WHEN DOES SYMMETRY REDUCE EFFECTIVE RANK IN ZEROTH-ORDER OPTIMIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning large neural networks is often limited by the substantial memory overhead of backpropagation. To address this bottleneck, memory-efficient zeroth-order optimization (MeZO) enables fine-tuning using only function evaluations, but classical zeroth-order optimization (ZO) query-complexity bounds typically scale with the ambient parameter dimension. Recent theory suggests that ZO efficiency can instead depend on the effective rank, a spectral measure of how curvature is distributed across parameter directions. In this work, we show that, for smooth objectives with appropriate symmetries, the effective rank of the Hessian is bounded by the codimension of the symmetry orbit together with an explicit correction term that quantifies nonstationarity. Our bound is tight, with equality achieved by an explicit family of smooth invariant objectives. At stationary points, it recovers the exact Hessian null directions induced by symmetry. Lastly, we establish finite-sample convergence of ZO-SGD under mild conditions and obtain dimension-independent function-query bounds without assuming a dominate low-effective-rank Hessian matrix. In experiments, empirical evidences are provided to illustrate the theoretical mechanism and the connection between effective rank and ZO optimization.

1 INTRODUCTION

Fine-tuning large neural networks is often constrained by the memory required for backpropagation. This has motivated memory-efficient zeroth-order (ZO) fine-tuning, which estimates parameter updates using forward function evaluations and avoids retaining activations for backpropagation (Malladi et al., 2023; Zhang et al., 2024; Chen et al., 2025; Liu et al., 2025a; Shang et al., 2026). Yet classical complexity analyses and minimax lower bounds show that ZO query complexity can grow with the ambient parameter dimension d over broad convex problem classes (Nesterov & Spokoiny, 2017; Jamieson et al., 2012; Shamir, 2013; Duchi et al., 2015; Shamir, 2017; Yu et al., 2024). The practical effectiveness of ZO methods in such high-dimensional settings suggests a gap between worst-case dimension dependence and the effective structure of large-model fine-tuning objectives.

Existing theory suggests that the *local effective rank* provides an informative characterization of the dimension dependence of ZO methods. In particular, Malladi et al. (2023) establish effective-rank-dependent convergence guarantees for memory-efficient zeroth-order optimizer (MeZO) under a low-effective-rank matrix smoothness assumption. A growing body of empirical and theoretical work also provides evidence of low-dimensional structure in neural Hessians (Sagun et al., 2017; Ghorbani et al., 2019; Papyan, 2020; Singh et al., 2021). However, these observations do not establish when the required low-effective-rank condition holds, leading to our central question:

> *What properties of neural networks give rise to the low effective rank observed during ZO fine-tuning?*

To answer this question, we observe that modern neural architectures, including Transformers, contain substantial parameter redundancy, with many distinct parameter configurations representing the same function (as illustrated in Examples 3 and 4). A structured source of such redundancy is the *continuous parameter symmetry*, under which transformations of the parameters leave the represented function unchanged. We show that continuous parameter symmetries provide a direct

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.

mechanism for low-dimensional curvature structure. Specifically, we consider smooth objectives that are invariant under a smooth action of a Lie group G:

$$f(a \cdot \theta) = f(\theta), \qquad a \in G.$$

Such symmetries generate continuous families of equivalent parameters, creating redundant directions that lie in the Hessian kernel at stationary points (Neyshabur et al., 2015; Absil et al., 2008; Simsek et al., 2021).

However, parameter symmetry alone does not imply low effective rank or faster ZO convergence. The rotationally invariant quadratic in Example 1 has codimension-one symmetry orbits away from the origin, yet retains Hessian effective rank d and dimension-dependent spherical ZO convergence. Addressing this gap motivates our main contributions:

1. **A sharp characterization of effective rank induced by symmetry.** We prove that, smooth invariant objectives (Assumption 1), the effective rank near a stationary point is bounded by the codimension of the symmetry orbit at that point, together with a non-stationarity correction term (Theorem 1). This correction vanishes at stationarity, implying that the effective rank is heavily determined by the codimension of its symmetry orbit at stationarity. Moreover, we show that this bound is tight by constructing a smooth invariant objective that attain this bound (Proposition 2).
2. **A finite-sample convergence bound for ZO-SGD.** Under the boundedness assumption (Assumption 2) and the Polyak–Łojasiewicz (PL) condition (Assumption 3), we derive finite-sample convergence error bounds for the classical ZO-SGD method (Theorem 2 and corollary 1). Importantly, these bounds depend on the symmetry-induced effective rank rather than the ambient dimension d, thereby capturing the dimension reduction induced by symmetry.
3. **Synthetic and real-data optimization experiments.** We empirically validate the proposed mechanism through both synthetic and real-data optimization experiments (Section 5). Specifically, we examine how the effective rank evolves along ZO training and provide empirical evidence for the progression

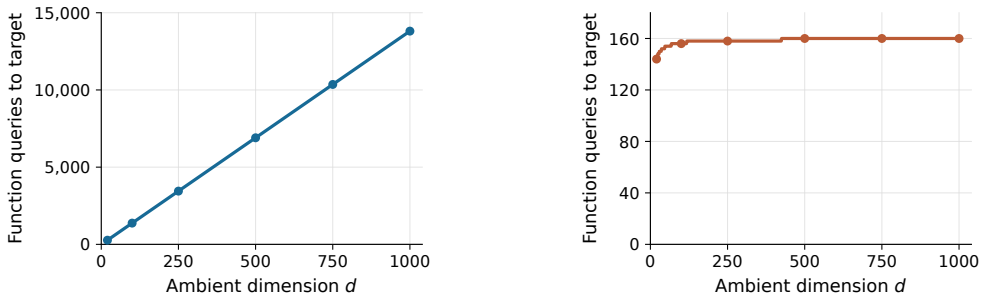
   $$\text{ZO training} \implies \text{approach to stationarity} \implies \text{lower effective rank}.$$

   This behavior is consistently observed in both the radial quartic objective and the transformer neural network experiment, providing empirical support for the connection between effective rank and ZO optimization.

2 Related Work

ZO applications in high-dimensional optimization. Zeroth-order optimization has enabled memory-efficient adaptation of large language models using only forward evaluations. MeZO demonstrates that models with billions of parameters can be fine-tuned with memory requirements comparable to inference (Malladi et al., 2023), and Zhang et al. (2024) benchmark a range of ZO optimizers across model families, tasks, and fine-tuning schemes. Subsequent methods improve ZO fine-tuning through different forms of structure: Sparse MeZO perturbs selected parameter subsets (Liu et al., 2025a), LOZO constructs low-rank gradient estimates (Chen et al., 2025), and SubZero estimates updates in random subspaces (Yu et al., 2025). Other approaches combine multiple queries with momentum (Peng et al., 2025) or adapt ZO updates to quantized models (Zhou et al., 2025). These developments motivate understanding which properties of high-dimensional fine-tuning objectives permit efficient optimization from function evaluations.

ZO complexity and lower bounds. Classical analyses of randomized gradient-free methods yield query bounds that depend explicitly on the ambient dimension d, in both convex and nonconvex settings (Nesterov & Spokoiny, 2017; Ghadimi & Lan, 2013). For Euclidean stochastic convex optimization with paired evaluations of the same sampled function, optimal rates give a query complexity of $\Theta(d/\epsilon^2)$ for expected suboptimality ϵ, with fixed domain and Lipschitz parameters (Duchi et al., 2015; Shamir, 2017). Under noisy function-value oracles, lower bounds also establish dimension-dependent costs, including quadratic dependence on d for smooth strongly convex classes (Shamir, 2013). These worst-case results concern specific function classes and oracle models. Refined analyses exploit additional curvature structure: Yue et al. (2023) obtain weaker dimension



(a) Example 1: $f(\theta) = \frac{1}{2}\|\theta\|^2$. (b) Example 2: $f(\theta) = \frac{1}{2}\|P(\theta - \theta^\star)\|^2$.

Figure 1: Query complexity for reaching $\mathbb{E}f(\theta_T) \leq 10^{-3} f(\theta_0)$ using ZO method. **(Left)** The rotationally invariant quadratic requires $\Theta(d\log(1/\epsilon))$ queries. **(Right)** The translation-invariant quadratic requires $O(r\log(1/\epsilon))$ queries and is independent of d.

dependence using spectral measures of the Hessian, while MeZO bounds the slowdown relative to SGD by the effective rank of a common Hessian upper matrix within each update neighborhood (Malladi et al., 2023). In our work, we aim to understand when the complexity of ZO optimization can be characterized by intrinsic structural quantities rather than the ambient dimension alone.

