/2} > 0$, so ψ is strictly increasing in x_d; moreover, since $|\beta\Lambda| \le 2a$, $\psi > 0$ for $x_d > 2a$ and $\psi < 0$ for $x_d < -2a$. Hence ψ has a unique zero in x_d, which lies in $[-2a, 2a]$ and depends continuously on $(x_1,\dots,x_{d-1})$ by the implicit function theorem; its zero set is the valley floor. Finally, $f \ge 0$ with equality if and only if $\psi = 0$ and $x_1 = 0$, and $\Lambda = 0$ when $x_1 = 0$ because $b(-k) = 0$ for $k \ge 2$; hence the minimizers are the points with $x_1 = 0$ and $x_d = 0$, and $f^* = 0$.

(ii) Since $\nabla\psi = (1-\beta'\Lambda)e_d - \beta\nabla\Lambda$ and $\nabla\Lambda \perp e_d$,

$$\|\nabla\psi\|^2 = (1-\beta'\Lambda)^2 + \beta^2\|\nabla\Lambda\|^2 \in \left[\tfrac{\mu_0}{2},\ \left(1+\tfrac{3\sqrt{3}}{4}a\right)^2 + G_{a,\omega}^2\right], \tag{25}$$

where the lower bound uses $(\partial_d\psi)^2 = (1-\beta'\Lambda)^2 \ge \mu_0/2$, and the upper bound uses $|\beta'\Lambda| \le \frac{3\sqrt{3}}{4}a$, $\beta \le 1$, and (24). Since $\nabla f = 2\psi\nabla\psi + \kappa x_1 e_1$,

$$\begin{aligned}
\|\nabla f\|^2 &= 4\psi^2\|\nabla\psi\|^2 + 4\kappa x_1\psi\,\partial_1\psi + \kappa^2 x_1^2 \ge 4\psi^2(\partial_d\psi)^2 + (2\psi\,\partial_1\psi + \kappa x_1)^2\\
&\ge 2\mu_0\psi^2 + \tfrac{\kappa^2}{4}x_1^2 - \tfrac{4}{3}\psi^2(\partial_1\psi)^2,
\end{aligned}$$

where the last step uses $(A+B)^2 \ge \frac{1}{4}B^2 - \frac{1}{3}A^2$. By Lemma B.1, $|\partial_1\psi| = \beta|\partial_1\Lambda| \le 2a\sup|b'| = \frac{192}{25\sqrt{5}}a < 3.44a$, so $\frac{4}{3}\psi^2(\partial_1\psi)^2 < \frac{4}{3}(3.44a)^2\psi^2 < 16a^2\psi^2$ and

$$\|\nabla f\|^2 \ge (2\mu_0 - 16a^2)\psi^2 + \tfrac{\kappa^2}{4}x_1^2 \ge 2\mu\big(\psi^2 + \tfrac{\kappa}{2}x_1^2\big) = 2\mu(f - f^*)$$

with $\mu = \min\{\mu_0 - 8a^2, \kappa/4\}$, which is (6); here $\mu > 0$ because $a < \frac{4}{8+3\sqrt{3}}$ is equivalent to $\mu_0 > 8a^2$.

Moreover, we have $\nabla^2 f = 2\nabla\psi\nabla\psi^\top + 2\psi\nabla^2\psi + \kappa e_1 e_1^\top$ and

$$\nabla^2\psi = -\beta\nabla^2\Lambda - \beta'\big(\nabla\Lambda e_d^\top + e_d\nabla\Lambda^\top\big) - \beta''\Lambda e_d e_d^\top.$$

The middle matrix has the nonzero eigenvalues $\pm\|\nabla\Lambda\|$, so its trace norm is at most $2G_{a,\omega}$ by (24); hence

$$\|\nabla^2\psi\|_* \le \beta K_{a,\omega} + 2|\beta'|G_{a,\omega} + 2a|\beta''|.$$

Together with $|\psi| \le |x_d| + \beta|\Lambda| \le |x_d| + 2a$ and (23),

$$\begin{aligned}|\psi|\|\nabla^2\psi\|_* &\le (|x_d\beta|K_{a,\omega} + 2|x_d\beta'|G_{a,\omega} + 2a|x_d\beta''|) + 2a(\beta K_{a,\omega} + 2|\beta'|G_{a,\omega} + 2a|\beta''|)\\ &\le \Big(\tfrac{K_{a,\omega}}{2} + G_{a,\omega} + 2a\Big) + 2a\Big(K_{a,\omega} + \tfrac{3\sqrt{3}}{4}G_{a,\omega} + 4a\Big),\end{aligned}$$

where β and its derivatives are evaluated at x_d. Combining this with (25) gives $\|\nabla^2 f\|_* \le 2\|\nabla\psi\|^2 + 2|\psi|\|\nabla^2\psi\|_* + \kappa \le H$, which is (3). □

The comparison with envelope conditions in Section 5 uses the following elementary relation between an envelope and the trace norm.