Symmetry and Hessian rank. Group actions and quotient spaces describe equivalent parameterizations in manifold optimization and matrix factorization (Lee, 2013; Absil et al., 2008; Mishra et al., 2014). In neural networks, positive homogeneity and continuous parameter symmetries produce families of functionally equivalent models (Neyshabur et al., 2015; Simsek et al., 2021; Grigsby et al., 2023). Empirical studies report neural Hessian spectra with a small number of large eigenvalues and a bulk concentrated near zero (Sagun et al., 2017; Ghorbani et al., 2019). Existing analyses connect spectral outliers to class structure in the data (Papyan, 2020) and derive architecture- and data-dependent Hessian rank bounds for deep linear networks (Singh et al., 2021). In our work, we aim to understand the effectiveness of ZO methods in high-dimensional optimization through the lenses of symmetry and Hessian rank.

3 Motivating Examples and Preliminaries

We consider minimizing the smooth objective $f(\theta) = \mathbb{E}_\xi[\mathcal{L}(\theta;\xi)]$, $\theta \in \mathbb{R}^d$, using only function evaluations. Throughout this work, we use the spherical two-point estimator

$$\widehat{g}_\mu = d\frac{\mathcal{L}(\theta+\mu v;\xi) - \mathcal{L}(\theta-\mu v;\xi)}{2\mu} v, \qquad \theta^+ = \theta - \eta\widehat{g}_\mu, \tag{1}$$

where $\mu > 0$, $v \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and v is independent of ξ. We are interested in how the continuous parameter symmetry will affect the performance of ZO-SGD using Equation (1).

3.1 Motivating Examples

We begin with two examples that isolate the role of parameter symmetry in zeroth-order optimization. Our first example illustrates that continuous parameter symmetry alone does not guarantee faster ZO convergence, even when the symmetry orbits have small codimension.

Example 1. *For $d \geq 2$, consider $f(\theta) = \|\theta\|^2/2$. This objective is invariant under rotations; that is, for an arbitrary orthogonal matrix O,*

$$f(O\cdot\theta) = \theta^\top O^\top O\theta/2 = \|\theta\|^2/2,$$

and each nonzero point lies on a symmetry orbit of dimension $d-1$ and codimension one. Nevertheless, its Hessian is I_d, with effective rank d. At the optimal constant step size $\eta = 1/d$, the update in Equation (1) satisfies

$$\mathbb{E}[f(\theta^+) \mid \theta] = (1 - 1/d) f(\theta).$$

Consequently, for $f(\theta_0) > 0$, reaching $\mathbb{E}f(\theta_T) \le \epsilon f(\theta_0)$ requires $\Theta(d\log(1/\epsilon))$ function queries, for $0 < \epsilon \le 1/2$. At the minimizer, the orbit is a point with codimension d.

The next example shows a contrasting regime in which continuous parameter symmetry induces globally flat directions, and the ZO convergence rate depends on the orbit codimension rather than the ambient dimension.

Example 2. *Let P be an orthogonal projector of rank $1 \le r < d$, and define*

$$f(\theta) = \tfrac{1}{2}\left\|P(\theta - \theta^\star)\right\|^2.$$

The objective is invariant under every translation $\theta \mapsto \theta + h$ with $h \in \ker P$. These orbits are affine subspaces of dimension $d - r$ and codimension r. The Hessian is P, with effective rank r. At the optimal constant step size $\eta = (d+2)/(d(r+2))$, the same spherical two-point update satisfies

$$\mathbb{E}[f(\theta^+) \mid \theta] = \left(1 - \frac{d+2}{d(r+2)}\right) f(\theta).$$

It therefore reaches $\mathbb{E}f(\theta_T) \le \epsilon f(\theta_0)$ in $O(r\log(1/\epsilon))$ function queries.

We illustrate these two examples in Figure 1. These examples motivate us to understand when continuous parameter symmetry translates into low effective rank and favorable dimension dependence in ZO optimization. To this end, we characterize the relationship between symmetry-induced tangent directions and the local curvature of the objective.

3.2 Functions with Continuous Parameter Symmetry

We begin by formalizing the class of smooth objectives with continuous parameter symmetry.

Assumption 1 (Smooth invariant objectives)**.** *On an open region $U \subseteq \mathbb{R}^d$, the objective f is C^2 and invariant under a smooth action of a Lie group G:*

$$f(a \cdot \theta) = f(\theta) \quad \text{whenever } \theta,\ a \cdot \theta \in U.$$

We refer to G as a continuous parameter symmetry of f on U.

Many modern architectures, including Transformers, admit continuous parameter symmetries. This property has been widely recognized in existing literature (Tran et al., 2025; Wang & Wang, 2026; Zhang et al., 2025; Liu et al., 2025b; Fishman et al., 2025; Li & Shen, 2026). We introduce following two examples to further elaborate this structure.

Example 3 (Gauge symmetry in Transformer attention)**.** *The contribution of one attention head (Vaswani et al., 2017) can be written as*

$$F(X) = A(X)(XW_V)W_O, \qquad W_V \in \mathbb{R}^{p\times h}, \quad W_O \in \mathbb{R}^{h\times p},$$

where $A(X)$ is the attention matrix determined by the query and key projections, and W_O is the corresponding block of the output projection. For any invertible $S \in \mathrm{GL}(h)$, the transformation

$$(W_V, W_O) \longmapsto (W_V S^{-1}, SW_O)$$

preserves $F(X)$ for every input. These transformations form a smooth group action and leave the full model's predictions unchanged when the remaining parameters are fixed.

Example 4 (Scaling symmetry in SwiGLU layers)**.** *A SwiGLU feed-forward block (Shazeer, 2020) has the form*

$$F(x) = \left[\phi(xW_g) \odot (xW_u)\right]W_d, \qquad \phi(t) = \frac{t}{1+e^{-t}},$$

where ϕ acts elementwise, $\odot$ denotes elementwise multiplication, and the hidden width is h. For any positive diagonal matrix $C = \mathrm{diag}(c_1, \dots, c_h)$, the transformation

$$(W_g, W_u, W_d) \longmapsto (W_g, W_u C, C^{-1}W_d)$$

leaves $F(x)$ unchanged: scaling the linear branch scales each hidden channel, and the inverse scaling in W_d cancels it. This gives a smooth action of the group $(\mathbb{R}_{>0})^h$.

To describe the geometry of these symmetries, let $\mathfrak{g}$ be the Lie algebra of G and define

$$\mathcal{O}_\theta = G \cdot \theta, \qquad \xi_X(\theta) = \frac{d}{dt}\Big|_{t=0} \exp(tX) \cdot \theta, \qquad T_\theta = \{\xi_X(\theta) : X \in \mathfrak{g}\}.$$

The map $X \mapsto \xi_X(\theta)$ is linear, and T_θ is the tangent space to the orbit $\mathcal{O}_\theta$ at θ. We write $m = \dim T_\theta$ and $k = d - m$ for the orbit dimension and codimension, respectively, and let P_T denote the orthogonal projector onto T_θ. For a linear action, $\xi_X(\theta) = A_X\theta$ and $D\xi_X(\theta) = A_X$. Write $g = \nabla f(\theta)$ and $M = \nabla^2 f(\theta)$. We use Euclidean vector norms and the matrix operator norm $\|\cdot\|_{\mathrm{op}}$.

Definition 1 (Effective rank)**.** *For a symmetric matrix H, its positive part is defined as $H_+ = V \operatorname{diag}(\max\{\lambda_i, 0\})V^\top$. Then the effective rank of H is defined as*

$$r_{\mathrm{eff}}(H) = \frac{\operatorname{tr}(H_+)}{\|H_+\|_{\mathrm{op}}} \qquad \text{if } H_+ \neq 0,$$

and set $r_{\mathrm{eff}}(H) = 0$ when $H_+ = 0$.

We note that a substantial body of existing work has studied this or closely related notions of effective rank, including the covariance effective rank (Koltchinskii & Lounici, 2017), intrinsic dimension (Tropp, 2015), and, in particular, the effective rank of local Hessian upper bounds in ZO analysis (Malladi et al., 2023). Our definition follows that of Malladi et al. (2023), with the modification of replacing the Hessian by H_+, which accommodates indefinite Hessians. In the next section, we derive an effective-rank bound for the Hessian in terms of the codimension of the symmetry orbit and discuss the distinction between Example 1 and Example 2.

4 Main Theoretical Results

In this section, we will establish an effective-rank bound at stationary points and revisit the distinction between Example 1 and Example 2 under Assumption 1; then we extend it to nonstationary points. After establishing tightness, we derive finite-sample convergence guarantees for ZO-SGD.

4.1 An Upper Bound on Effective Rank from Symmetry

We begin with discussing the orbit codimension induced by continuous parameter symmetry at stationary points. It is formalized in the following proposition.