Lemma B.4 (A common envelope bounds the trace norm)**.** *Let $A \succeq 0$ and let B be symmetric with $-A \preceq B \preceq A$. Then $\|B\|_* \le \mathrm{Tr}(A)$.*

Proof. Write $B = B_+ - B_-$ with $B_\pm \succeq 0$ and $B_+B_- = 0$, and let $P_\pm$ be the orthogonal projectors onto the ranges of $B_\pm$, so that $P_+ + P_- \preceq I_d$. Since $A \succeq 0$, $A \succeq B$, and $A \succeq -B$,

$$\begin{aligned}\mathrm{Tr}(A) &\ge \mathrm{Tr}(P_+AP_+) + \mathrm{Tr}(P_-AP_-) \ge \mathrm{Tr}(P_+BP_+) + \mathrm{Tr}(P_-(-B)P_-)\\ &= \mathrm{Tr}(B_+) + \mathrm{Tr}(B_-) = \|B\|_*.\end{aligned}$$

□

We now prove the two-sided bound on the trace of a common envelope on the Twisted Valley.

Proof of Proposition 5.2. (i) Let K be the set of integers $k \in \{2,\dots,d-1\}$ with $|k - x_1| \le \rho$. If $[x_1-\rho, x_1+\rho] \supseteq [2, d-1]$, then $|K| = d-2$. Otherwise $x_1 - \rho \ge 2$ or $x_1 + \rho \le d-1$, and since $2 \le x_1 \le d-1$ the set $[x_1-\rho, x_1+\rho] \cap [2, d-1]$ contains $[x_1-\rho, x_1]$ or $[x_1, x_1+\rho]$, an interval of length ρ, which contains at least $\lfloor\rho\rfloor$ integers. Hence $|K| \ge \min\{\lfloor\rho\rfloor, d-2\}$.

Fix $k \in K$ and let $y \in \mathbb{R}^d$ be the point with $y_1 = k$, with y_k the point of $\frac{\pi}{\omega}\mathbb{Z}$ closest to x_k, with $y_d = 0$, and with $y_j = x_j$ for the remaining j. Then

$$\|y - x\|^2 = (k - x_1)^2 + (y_k - x_k)^2 + x_d^2 \le \rho^2 + \frac{\pi^2}{4\omega^2} + x_d^2 = r^2,$$

so $\nabla^2 f(y) \preceq A$ by hypothesis. Since $b(k-j) = 0$ for every integer $j \ne k$ and $b(0) = 1$, $b'(0) = 0$, we have $\Lambda(y) = a\sin(\omega y_k) = 0$, $\partial_1\Lambda(y) = a\sin(\omega y_k)b'(0) = 0$, $\partial_k\Lambda(y) = a\omega\cos(\omega y_k) = \pm a\omega$, and $\partial_j\Lambda(y) = 0$ for the remaining j. Hence $\psi(y) = y_d - \beta(0)\Lambda(y) = 0$, $\partial_d\psi(y) = 1 - \beta'(0)\Lambda(y) = 1$, and $\nabla\psi(y) = e_d \mp a\omega\, e_k$. Because $\psi(y) = 0$, the Hessian of $f = \psi^2 + \frac{\kappa}{2}x_1^2$ at y is

$$\nabla^2 f(y) = 2\nabla\psi(y)\nabla\psi(y)^\top + \kappa e_1 e_1^\top, \qquad e_k^\top\nabla^2 f(y) e_k = 2a^2\omega^2.$$

The relation $\nabla^2 f(y) \preceq A$ gives $A_{kk} \ge e_k^\top\nabla^2 f(y) e_k = 2a^2\omega^2$. Since $A \succeq 0$, every diagonal entry of A is nonnegative, so we obtain

$$\mathrm{Tr}(A) \ge \sum_{k\in K} A_{kk} \ge 2a^2\omega^2|K|,$$

which is (9).