Proposition 1. *Let f be C^2 and locally invariant under a smooth Lie-group action near $\theta_\star$. If $\nabla f(\theta_\star) = 0$, then*

$$T_{\theta_\star} \subseteq \ker \nabla^2 f(\theta_\star), \qquad \operatorname{rank}(\nabla^2 f(\theta_\star)) \leq k_{\theta_\star}.$$

In particular, $r_{\mathrm{eff}}((\nabla^2 f(\theta_\star))) \leq k_{\theta_\star}$.

Proof. Differentiating invariance and using stationarity shows that every orbit tangent vector lies in the Hessian kernel. The rank and effective-rank bounds follow; see Appendix A.2 for details. □

Remark 1. *Proposition 1 clarifies the contrast between Examples 1 and 2. Both objectives are smooth quadratics with continuous parameter symmetries and constant Hessians. The key difference lies in the geometry of their stationary orbits. In Example 1, the nonzero orbits are concentric spheres of dimension $d-1$. As their radius tends to zero, these spheres shrink to the singleton orbit $\{0\}$ at the unique stationary point, where no nonzero orbit-tangent directions remain. In Example 2, the translation orbits are parallel affine subspaces of dimension $d-r$. The stationary set $\theta^\star + \ker P$ is itself one entire orbit of the same dimension, whose $d-r$ tangent directions lie in the Hessian kernel. Table 1 summarizes these differences between the two examples.*

However, low effective rank at stationary points does not automatically present faster ZO convergence, since optimization updates generally occur at nonstationary points. We therefore examine how symmetry affects the effective rank away from stationarity. To quantify the remaining curvature, we differentiate invariance along an orbit and then in the base point, obtaining

$$\langle g, \xi_X(\theta)\rangle = 0, \qquad M\xi_X(\theta) = -[D\xi_X(\theta)]^\top g. \tag{2}$$

Table 1: Comparison of Examples 1 and 2. Here P is an orthogonal projector of rank $1 \le r < d$. Query bounds refer to reaching $\mathbb{E}f(\theta_T) \le \epsilon f(\theta_0)$ for $f(\theta_0) > 0$ and $0 < \epsilon \le 1/2$, using the respective optimal constant step sizes stated in the examples. Although both examples have orbit codimension strictly below d at *nonstationary points*, their orbit codimensions at *stationary points* is more important in explaining the different dimension dependence of the ZO query bounds.

Property	$\frac{1}{2}\|\theta\|^2$	$\frac{1}{2}\|P(\theta-\theta^\star)\|^2$
Symmetry action	$\theta \mapsto O\theta, \quad O \in \mathrm{O}(d)$	$\theta \mapsto \theta + h, \quad h \in \ker P$
Stationary set	$\{0\}$	$\theta^\star + \ker P$
Orbit codimension at nonstationary points	1	r
Orbit codimension at stationary points	d	r
Hessian effective rank, at every point	d	r
Function-query bound	$\Theta(d\log(1/\epsilon))$	$O(r\log(1/\epsilon))$

The second identity controls the entire Hessian image of a tangent vector, including its normal component. We use it to bound the tangent Hessian residual in terms of the gradient norm. For a nonstationary point, we define the generator conditioning

$$\kappa_\theta = \sup_{\substack{q \in T_\theta \\ \|q\|=1}} \inf_{\substack{X \in \mathfrak{g} \\ \xi_X(\theta)=q}} \|D\xi_X(\theta)\|_{\mathrm{op}}, \tag{3}$$

with $\kappa_\theta = 0$ if $m = 0$. Also define the measurable tangent Hessian residual $\delta_\theta = \|MP_T\|_{\mathrm{op}}$. Equation (2) implies $\delta_\theta \le \kappa_\theta \|g\|$. The following result presents the effective-rank bound near an invariant orbit:

Theorem 1. *Suppose Assumption 1 holds on an open neighborhood U of θ, with $\kappa_\theta < \infty$. Let $m = \dim T_\theta$, $k = d - m$, $H_0 = M_+$, and $\ell = \|H_0\|_{\mathrm{op}}$. Then*

$$\mathrm{tr}(H_0) \le k\ell + m\delta_\theta \le k\ell + m\kappa_\theta \|g\|. \tag{4}$$

If $\ell > 0$ and $m\kappa_\theta \|g\| \le \tau\ell$, then $r_{\mathrm{eff}}(H_0) \le \min\{d, k+\tau\}$. The same conclusion holds under the weaker, directly testable condition $m\delta_\theta \le \tau\ell$.

Proof. The proof bounds the tangent contribution to $\mathrm{tr}(H_0)$ using invariance and the normal contribution using the orbit codimension. See Appendix A.3 for details. □

Remark 2. *At a stationary point ($g = 0$), the tangent contribution vanishes and $r_{\mathrm{eff}}(H_0) \le k$, which recovers our result from Proposition 1. Away from stationarity, it introduces the correction term $m\kappa_\theta\|g\|$, which quantifies how much the effective rank differs from that of stationary points.*

We next show that the dependence on both orbit codimension and nonstationarity in Theorem 1 is tight; that is, there exists a function such that the equality holds. The following family attains the bound exactly for arbitrary positive orbit dimension and codimension.

Proposition 2. *Let $m, k \ge 1$, $d = m + k$, and $(x, z) \in \mathbb{R}^{m+1} \times \mathbb{R}^{k-1}$. For $0 \le \varepsilon \le 1$, define*

$$f_\varepsilon(x, z) = \frac{\varepsilon}{2}\|x\|^2 + \frac{1-\varepsilon}{8}(\|x\|^2 - 1)^2 + \frac{1}{2}\|z\|^2. \tag{5}$$

This objective is smooth and invariant under $\mathrm{SO}(m+1)$ acting on x and fixing z. At $\theta_0 = (e_1, 0)$, the orbit has dimension m and codimension k, and

$$\nabla^2 f_\varepsilon(\theta_0) = \mathrm{diag}(1, \varepsilon I_m, I_{k-1}), \qquad \kappa_{\theta_0} = \ell = 1, \qquad \|\nabla f_\varepsilon(\theta_0)\| = \varepsilon.$$

Thus both inequalities in Equation (4) hold with equality and $r_{\mathrm{eff}}(H_0) = k + m\varepsilon$. For every $\tau \ge 0$, choosing $\varepsilon = \min\{\tau/m, 1\}$ gives $m\kappa_{\theta_0}\|\nabla f_\varepsilon(\theta_0)\| \le \tau\ell$ and

$$r_{\mathrm{eff}}(H_0) = \min\{d, k + \tau\}. \tag{6}$$

The proof is in Appendix A.3. At $\varepsilon = 0$, the stationary orbit attains $r_{\mathrm{eff}}(H_0) = k$; at $\varepsilon = 1$, the objective is a scalar quadratic with $r_{\mathrm{eff}}(H_0) = d$. These endpoints and every intermediate value are attained at finite dimension. The case $m = 0$, $k = d$, is attained by a scalar positive definite quadratic with the trivial group action.

4.2 FINITE-SAMPLE CONVERGENCE OF ZO-SGD

We now connect Theorem 1 to the classical ZO-SGD update. The boundedness assumption below is supposed to hold on a small convex neighborhood near the stationarity.

Assumption 2. *Let $U \subseteq \mathbb{R}^d$ be nonempty, open, and convex, and suppose Assumption 1 holds at every $y \in U$. Let $M_y = \nabla^2 f(y)$, $\ell_y = \|(M_y)_+\|_{\mathrm{op}}$, $m_y = \dim T_y$, $k_y = d - m_y$, and $\delta_y = \|M_y P_{T_y}\|_{\mathrm{op}}$. There exist constants $L > 0$, $1 \le K \le d$, and $\tau \ge 0$ such that,*

$$\ell_y \le L, \qquad k_y \le K, \qquad \text{and} \qquad m_y \delta_y \le \tau L \tag{7}$$

for every $y \in U$.

Applying Theorem 1 pointwise under Assumption 2 gives

$$\nabla^2 f(y) \preceq L I_d, \qquad \mathrm{tr}((\nabla^2 f(y))_+) \le (K + \tau)L, \qquad y \in U. \tag{8}$$

We note that the first inequality, $\nabla^2 f(y) \preceq L I_d$, corresponds to the local r-effective rank assumption in Malladi et al. (2023) when $H(\theta)$ is chosen as $L I_d$. However, unlike their assumption, we do not require this matrix to have effective rank at most r. Our analysis instead uses $\mathrm{tr}((\nabla^2 f(y))_+) \le L(K + \tau)$, derived from the orbit codimension and nonstationarity correction. This bound controls the geometric contribution to the query complexity without requiring a common upper-bounding matrix with small effective rank.

Assumption 3 (Polyak–Łojasiewicz)**.** *Let $U \subseteq \mathbb{R}^d$ be nonempty, open, and convex, and suppose Assumption 1 holds at every $y \in U$. There are a finite lower bound $f^\star$ for f on U and a constant $\nu > 0$ such that*

$$\tfrac{1}{2} \|\nabla f(y)\|^2 \ge \nu (f(y) - f^\star) \tag{9}$$

holds for every $y \in U$

Without loss of generality, we can always select a valid constant ν satisfying $0 < \nu \le L$; otherwise, we can replace a supplied PL constant by its minimum with L if needed. The PL condition does not require strong convexity or a unique minimizer (Karimi et al., 2016). It turns the gradient term into a contraction of the objective gap. Set

$$D = \frac{d}{d+2} L(r+2) \le L(K + \tau + 2). \tag{10}$$

Then we have the following one-step iteration upper bound.

Theorem 2. *Suppose Assumptions 2 and 3 hold. Fix $\theta \in U$ and $\mu > 0$ with $\overline{B}(\theta, \mu) \subset U$. Let $Z = \nabla \mathcal{L}(\theta; \xi)$ satisfy $\mathbb{E} Z = g = \nabla f(\theta)$ and $\mathbb{E} \|Z - g\|^2 \le \sigma^2$, independently of $v \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Almost surely in ξ, suppose $\mathcal{L}(\cdot; \xi)$ is C^2 with ρ_s-Lipschitz Hessian on a neighborhood of this closed query ball. For $0 < \eta \le 1/(4D)$, define the analysis constants*

$$e_\mu = \frac{d \rho_s \mu^2}{6}, \qquad \lambda = \min\left\{ \frac{\eta \nu}{8}, \frac{D}{12 L d} \right\}, \qquad A_{\eta,\lambda} = \frac{\eta L}{\lambda^4} + \frac{9}{\lambda^2}. \tag{11}$$

If $\eta e_\mu \le \lambda \mu$ and the update $\theta^+ = \theta - \eta \widehat{g}_\mu$ lies in U almost surely, then

$$\mathbb{E}[f(\theta^+) - f^\star] \le \left(1 - \frac{\eta \nu}{2}\right)(f(\theta) - f^\star) + \eta^2 D \sigma^2 + \eta A_{\eta,\lambda} e_\mu^2. \tag{12}$$

Proof. Full details are given in Appendix A.5. □

We are now ready to state our dimension-independent function-query complexity guarantee.

Corollary 1 (Dimension-independent function-query complexity)**.** *Suppose Assumptions 2 and 3 hold on the same U. Starting from deterministic $\theta_0 \in U$, use the classical ZO-SGD with fixed $\eta, \mu > 0$. Assume that, through a deterministic horizon T, the iterates lie in U, the closed query balls lie in U, and the conditions of Theorem 2 hold almost surely. Then*

$$\mathbb{E}[f(\theta_T) - f^\star] \le e^{-\eta \nu T/2} \Delta_0 + \frac{2\eta D \sigma^2 + 2 A_{\eta,\lambda} e_\mu^2}{\nu}, \tag{13}$$

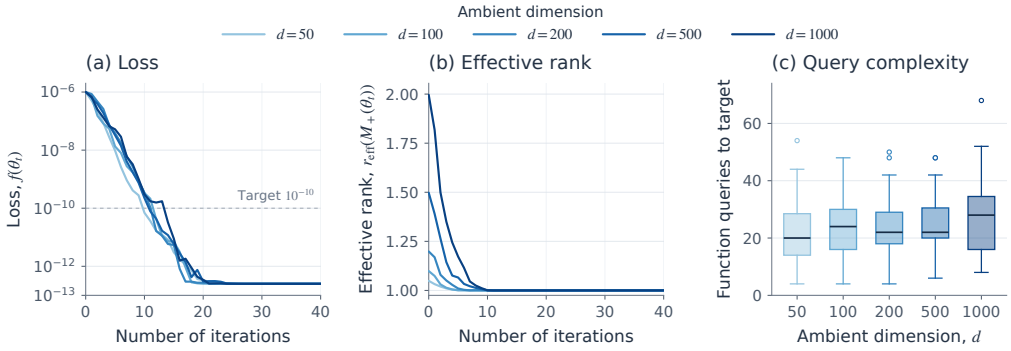


Figure 2: Spherical two-point ZO near a stationary symmetry orbit. (a) Loss and (b) positive-Hessian effective rank: medians over 40 seeds per dimension; darker blue denotes larger d. (c) Equivalent function queries to first reach $f \le 10^{-10}$. Protocol and boxplot definitions are in Appendix B.1.

where $\Delta_0 = f(\theta_0) - f^\star$. *For* $0 < \epsilon \le \Delta_0$, *choose*

$$\eta = \frac{1}{8D}\frac{\nu\epsilon}{\sigma^2 + \nu\epsilon}, \tag{14}$$

$$\eta e_\mu \le \lambda\mu, \qquad A_{\eta,\lambda} e_\mu^2 \le \frac{\nu\epsilon}{8}, \qquad \mu > 0. \tag{15}$$

Moreover, the ZO-SGD algorithm achieves $\mathbb{E}[f(\theta_T) - f^\star] \le \epsilon$ *after*

$$Q = 2\left\lceil \frac{16L(K+\tau+2)}{\nu}\left(1 + \frac{\sigma^2}{\nu\epsilon}\right)\log\frac{2\Delta_0}{\epsilon}\right\rceil, \qquad T = Q/2, \tag{16}$$

function evaluations.

Remark 3 (Discussion on dimension independence)**.** *For any fixed target accuracy* ϵ*, the query bound in Equation (16) has no explicit dependence on the ambient dimension* d*, provided that* $K, \tau, L/\nu, \sigma^2/\nu$*, and* Δ_0 *are bounded by dimension-independent constants. The finite-difference radius may shrink with* d*, without increasing function-query complexity.*

For $\rho_s > 0$, the smoothing inequalities are satisfied by

$$0 < \mu \le \min\left\{\frac{6\lambda}{\eta d \rho_s}, \left(\frac{9\nu\epsilon}{2A_{\eta,\lambda} d^2 \rho_s^2}\right)^{1/4}\right\}; \tag{17}$$

for $\rho_s = 0$, they impose no additional radius restriction. The query balls must still lie in U. The proof of Corollary 1, including integrability and both error floors, is in Appendix A.6.

5 Experiments

Motivated by Theorem 1, we empirically investigate the connection between Hessian effective rank and ZO optimization. In particular, we examine the following progression:

$$\text{ZO training} \implies \text{approach to stationarity} \implies \text{lower effective rank.} \tag{P}$$

The synthetic experiment provides a controlled illustration of the theoretical mechanism, while the neural experiment examines whether similar low-effective-rank behavior occurs in practice.

5.1 ZO Near a Stationary Symmetry Orbit

We first consider the radial quartic $f(\theta) = \frac{(\|\theta\|^2-1)^2}{4}$, which is rotationally invariant and whose unit sphere forms a stationary orbit of codimension one. We simulate the exact radial dynamics induced by spherical two-point ZO in dimensions ranging from 50 to 1000, using the same initialization gap and learning rate across dimensions. Full experimental details are provided in Appendix B.1.

As shown in Figure 2(a,b), the loss decreases on comparable iteration scales across dimensions, while the effective rank converges toward one. At initialization, the effective rank ranges from 1.05 to 2.00, remaining far below the ambient dimension d. This indicates that near a stationary symmetry orbit, the effective rank can be intrinsically low even in high-dimensional spaces. Figure 2(c) shows similar query counts across dimensions: increasing d by a factor of twenty (from 50 to 1000) raises the median queries only from 20 to 28. This weak dimension dependence is consistent with the low effective rank near stationarity.

5.2 ZO Training of a Symmetric Neural Model

We next examine the connection between parameter symmetry, effective rank, and ZO optimization in a Transformer. Its attention layers admit continuous parameter symmetries of the form described in Example 3. We train a two-layer Transformer with $d = 28{,}930$ parameters on the SMS spam classification task using spherical two-point ZO with the cross-entropy objective. We conduct ten independent runs from different random initializations, without pretraining. Appendix B.2 provides details of the model architecture, optimization setup, and effective-rank measurement.