(ii) Let

$$J \coloneqq \{1, d\} \cup \{k \in \{2,\dots,d-1\} : |k - x_1| < r + 1\}$$

and $P_J \coloneqq \sum_{j\in J} e_j e_j^\top$. The open interval $(x_1 - r - 1, x_1 + r + 1)$ contains at most $2r + 3$ integers, so $|J| \le \min\{2r + 5, d\}$. Let $\|y - x\| \le r$. By Lemma B.3, $\partial_j \Lambda(y) = 0$ and $\partial_{ij}\Lambda(y) = 0$ for all i whenever $j \notin \{1\} \cup I(y_1)$; since $\psi = y_d - \beta(y_d)\Lambda$ and $f = \psi^2 + \frac{\kappa}{2} y_1^2$, every row and column of $\nabla^2 f(y)$ outside $\{1, d\} \cup I(y_1)$ vanishes, and $I(y_1) \subseteq J$ because $|y_1 - k| < 1$ implies $|k - x_1| < 1 + r$; hence $\nabla^2 f(y) = P_J \nabla^2 f(y) P_J$. By Proposition 5.1, $\|\nabla^2 f(y)\|_{\mathrm{op}} \le \|\nabla^2 f(y)\|_* \le H$, so for every $z \in \mathbb{R}^d$, we have

$$\left|z^\top \nabla^2 f(y) z\right| = \left|(P_J z)^\top \nabla^2 f(y)(P_J z)\right| \le H \|P_J z\|^2 = H z^\top P_J z.$$

Thus $A \coloneqq H P_J$ satisfies $-A \preceq \nabla^2 f(y) \preceq A$ for all such y, and $\mathrm{Tr}(A) = H|J| \le H \min\{2r + 5, d\}$. □

C Transverse Gap in the Twisted Valley Experiment

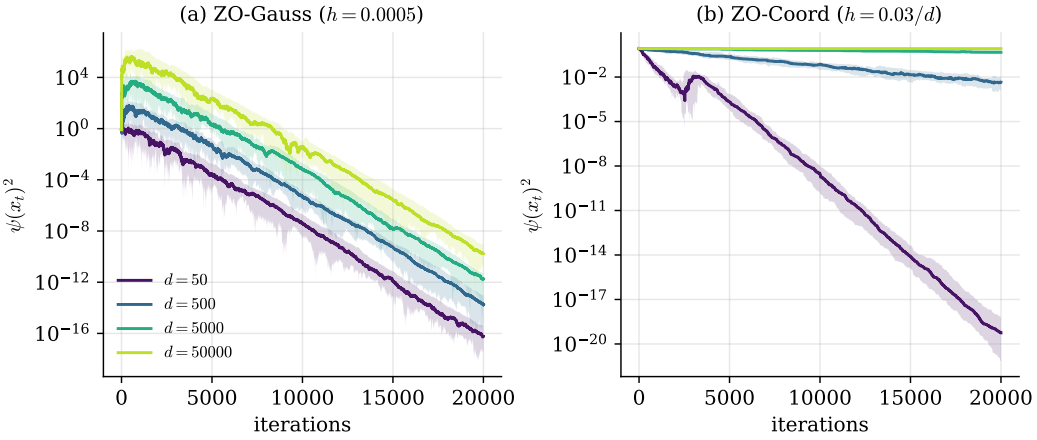

Figure 5: **Transverse gap $\psi(x_t)^2$ along the runs of Figures 2 and 3** (medians over 20 runs, bands the 10th to 90th percentiles).

Figure 5 shows the first term of the Twisted Valley (8) along the runs of Section 6.1, that is, the squared distance of the iterate from the valley floor in the transverse coordinate. While the clock coordinate is large, the gradient κx_1 of the quadratic term enters the two-point estimate and moves x_d by a random amount of order $h\kappa|x_1|$ per iteration. The transverse gap of ZO-Gauss therefore first grows, to about 4×10^5 at $d = 50000$, and then decreases as x_1 decreases. The curves for the four dimensions are shifted by the same number of iterations as the optimality gaps in Figure 2, and at $T = 20000$ the transverse gap is between 6×10^{-17} ($d = 50$) and 2×10^{-10} ($d = 50000$). For ZO-Coord the transverse gap decreases to machine precision only at $d = 50$; it ends at 4.5×10^{-3} at $d = 500$ and stays above 0.4 for $d \ge 5000$.