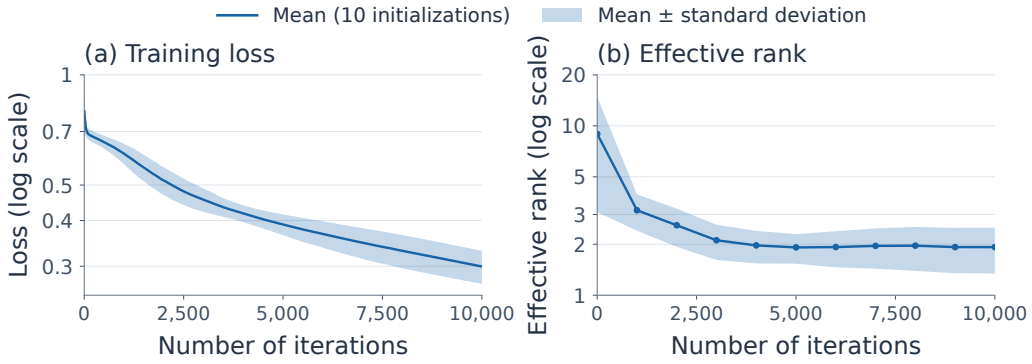


Figure 3: ZO training from ten random initializations of a 28,930-parameter Transformer. Curves show means and shaded bands one standard deviation across 10 independent runs.

As shown in Figure 3, we observe the same progression summarized in P: the training loss decreases as the effective rank becomes substantially smaller than d. In particular, the mean effective rank decreases from 8.97 at initialization to 1.94 at the end of training. This behavior indicates that, as the optimization trajectory approaches stationarity, low effective rank naturally emerges in the objective geometry.

6 Conclusion

In this work, we studied how continuous parameter symmetry can reduce the effective rank of Hessian and improve the efficiency of zeroth-order optimization. We established an effective-rank upper bound in terms of orbit codimension and nonstationarity (Theorem 1) and constructed a smooth invariant family that attains it exactly (Proposition 2), showing that our derived upper bound is tight. Under mild assumptions (Assumptions 2 and 3), with query balls and iterates remaining in that set almost surely, we derived finite-sample descent and convergence guarantees for ZO-SGD (Theorem 2 and Corollary 1), explicitly accounting for sampling noise and finite-difference error. These guarantees yield dimension-independent function-query complexity for without assuming a common low-effective-rank Hessian upper matrix. In experiments, empirical evidences are also provided to illustrate the theoretical mechanism and the connection between effective rank and ZO optimization. However, our results remain subject to several limitations: the theoretical guarantees rely on uniform local geometry, the PL condition, and the assumed containment of both queries and iterates. Future work will aim to establish verifiable conditions under which these assumptions hold and to empirically evaluate the predicted benefits in large-scale language model fine-tuning.

AI DISCLOSURE SECTION

In this work, we used generative AI tools for providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, and implementing methods. We have not used generative AI tools for formulating mathematical claims, proposing or refining hypotheses, helping develop theoretical models or conceptual frameworks, supporting qualitative and thematic data analysis, interpreting results, assisting with translation, or cleaning and reformatting dataset. Generating synthetic data sets are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature, formatting references, proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order finetuning for language models with low-rank structures. In *International Conference on Learning Representations*, 2025. arXiv:2410.07698.

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. doi: 10.1109/TIT.2015.2409256.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling FP8 training to trillion-token LLMs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f48b5133e89854a9e97cc22a6db83f25-Abstract-Conference.html.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2232–2241, 2019.

Elisenda Grigsby, Kathryn Lindsey, and David Rolnick. Hidden symmetries of relu networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11734–11760, 2023.

Kevin G. Jamieson, Robert D. Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, 2012.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851 of *Lecture Notes in Computer Science*, pp. 795–811. Springer, 2016. doi: 10.1007/978-3-319-46128-1_50.

Vladimir Koltchinskii and Karim Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157, 2017. doi: 10.1214/16-AOS1437. URL https://arxiv.org/abs/1504.07333.

John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2 edition, 2013.

Tianyi Li and Zhiqiang Shen. Scaling linear mode connectivity and merging to billion parameter pretrained transformers. *arXiv preprint arXiv:2606.23607*, 2026. URL https://arxiv.org/abs/2606.23607.

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*, 2025a. arXiv:2402.15751.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e5b1c0d4866f72393c522c8a00eed4eb-Abstract-Conference.html.

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

Bamdev Mishra, Gilles Meyer, Silvere Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29(3–4):591–621, 2014. doi: 10.1007/s00180-013-0464-z.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020. URL https://jmlr.org/papers/v21/20-933.html.

Yuezhang Peng, Yuxin Liu, Fei Wen, and Xie Chen. MUZO: Leveraging multiple queries and momentum for zeroth-order fine-tuning of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8566–8584. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.432.

Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 3–24. PMLR, 2013.

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

Sifeng Shang, Jiayi Zhou, Chenyu Lin, Minxian Li, and Kaiyang Zhou. Fine-tuning quantized neural networks with zeroth-order optimization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d17ef435b405a25715898c26dab3ed67-Abstract-Conference.html.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. URL https://arxiv.org/abs/2002.05202.

Berfin Simsek, Francois Ged, Arthur Jacot, Francesco Spadaro, Clement Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9722–9732, 2021.

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. In *Advances in Neural Information Processing Systems*, 2021.

Viet-Hoang Tran, Thieu N. Vo, An Nguyen, Tho-Huu Tran, Minh-Khoi Nguyen-Nhat, Thanh Tran, Duy-Tung Pham, and Tan Nguyen. Equivariant neural functional networks for transformers. In *International Conference on Learning Representations*, 2025.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015. doi: 10.1561/2200000048. URL https://arxiv.org/abs/1501.01571.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Hong Wang and Kelly Wang. Complete characterization of gauge symmetries in transformer architectures. In *Proceedings of the 4th (2025) and 3rd (2024) NeurIPS Workshops on Symmetry and Geometry in Neural Representations*, volume 282 of *Proceedings of Machine Learning Research*, pp. 622–663. PMLR, 2026. URL https://proceedings.mlr.press/v282/wang26a.html.

Qian Yu, Yining Wang, Baihe Huang, Qi Lei, and Jason D. Lee. Stochastic zeroth-order optimization under strongly convexity and lipschitz hessian: Minimax sample complexity. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of llms in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4475–4485, 2025.

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

Binchi Zhang, Zaiyi Zheng, Zhengzhang Chen, and Jundong Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77090–77106. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25dk.html.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59173–59190. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ad.html.

Jiajun Zhou, Yifan Yang, Kai Zhen, Ziyue Liu, Yequan Zhao, Ershad Banijamali, Athanasios Mouchtaris, Ngai Wong, and Zheng Zhang. QuZO: Quantized zeroth-order fine-tuning for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5341–5359. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.271.

APPENDIX CONTENTS

A Theoretical Results **13**

A.1 Supporting Lemmas 13

A.2 Proof of Proposition 1 14

A.3 Proof of Theorem 1 14

A.4 Proof of Proposition 2 15

A.5 Proof of Theorem 2 15

A.6 Proof of Corollary 1 17

B Experimental Details **17**

B.1 Radial Quartic Objectives 17

B.2 Two-Layer Transformer 18

A THEORETICAL RESULTS

A.1 SUPPORTING LEMMAS

Infinitesimal identity. For fixed $X \in \mathfrak{g}$, differentiate $f(\exp(tX) \cdot y) = f(y)$ at $t = 0$. This gives $\langle \nabla f(y), \xi_X(y) \rangle = 0$ on an open neighborhood. Differentiating that identity in an arbitrary direction w gives

$$w^\top \nabla^2 f(y) \xi_X(y) + \langle \nabla f(y), D\xi_X(y) w \rangle = 0,$$

which proves Equation (2).

There is an equivalent scalar orbit-curvature identity. For an orbit curve γ with $\gamma(0) = \theta$ and $\gamma'(0) = v$,

$$v^\top M v = -\langle g, II_\theta(v, v) \rangle . \tag{18}$$

Here $II_\theta(v, v)$ is the normal component of $\gamma''(0)$, on a locally embedded orbit. The tangent component is orthogonal to g. Equation (18) follows by differentiating $f(\gamma(t))$ twice; it explains how nonzero tangent curvature can coexist with invariance.

Lemma 1. *Let Z be independent of $v \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, with $\mathbb{E}\|Z\|^2 < \infty$, and let $H \succeq 0$ be fixed. For the first-order term $X = d(v^\top Z)v$,*

$$\mathbb{E}[X^\top H X] = c_d \{ \mathrm{tr}(H) \mathbb{E}\|Z\|^2 + 2\mathbb{E}[Z^\top H Z] \}, \tag{19}$$

where $c_d = d/(d+2)$.

Proof. Condition on $Z = z$. Rotational invariance and $\|v\| = 1$ give

$$\mathbb{E}[v_a v_b v_i v_j] = \frac{\delta_{ab}\delta_{ij} + \delta_{ai}\delta_{bj} + \delta_{aj}\delta_{bi}}{d(d+2)}.$$

For example, sign symmetry eliminates odd moments, rotational invariance gives $\mathbb{E}v_1^4 = 3\mathbb{E}v_1^2 v_2^2$, and expanding $\mathbb{E}(\sum_i v_i^2)^2 = 1$ determines the denominator. Expanding $d^2(v^\top z)^2(v^\top H v)$ therefore gives $\frac{d}{d+2}\{\mathrm{tr}(H)\|z\|^2 + 2z^\top H z\}$. Finally average over Z; the formula is immediate for $d = 1$.

□

Lemma 2. *Fix $x, z \in \mathbb{R}^d$ and $h > 0$. Suppose f is C^2 on a neighborhood of the closed ball with center $c = x - hdz/2$ and radius $R = hd\|z\|/2$. On this ball, suppose $\nabla^2 f \preceq LI_d$ and $\mathrm{tr}(\nabla^2 f) \le B$, where $L > 0$ and $B \ge 0$. Then*

$$\mathbb{E}_v f(x - hd(v^\top z)v) \le f(x) - h\nabla f(x)^\top z + \frac{h^2}{2}\frac{d}{d+2}(B + 2L)\|z\|^2, \quad v \sim \mathrm{Unif}(\mathbb{S}^{d-1}). \tag{20}$$

For $d = 1$, the same conclusion holds with coefficient L in place of $d(B+2L)/(d+2)$; the displayed coefficient is valid whenever $B \ge L$.

Proof. The case $z = 0$ is immediate, and $d = 1$ follows from scalar Taylor integration on the segment. Suppose $d \geq 2$ and $z \neq 0$, and put $e = z/\|z\|$. The endpoint is

$$x - hd(v^\top z)v = c + Rw, \qquad w = e - 2(v^\top e)v \in \mathbb{S}^{d-1}.$$

Let u be the harmonic extension to the ball of the boundary values of f. Since the Laplacian of f is at most B, the maximum principle gives

$$u(y) \leq f(y) + \frac{B}{2d}(R^2 - \|y - c\|^2). \tag{21}$$

Indeed, the right side has nonpositive Laplacian and equals u on the boundary.

For $d = 2$, the reflected point w is uniform on the circle, so its boundary average is $u(c)$. For $d \geq 3$, set $\alpha = (d-2)/2$. Relative to normalized surface measure, the density of w is $\|w - e\|^{2-d}$. To see this, $(v^\top e)^2$ has the beta law with parameters $1/2, (d-1)/2$, while $w^\top e = 1 - 2(v^\top e)^2$; comparing its density with that of the first coordinate of a uniform point on the sphere gives the claimed factor. The Poisson kernel at te is $P_d(t, w) = (1 - t^2)/\|w - te\|^d$. Direct differentiation gives

$$\frac{d}{dt}\left(t^\alpha \|w - te\|^{2-d}\right) = \alpha t^{\alpha-1} P_d(t, w).$$

Integration from zero to one therefore represents the reflected density as a mixture of Poisson kernels. Consequently,

$$\mathbb{E}_v f(c + Rw) = \mathbb{E}_S u(c + RSe), \tag{22}$$

where S has density $\alpha t^{\alpha-1}$ on $(0, 1)$ for $d \geq 3$, and $S = 0$ for $d = 2$. The integrable boundary singularity causes no difficulty because the boundary values of f are continuous and bounded. In both cases,

$$\mathbb{E}(1 - S) = \frac{2}{d}, \qquad \mathbb{E}(1 - S)^2 = \frac{8}{d(d+2)}, \qquad \mathbb{E}(1 - S^2) = \frac{4}{d+2}.$$

Apply Equation (21) at $c + RSe = x - R(1 - S)e$, and use $\nabla^2 f \preceq LI_d$ along the axis segment:

$$f(x - R(1 - S)e) \leq f(x) - R(1 - S)\nabla f(x)^\top e + \tfrac{L}{2}R^2(1 - S)^2.$$

Substituting the three moments and $R = hd\|z\|/2$ proves Equation (20). No independence between the endpoint Hessian and v is used. □

A.2 Proof of Proposition 1

Proof. By Equation (2), stationarity implies $\nabla^2 f(\theta_\star)\xi_X(\theta_\star) = 0$ for every $X \in \mathfrak{g}$. These vectors span the orbit tangent space, so it is contained in the Hessian kernel. Rank-nullity gives the rank bound. Taking the positive part cannot increase matrix rank, and a PSD matrix has effective rank no larger than its rank. This also proves the effective-rank statement for indefinite Hessians. □

A.3 Proof of Theorem 1

Proof of Theorem 1. For unit $q \in T_\theta$, Equation (2) and the infimum in Equation (3) give $\|Mq\| \leq \kappa_\theta \|g\|$. Taking the supremum over unit tangent vectors yields $\delta_\theta \leq \kappa_\theta \|g\|$; when $m = 0$, $\delta_\theta = 0$. If $H_0 = 0$, the trace bound follows immediately. Otherwise, $\ell > 0$. Moreover, $H_0^2 \preceq M^2$ by spectral calculus, so

$$q^\top H_0 q \leq \|H_0 q\| \leq \|Mq\| \leq \delta_\theta.$$

Let $q_1, \ldots, q_m$ be an orthonormal tangent basis and $n_1, \ldots, n_k$ an orthonormal normal basis. Each normal direction satisfies $n_j^\top H_0 n_j \leq \ell$. Consequently,

$$\operatorname{tr}(H_0) = \sum_{i=1}^{m} q_i^\top H_0 q_i + \sum_{j=1}^{k} n_j^\top H_0 n_j \leq m\delta_\theta + k\ell \leq m\kappa_\theta \|g\| + k\ell,$$

which proves Equation (4). When $m = 0$, the tangent sum is empty. Dividing by ℓ gives

$$r_{\text{eff}}(H_0) \leq k + \frac{m\delta_\theta}{\ell} \leq k + \frac{m\kappa_\theta \|g\|}{\ell}.$$

Also, $r_{\text{eff}}(H_0) \leq d$, since every eigenvalue of H_0 is at most ℓ. Either stated condition therefore yields $r_{\text{eff}}(H_0) \leq \min\{d, k + \tau\}$. □

The tangent estimate also gives $\operatorname{tr}(P_T H_0 P_T) \le m \|M P_T\|_{\mathrm{op}}$. No positive-semidefiniteness of M or constant orbit dimension in a neighborhood is required.

For the radial quartic in Equation (30), $M = 2\theta\theta^\top + (r^2-1)I_d$. At $r > 1$, $M \succeq 0$, $\ell = 3r^2 - 1$, and the tangent eigenvalue equals $\delta_\theta = r^2 - 1$. For a unit tangent q, the skew generator $A = (q\theta^\top - \theta q^\top)/r^2$ satisfies $A\theta = q$ and $\|A\|_{\mathrm{op}} = 1/r$; conversely $\|A\|_{\mathrm{op}} \ge 1/r$ follows from $A\theta = q$. Thus $\kappa_\theta = 1/r$ and $\kappa_\theta \|g\| = r^2 - 1$. Both inequalities in Equation (4) are attained. The constant multiplying $m\kappa_\theta \|g\|$ cannot be decreased uniformly.

A.4 Proof of Proposition 2

Proof. For the objective in Equation (5), rotations of x preserve $\|x\|$ and fix z, so the objective is invariant under the stated action. The orbit of $\theta_0 = (e_1, 0)$ is $S^m \times \{0\}$, which has dimension m and codimension k. Direct differentiation gives

$$g = (\varepsilon e_1, 0), \qquad M = \operatorname{diag}(1, \varepsilon I_m, I_{k-1}) \succeq 0.$$

Consequently $H_0 = M$, $\ell = 1$, $\delta_{\theta_0} = \varepsilon$, and $\|g\| = \varepsilon$. For a unit $q \in \mathbb{R}^{m+1}$ with $q^\top e_1 = 0$, the generator $A = \operatorname{diag}(qe_1^\top - e_1 q^\top, 0)$ satisfies $A\theta_0 = (q, 0)$ and $\|A\|_{\mathrm{op}} = 1$. Every generator producing a unit tangent vector has operator norm at least one, since $\|\theta_0\| = 1$. Thus $\kappa_{\theta_0} = 1$, and both inequalities in Equation (4) hold with equality:

$$r_{\mathrm{eff}}(H_0) = k + m\varepsilon = k + \frac{m\delta_{\theta_0}}{\ell} = k + \frac{m\kappa_{\theta_0}\|g\|}{\ell}.$$