D Baselines on the Twisted Valley

Figure 6 compares ZO-Gauss with further two-point methods in the setting of Section 6.1, from $x_1 = d - 1$, at $2T = 4 \times 10^4$ function queries. The control run applies ZO-Gauss with the step size of Figure 2 to the function with $\Lambda \equiv 0$, that is, $f(x) = x_d^2 + \frac{\kappa}{2} x_1^2$, whose valley does not rotate. The averaged Gaussian method averages $k = 10$ independent two-point Gaussian estimators per iteration and spends $2k$ queries per iteration. The sparse perturbation method is a uniformly random variant of the magnitude-based mask of Sparse MeZO (Liu et al., 2025). It draws a uniformly random set S of k coordinates and a Gaussian vector u_S supported on S, and uses the unbiased estimator $\frac{d}{k} \frac{f(x+\nu u_S) - f(x-\nu u_S)}{2\nu} u_S$, here with $k = d/10$ and with $k = 10$. ZO-Coord is the case $k = 1$ with u_S replaced by a uniformly random basis vector e_i. ZO-Coord uses $h = 0.3/d$, ten times the step size of Figure 3. For every other method except the two runs with $h = 5 \times 10^{-4}$, the step size is

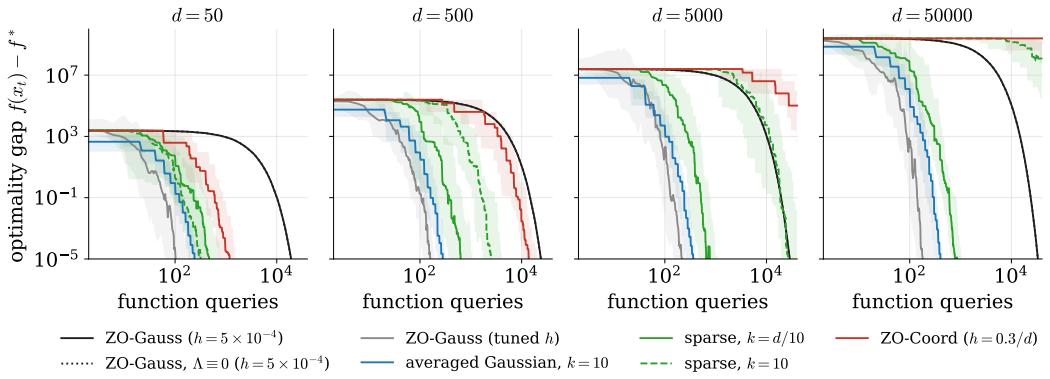

Figure 6: **Baselines on the Twisted Valley** in the setting of Section 6.1, from $x_1 = d - 1$, at equal numbers of function queries (medians over 20 runs, bands the 10th to 90th percentiles; logarithmic query axis). Black: ZO-Gauss with the step size 5×10^{-4} of Figure 2 (solid) and the same on the function with $\Lambda \equiv 0$ (dotted, hidden behind the solid line); gray: ZO-Gauss with the tuned step size; blue: averaged Gaussian with $k = 10$; green: sparse perturbation with $k = d/10$ (solid) and $k = 10$ (dashed); red: ZO-Coord with $h = 0.3/d$.

the value with the smallest median final gap in the grid $\{10^{-4}, 3 \times 10^{-4}, \dots, 1\}$, multiplied by k/d for the sparse perturbation method. ZO-Coord cannot share the step size 5×10^{-4} of ZO-Gauss. Its update along the sampled axis is hd times the finite difference, so the quadratic term of the Twisted Valley alone requires $hd\kappa < 2$, and for $d \geq 5000$ the iterates diverge as soon as the first axis is sampled.