For any prescribed $\tau \ge 0$, choose $\varepsilon = \min\{\tau/m, 1\}$. Then $m\kappa_{\theta_0}\|g\| \le \tau\ell$ and $r_{\mathrm{eff}}(H_0) = \min\{d, k+\tau\}$. Equality therefore holds at finite dimension for every $m, k \ge 1$ and $\tau \ge 0$. In particular, $\varepsilon = 0$ attains the stationary bound $r_{\mathrm{eff}}(H_0) = k$, while $\varepsilon = 1$ attains the ambient bound $r_{\mathrm{eff}}(H_0) = d$. The case $m = 0$, $k = d$, is attained by a positive definite scalar quadratic with the trivial group action. □

A.5 Proof of Theorem 2

Proof. Write $x = \theta$, $g = \nabla f(x)$, $G = \|g\|$, and $\Delta = f(x) - f^\star$. By Theorem 1 and assumption 2, $\operatorname{tr}(\nabla^2 f) \le \operatorname{tr}((\nabla^2 f)_+) \le rL$ and $\nabla^2 f \preceq LI_d$ on U. Convexity of U therefore gives the scalar upper model

$$f(y) \le f(x) + g^\top(y - x) + \tfrac{L}{2}\|y - x\|^2, \qquad x, y \in U. \tag{23}$$

The same upper curvature bound implies the interpolation inequality

$$f((1-t)x + ty) \ge (1-t)f(x) + tf(y) - \tfrac{L}{2}t(1-t)\|y - x\|^2, \quad 0 \le t \le 1. \tag{24}$$

This follows from concavity of $f(y) - L\|y\|^2/2$; it does not require a lower Hessian bound.

Oracle remainder and integrability. Taylor expansion on the query ball gives, almost surely,

$$\frac{\mathcal{L}(x + \mu v; \xi) - \mathcal{L}(x - \mu v; \xi)}{2\mu} = v^\top Z + R_\xi(v), \qquad |R_\xi(v)| \le \rho_s \mu^2/6.$$

Set $e = e_\mu$, $q = e/d$, and $X = d(v^\top Z)v$, so $\widehat{g}_\mu = X + E_\xi(v)$, where $E_\xi(v) = dR_\xi(v)v$ and $\|E_\xi(v)\| \le e$. Because $\mathbb{E}\|X\|^2 = d\mathbb{E}\|Z\|^2$,

$$\mathbb{E}\|\widehat{g}_\mu\|^2 \le 2d(G^2 + \sigma^2) + 2e^2 < \infty.$$

The upper model at the actual endpoint and the lower bound $f^\star$ imply integrability of its objective gap. Thus all subsequent conditional expectations are well defined.

A reference ball inside the same domain. The constants satisfy

$$D \ge L \ge \nu > 0, \qquad 0 < \lambda \le \frac{1}{32}, \qquad 3\lambda L d \le \frac{D}{4}, \qquad \lambda^2 \le \frac{\eta D}{8}.$$

Put $\beta = 1 - 2\lambda$, so $1/2 \le \beta < 1$. Fix a sample ξ from the probability-one set where the oracle is regular and update containment holds for almost every direction, and write $z = Z(\xi)$, $a = v^\top z$, $b = v^\top g$. Introduce only for the proof

$$y = x - \eta d(a + R_\xi(v))v, \qquad y_\beta = x - \beta\eta d a v.$$

On directions with $|a| \ge q/\lambda$, excluding a null set if $e = 0$,

$$y_\beta = (1-t)x + ty, \qquad t = \frac{\beta a}{a + R_\xi(v)}, \qquad 0 < t \le \frac{\beta}{1-\lambda} < 1.$$

For $|a| < q/\lambda$, $\|y_\beta - x\| \le \beta\eta e/\lambda \le \beta\mu < \mu$. Hence the reference endpoints lie in U: the first case uses convexity and the actual update, and the second uses the query ball. This also holds for all limiting directions. More explicitly, continuity of the actual endpoint in v and full support of spherical sampling put every actual endpoint in $\overline{U}$; the strict bound on t, together with $x \in U$, puts the first reference endpoint in U. The second is in the query ball directly. The reference sphere is compact, and its closed ball, its convex hull, is contained in U. Thus Lemma 2 applies to y_β, with $h = \beta\eta$ and $B = rL$. In particular,

$$\frac{1}{\beta}\mathbb{E}_v[f(y_\beta) - f(x)] \le -\eta g^\top z + \frac{\beta\eta^2 D}{2}\|z\|^2. \tag{25}$$

No separate assumption of containment for an ideal directional-gradient update has been introduced.

Comparison on directions with a large projection. Let $F_v = f(y_\beta) - f(x)$. On $|a| \ge q/\lambda$, rearranging Equation (24) gives

$$f(y) - f(x) \le \frac{F_v}{\beta} + \frac{R_\xi(v)}{\beta a}F_v + \frac{L\eta^2 d^2}{2}(a + R_\xi(v))(2\lambda a + R_\xi(v)).$$

The lower bound $f^\star$ and Equation (23) give

$$|F_v| \le \Delta + \beta\eta d|ab| + \tfrac{L}{2}\beta^2\eta^2 d^2 a^2.$$

Since $|R_\xi(v)/a| \le \lambda$ and $\lambda \le 1/4$, it follows that

$$f(y) - f(x) \le \frac{F_v}{\beta} + 2\lambda\Delta + \lambda\eta d|ab| + 3\lambda L\eta^2 d^2 a^2. \tag{26}$$

For the last coefficient, the extrapolation term is at most $\frac{3}{2}\lambda(1+\lambda)L\eta^2 d^2 a^2$, and the preceding remainder contributes at most $\lambda L\eta^2 d^2 a^2/2$; their sum is at most $3\lambda L\eta^2 d^2 a^2$.

Small projections and small sample gradients. On $|a| < q/\lambda$, use $F_v \ge -\Delta$ and the scalar upper model directly. The bound $\|y - x\| \le (1+\lambda)\eta e/\lambda$ yields

$$f(y) - f(x) \le \frac{F_v}{\beta} + 2\Delta + \frac{2\eta G e}{\lambda} + \frac{L\eta^2 e^2}{\lambda^2}. \tag{27}$$

For a uniform spherical direction and $z \ne 0$,

$$\mathbb{P}_v(|v^\top z| < a_0) \le \min\{1, \sqrt{d}\, a_0/\|z\|\}, \qquad a_0 > 0. \tag{28}$$

For $d \ge 3$, this follows from the coordinate density $\Gamma(d/2)(1-t^2)^{(d-3)/2}/[\sqrt{\pi}\Gamma((d-1)/2)]$, whose maximum is at most $\sqrt{d}/2$. The case $d = 2$ follows from $(2/\pi)\arcsin t \le t$ on $[0,1]$, and $d = 1$ is immediate.

When $e > 0$, split samples at the deterministic threshold $z_0 = e/(\lambda^2\sqrt{d})$. For $\|z\| > z_0$, the probability of a small projection is at most λ, by Equation (28). Average Equations (26) and (27) over directions, and apply Equation (25) and $\mathbb{E}_v a^2 = \|z\|^2/d$, $\mathbb{E}_v|ab| \le \|z\| G/d$. This gives

$$\begin{aligned}\mathbb{E}_v[f(y) - f(x)] &\le -\eta g^\top z + \eta^2(D/2 + 3\lambda L d)\|z\|^2 + \lambda\eta G\|z\| \\ &\quad + 4\lambda\Delta + 2\eta G e/\lambda + L\eta^2 e^2/\lambda^2 \\ &\le -\eta g^\top z + \eta^2 D\|z\|^2 + \tfrac{\eta}{4}G^2 + 4\lambda\Delta + 9\eta e^2/\lambda^2.\end{aligned}$$

The second line uses

$$\lambda G \|z\| \le G^2/8 + 2\lambda^2 \|z\|^2, \qquad 2Ge/\lambda \le G^2/8 + 8e^2/\lambda^2,$$

with $3\lambda Ld \le D/4$, $2\lambda^2 \le \eta D/4$, and $\eta L \le 1$. For $\|z\| \le z_0$, direct spherical averaging of the scalar upper model gives instead

$$\begin{aligned}\mathbb{E}_v[f(y) - f(x)] &\le -\eta g^\top z + \eta G e + L\eta^2(d\|z\|^2 + e^2) \\ &\le -\eta g^\top z + \eta^2 L d z_0^2 + \tfrac{\eta}{8}G^2 + 3\eta e^2.\end{aligned}$$

Averaging these two sample classes, using $\mathbb{E}Z = g$ and $\mathbb{E}\|Z\|^2 \le G^2 + \sigma^2$, and substituting $dz_0^2 = e^2/\lambda^4$, yields

$$\mathbb{E}[f(y) - f^\star] \le (1 + 4\lambda)\Delta - \left(\frac{3\eta}{4} - \eta^2 D\right) G^2 + \eta^2 D\sigma^2 + \eta\left(\frac{\eta L}{\lambda^4} + \frac{9}{\lambda^2}\right) e^2. \tag{29}$$

When $e = 0$, use $z_0 = 0$, treat $z = 0$ directly (the update is zero), and apply the large-projection argument elsewhere; the same inequality holds. These arguments require only the second moment of Z.