The control coincides with ZO-Gauss at every d, so the rotation of the valley does not slow ZO-Gauss at the step size of Figure 2. With the tuned step size, ZO-Gauss reaches the gap 10^{-12} within 200 to 330 queries for every d, the averaged Gaussian method within 450 to 600, and the sparse perturbation method with $k = d/10$ within 900 to 1200. The step size 5×10^{-4} of Figure 2 is about $\frac{1}{15H}$ for the bound $H < 133$ of Proposition 5.1, a bound far above the curvature met along the trajectory. The sparse perturbation method with $k = 10$ and ZO-Coord probe the first axis with probability k/d per iteration, and their final gaps grow with d. Both reach machine precision at $d = 50$ and $d = 500$, but at $d = 5000$ they end at 1.6×10^{-9} and 1.0×10^{5}, respectively, and at $d = 50000$ at 1.1×10^{8} and 2.5×10^{9}. What separates the methods is thus whether the probe covers every direction with a probability that does not depend on d, not the rotation of the curvature.

E Empirical Diagnostic: Hessian Spectra of the Adapted Models

We examine the curvature of the two models adapted in Section 6.2 at $d = 2453$. Because the backbone is frozen and the features are precomputed, the target training loss is an explicit function of the adapter and head parameters, and its exact Hessian on the 2000-image training subset is computed by automatic differentiation. We evaluate it at the ZO-Gauss iterate x after 4000 queries (seed 0) and at the iterate y reached 200 queries later, for which $\|y - x\| = 2.15$ (CNN) and 0.85 (ViT), compared with $\|x\| = 10.4$ and 7.0.

Figure 7 shows the absolute spectrum at x. The trace norm is 156 (CNN) and 60 (ViT), that is, 5.5 and 11.9 times the operator norm, whereas a flat spectrum would give the ratio $d = 2453$; equivalently, the trace norm at x is smaller than $d \, \|\nabla^2 f\|_{\mathrm{op}}$ by a factor of about 450 and 200. The spectrum has no gap: beyond the leading three eigenvalues, which carry 47% and 22% of the absolute mass, it decays roughly like $|\lambda_1|/j$, and 80% of the mass requires 56 and 218 eigenvalues. The curvature is therefore not low rank; what is small is the sum of the eigenvalues relative to d times the largest one, which is the quantity the trace norm measures. The Hessians are indefinite, with 616 and 710 negative eigenvalues out of 2453, so the two-sided form of Assumption 3.2 is needed. Between x and y the principal directions move moderately. With the columns of U_x and U_y

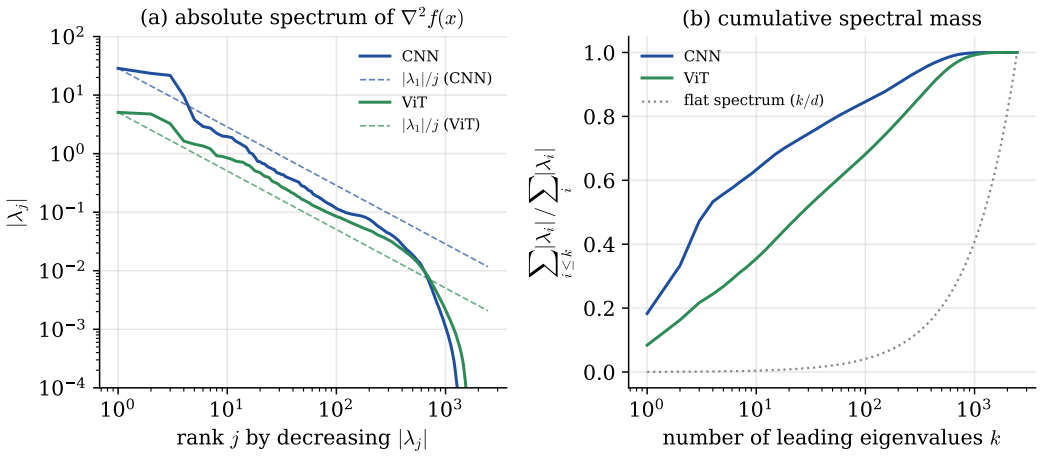

Figure 7: **Absolute spectrum and cumulative spectral mass** of the exact Hessian $\nabla^2 f(x)$ of the target training loss at the ZO-Gauss iterate x after 4000 queries, for the two models of Section 6.2 at $d = 2453$; dashed lines show a $1/j$ decay from the largest eigenvalue, the dotted line a flat spectrum.

orthonormal bases of the leading k-dimensional eigenspaces at x and y, the overlap $\|U_x^\top U_y\|_{\mathrm{F}}^2/k$ decreases from 0.88 ($k = 5$) to 0.54 ($k = 100$) for the CNN and from 0.99 to 0.87 for the ViT. These measurements at two points of one trajectory do not verify the bound of Assumption 3.2 at every point of the trajectory, and they do not identify curvature rotation as the cause of the performance gap in Figure 4.

F Details of the CIFAR-10 Experiment

Table 1: **Adapter width sweep.** Median final training loss and target test accuracy over five seeds, selected step sizes (h of Algorithm 1 for ZO-Gauss and of (10) for ZO-Coord), and norms of the exact Hessian of the training loss at initialization (x_0) and at the final ZO-Gauss iterate (x_T, seed 0).

		ZO-Gauss		ZO-Coord		$\|\nabla^2 f\|_*$		$\|\nabla^2 f\|_{\mathrm{op}}$	
Model	r / d	loss / acc	h	loss / acc	h	x_0	x_T	x_0	x_T
CNN	2 / 647	0.86 / 65.6%	3×10^{-3}	0.92 / 64.3%	10^{-3}	171	114	26.5	29.6
CNN	8 / 1421	0.90 / 64.4%	3×10^{-3}	1.05 / 59.4%	10^{-3}	330	140	42.9	31.2
CNN	16 / 2453	0.91 / 63.7%	3×10^{-3}	1.05 / 54.3%	3×10^{-4}	426	156	36.7	28.6
CNN	32 / 4517	0.91 / 64.2%	10^{-3}	1.21 / 50.7%	3×10^{-4}	556	177	42.6	33.9
CNN	64 / 8645	0.89 / 65.1%	10^{-3}	1.41 / 49.3%	3×10^{-4}	1001	220	37.0	36.8
ViT	2 / 647	0.88 / 65.2%	3×10^{-3}	0.91 / 63.0%	3×10^{-3}	79	35	4.2	4.2
ViT	8 / 1421	0.87 / 65.3%	3×10^{-3}	0.98 / 61.6%	3×10^{-3}	130	45	4.6	5.1
ViT	16 / 2453	0.88 / 64.6%	3×10^{-3}	1.02 / 59.6%	10^{-3}	199	60	5.1	5.1
ViT	32 / 4517	0.87 / 65.1%	3×10^{-3}	1.03 / 59.1%	10^{-3}	364	85	5.4	6.6
ViT	64 / 8645	0.90 / 63.9%	3×10^{-3}	1.15 / 57.1%	10^{-3}	384	143	6.7	9.4

The source and target tasks are the classes 0–4 and 5–9 of CIFAR-10. The backbones are a three-layer CNN and a Vision Transformer with patch size 4, width 64, and four blocks (ViT), pre-trained on the source task with first-order optimization and frozen for the target task. On the target task, a residual adapter $z \mapsto z + W_2 \tanh(W_1 z + b_1) + b_2$ with $W_1 \in \mathbb{R}^{r\times 64}$ and $W_2 \in \mathbb{R}^{64\times r}$, initialized as the identity by $W_2 = 0$, and a new five-way linear head are trained, which gives $d = 129r + 389$ trainable parameters. A function query is the cross-entropy on a minibatch of 128 images drawn from a fixed subset of 2000 target training images, shared by the two evaluations of a two-point estimator, with $\nu = 10^{-3}$. Table 1 lists the numbers behind Figure 4, whose panels (c) and (f) show

the Hessian norms against d. The pre-trained backbones reach 81.7% (CNN) and 76.1% (ViT) test accuracy on the source task. For every configuration (model, r, method), the step size is chosen by the final training loss of a seed-0 pilot run from the grid $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$, and five seeds are then run for 4000 queries. The step size selected for ZO-Coord is never larger than that of ZO-Gauss, since its update of hd times the finite difference is concentrated on a single axis. The exact Hessian of the training loss on the 2000-image subset is computed by automatic differentiation at the initial point and at the final ZO-Gauss iterate of seed 0. At initialization the trace norm is between 1.5 and 4.6 times larger than at the final iterate, since the randomly initialized adapter weights contribute curvature in every direction. The Hessian is measured at only two points of one trajectory, so these measurements neither verify Assumption 3.2, which concerns every point, nor determine its constant H.

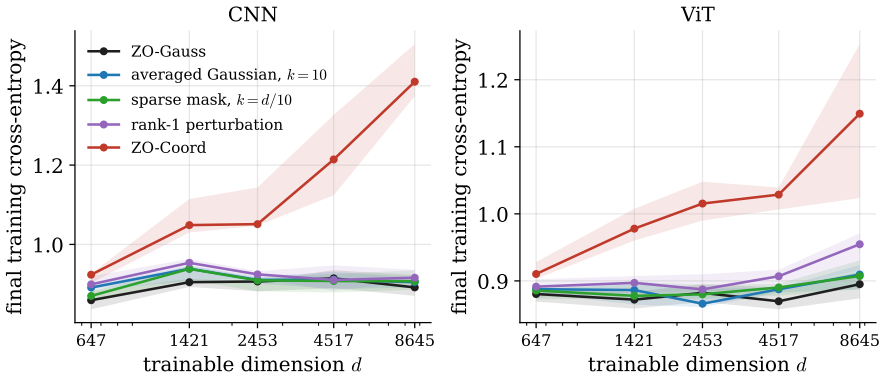

Figure 8: **Baselines in the adapter width sweep.** Median final training cross-entropy over five seeds against the trainable dimension d, with 10th to 90th percentile bands, for ZO-Gauss and ZO-Coord of Figure 4 and for the averaged Gaussian, sparse perturbation, and rank-one perturbation methods.

Figure 8 adds three baselines to the sweep, each with the step size chosen by the same pilot procedure from the grid $\{10^{-4}, 3\times10^{-4}, \dots, 10^{-1}\}$: the averaged Gaussian method with $k = 10$ directions and the sparse perturbation method with $k = d/10$, both as in Appendix D, and a rank-one perturbation. The last is the perturbation of the low-rank ZO method of Chen et al. (2025), which is distinct from DeepZero (Chen et al., 2024); it perturbs every weight matrix by $uv^\top$ with Gaussian u and v and every bias by a Gaussian vector, so that the perturbation has identity second moment as for ZO-Gauss. Their final training losses stay between 0.86 and 0.95 on the CNN and between 0.87 and 0.96 on the ViT over the whole range of d, as for ZO-Gauss, whereas that of ZO-Coord grows from 0.92 to 1.41 and from 0.91 to 1.15.

Reference runs. Table 2 and Figure 9 report three reference runs with the protocol above; they address whether the insensitivity of ZO-Gauss to d in Figure 4 reflects an ability to use the additional dimensions or only their inactivity. First, in the head-only model ($r = 0$, $d = 389$), in which the adapter reduces to a bias on the features, ZO-Gauss reaches 0.88 / 64.5% on the CNN and 0.88 / 64.8% on the ViT. These values are close to those it reaches for every r, so within the budget the adapter directions add nothing to the loss ZO-Gauss attains. Second, Adam on the same parameterization, with the learning rate chosen from the same grid by the same pilot procedure, is run for 2000 steps (the same number of passes over the data as 4000 queries) and for 20000 steps. It lowers the training loss monotonically in r and reaches zero on the 2000 training images for $r \geq 32$ (after 20000 steps for the CNN), while its test accuracy does not improve with r; the adapter directions therefore carry signal that a first-order method uses, and ZO-Gauss does not exploit it within the budget. ZO-Gauss does move the adapter: at the pilot step size, the displacement $\|W_1(x_T) - W_1(x_0)\|$ of the first adapter layer is 1.9 to 5.9 on the CNN and 1.0 to 8.4 on the ViT, comparable to or larger than that of the head, 1.5 to 3.3. At a fixed step size, however, this displacement grows roughly with the square root of the number of adapter parameters, as for a random walk, and the adapter lowers the loss no further than the head alone; note that the gradient with respect to W_1 vanishes at the identity initialization $W_2 = 0$. Third, with H_r the trace norm at initialization, the step size

Table 2: **Reference runs for the adapter width sweep.** Median final training loss and target test accuracy over five seeds for ZO-Gauss at the pilot step size (the values of Table 1, plus the head-only model $r = 0$), for ZO-Gauss at the step size $h = 1/(10H_r)$, proportional to that of Theorem 4.1, where H_r is the trace norm of the Hessian at initialization, and for Adam with minibatch 128 after 2000 and 20000 steps.

		ZO-Gauss, pilot h	ZO-Gauss, $h = 1/(10H_r)$			Adam	
Model	r / d	loss / acc	H_r	h	loss / acc	2000 steps	20000 steps
CNN	0 / 389	0.88 / 64.5%	–	–	–	0.66 / 71.7%	0.64 / 71.2%
CNN	2 / 647	0.86 / 65.6%	171	5.9×10^{-4}	0.94 / 63.0%	0.63 / 71.5%	0.57 / 70.3%
CNN	8 / 1421	0.90 / 64.4%	330	3.0×10^{-4}	1.07 / 54.6%	0.51 / 69.2%	0.39 / 65.9%
CNN	16 / 2453	0.91 / 63.7%	426	2.4×10^{-4}	1.08 / 55.7%	0.41 / 67.3%	0.16 / 63.1%
CNN	32 / 4517	0.91 / 64.2%	556	1.8×10^{-4}	1.09 / 56.5%	0.18 / 65.9%	0.00 / 64.1%
CNN	64 / 8645	0.89 / 65.1%	1001	1.0×10^{-4}	1.21 / 47.5%	0.06 / 66.3%	0.00 / 65.7%
ViT	0 / 389	0.88 / 64.8%	–	–	–	0.72 / 69.2%	0.71 / 68.8%
ViT	2 / 647	0.88 / 65.2%	79	1.3×10^{-3}	0.94 / 62.0%	0.65 / 65.8%	0.62 / 64.9%
ViT	8 / 1421	0.87 / 65.3%	130	7.7×10^{-4}	0.99 / 60.1%	0.45 / 62.3%	0.30 / 59.1%
ViT	16 / 2453	0.88 / 64.6%	199	5.0×10^{-4}	1.05 / 57.9%	0.20 / 60.6%	0.02 / 59.0%
ViT	32 / 4517	0.87 / 65.1%	364	2.7×10^{-4}	1.13 / 53.2%	0.00 / 61.7%	0.00 / 60.4%
ViT	64 / 8645	0.90 / 63.9%	384	2.6×10^{-4}	1.12 / 53.7%	0.00 / 64.2%	0.00 / 63.3%

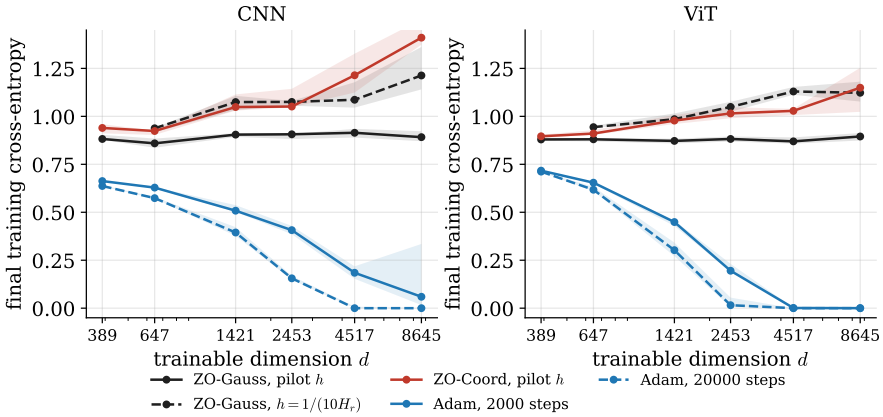

Figure 9: **Reference runs for the adapter width sweep.** Median final training cross-entropy over five seeds against d, with 10th to 90th percentile bands, for the runs of Table 2; $d = 389$ is the head-only model.

$h = 1/(10H_r)$, proportional to that of Theorem 4.1, makes the final loss of ZO-Gauss increase with H_r, close to that of ZO-Coord at its pilot step size. This is the behavior of the rate $\mathcal{O}(H/T)$ of Theorem 4.1, whose iteration count at a step size proportional to $1/H$ is proportional to H. It also shows that the pilot step sizes 10^{-3} to 3×10^{-3} of Table 1, which exceed $1/(10H)$ even for the trace norms at the final iterate, $H \geq 114$ (CNN) and $H \geq 35$ (ViT), lie outside the regime of the theorem. The sweep thus shares a feature of the Twisted Valley: most coordinates carry no signal that ZO-Gauss exploits within the budget, and they do not slow ZO-Gauss, whereas ZO-Coord samples them with probability $1 - 389/d$ per iteration. Unlike in the Twisted Valley, however, the set of about 389 active coordinates is fixed.