Finally, $\eta D \le 1/4$ gives $3\eta/4 - \eta^2 D \ge \eta/2$. PL and $4\lambda \le \eta\nu/2$ turn Equation (29) into Equation (12). This completes the proof. □

A.6 Proof of Corollary 1

Proof. Let $\mathcal{F}_t$ be the information before the fresh sample and direction, and put $\Delta_t = f(\theta_t) - f^\star \ge 0$. Applying Theorem 2 conditionally gives

$$\mathbb{E}[\Delta_{t+1} \mid \mathcal{F}_t] \le (1 - \eta\nu/2)\Delta_t + \eta^2 D\sigma^2 + \eta A_{\eta,\lambda} e_\mu^2.$$

The conditional integrability argument in that proof is valid for almost every history. Starting from deterministic θ_0, nonnegative total expectation and induction give $\mathbb{E}\Delta_t < \infty$ for every finite time. Summing the geometric series, whose sum is at most $2/(\eta\nu)$, proves Equation (13). We keep the original conditional sampling law and do not condition on a future non-exit event.

For $0 < \epsilon \le \Delta_0$, Equation (14) satisfies $\eta D \le 1/8$. The two error floors obey

$$\frac{2\eta D\sigma^2}{\nu} = \frac{\epsilon}{4}\frac{\sigma^2}{\sigma^2 + \nu\epsilon} \le \frac{\epsilon}{4}, \qquad \frac{2A_{\eta,\lambda}e_\mu^2}{\nu} \le \frac{\epsilon}{4}.$$

It suffices to take

$$T \ge \frac{2}{\eta\nu}\log\frac{2\Delta_0}{\epsilon} = \frac{16D}{\nu}\left(1 + \frac{\sigma^2}{\nu\epsilon}\right)\log\frac{2\Delta_0}{\epsilon}.$$

The inequality $D \le L(K + \tau + 2)$, integer rounding, and two function evaluations per step prove Equation (16). Solving $\eta e_\mu \le \lambda\mu$ and $A_{\eta,\lambda}e_\mu^2 \le \nu\epsilon/8$ gives Equation (17) for $\rho_s > 0$. If $\rho_s = 0$, the central difference is exact at every admissible radius and both inequalities are automatic. All parameters are fixed during a run; the auxiliary λ is not an algorithmic change.

Only the smoothing radius and domain admissibility, not the displayed function-query count, retain ambient-dimension dependence when the listed geometric and oracle constants are fixed. Uniform geometry, PL, and containment must hold on the stated neighborhood for the chosen process. The proof establishes neither a dimension-uniform initialization region nor automatic containment. The geometric condition is not inferred from a low-rank Hessian at a single point. □

B Experimental Details

B.1 Radial Quartic Objectives

For $f(\theta) = (\|\theta\|^2 - 1)^2/4$, write $s_t = \|\theta_t\|^2$. With v_t uniform on the unit sphere, the central difference is exactly

$$\frac{f(\theta_t + \mu v_t) - f(\theta_t - \mu v_t)}{2\mu} = (s_t + \mu^2 - 1)\langle\theta_t, v_t\rangle.$$

The ordinary ZO update multiplies this difference by dv_t. Hence its squared radius follows

$$s_{t+1} = s_t[1 + Z_t(a_t^2 - 2a_t)], \qquad a_t = \eta d(s_t + \mu^2 - 1), \qquad Z_t \sim \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{d-1}{2}\right),$$

with independent draws at each update. Simulating this scalar recurrence gives the same loss dynamics in distribution as ambient spherical two-point ZO. It retains the finite-difference correction μ^2 and does not project onto a symmetry orbit. Each update corresponds to two equivalent function evaluations. Rotations generate $(d-1)$-dimensional orbits away from the origin. For $r = \|\theta\| > 1$, the Hessian is $M = 2\theta\theta^\top + (r^2 - 1)I_d$, with radial eigenvalue $3r^2 - 1$ and $d - 1$ tangent eigenvalues $r^2 - 1$. Thus $\delta_\theta = r^2 - 1$, and

$$r_{\mathrm{eff}}(M) = 1 + \frac{(d-1)(r^2-1)}{3r^2 - 1}, \qquad r > 1, \tag{30}$$

which attains the symmetry bound exactly. On the stationary unit sphere, $M = 2\theta\theta^\top$ and the effective rank is one. The factor $d - 1$ accounts for the accumulation of weak positive tangent curvature. At fixed $r > 1$, effective rank grows linearly with d. If $0 < r^2 - 1 \le C/d$, with C independent of d, then $r_{\mathrm{eff}}(M_+) \le 1 + C/2$; hence a bound uniform in dimension requires a shrinking neighborhood of the sphere.

At initialization, we select θ_0 such that $f(\theta_0) = 10^{-6}$. Figure 2 uses a retrospective common-learning-rate slice, $\eta = 0.2$, of the archived grid $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.8\}$, with $\mu = 10^{-3}$ and 40 seeds per dimension $d \in \{50, 100, 200, 500, 1000\}$. The archived runs stopped at the first loss at most 10^{-10}, with a budget of 2,000 updates. For the trajectory panels, we replay the same random streams and continue every run through update 40, including updates after the first target crossing. All 200 first hitting times agree with the archived results; the largest is 34 updates. Panels (a,b) plot pointwise medians over all 40 seeds, without stopping or discarding trajectories after a target crossing. At every iterate, the Hessian has one radial eigenvalue $3s_t - 1$ and $d - 1$ tangent eigenvalues $s_t - 1$. We clip these eigenvalues at zero before computing $r_{\mathrm{eff}}(M_+(\theta_t))$; thus the rank diagnostic also applies to iterates inside the unit sphere. Panel (c) uses twice each run's first hitting time, excluding tuning and monitoring. Quartiles use linear interpolation; boxplot whiskers end at the most extreme observations within 1.5 interquartile ranges of the quartiles, with all other observations shown as circles. These counts describe equivalent update queries; the recurrence does not measure the runtime of ambient function evaluations. All 200 sampled runs reach the target. For $d = 50, 100, 200, 500, 1000$, the median query counts are $20, 24, 22, 22, 28$, respectively. The initial effective rank ranges from 1.05 to 2.00.

B.2 Two-Layer Transformer

The model has two Transformer encoder layers, embedding width 32, one attention head per layer, feedforward width 16, GELU activations, and zero dropout. It uses a 512-token vocabulary, learned positional embeddings, 48 input tokens plus a prepended classification token, a final layer normalization, and a two-logit classifier. All $d = 28{,}930$ parameters are optimized. The fixed training subset contains 1,006 SMS examples, balanced between spam and non-spam. The objective is the mean cross-entropy over this entire subset, with no added regularizer. We use ten different model initialization seeds, 7–16, with no Adam pretraining. Initial parameters are generated in FP32 on the CPU using the original PyTorch initialization, which copies one encoder layer's initial weights into distinct parameter tensors for the two layers. All ten complete parameter vectors were verified to be different and to reproduce their specified initialization seeds.

We use the spherical two-point estimator in Equation (1), with one independent unit-sphere direction per update, $\mu = 0.01$, and fixed learning rate $\eta = 0.001$. This numeric learning rate is carried over from the original model-seed-7 experiment and is shared across all initializations; it is not recalibrated for each initial Hessian. The matrix smoothness, PL, and almost-sure trajectory-containment assumptions on a common convex open set are not certified for this experiment. Model seeds 7–16 are paired in order with ZO direction seeds 7101–7110, giving one trajectory per initialization on the same dataset. All ten runs complete 10,000 updates, with no target-based stopping. Each run uses 20,000 update function evaluations, excluding monitoring and offline curvature diagnostics. Computation uses FP32 model parameters and updates with FP64 cross-entropy accumulation.

For each run, at initialization and every 1,000 updates, we estimate $r_{\mathrm{eff}}(M_+)$, where M is the full parameter Hessian of the original cross-entropy at the saved iterate. Four Rademacher probes

with common probe seed 39871 and fully reorthogonalized stochastic Lanczos quadrature estimate $\mathrm{tr}(M_+)$; the denominator is the largest observed positive Ritz value. The Lanczos depth starts at 120 and is increased over $160, 200, 240$ if the positive-trace estimate changes by more than five percent relative to the preceding check. This is an empirical depth-sensitivity diagnostic. Representative terminal parameter points were checked with FP64 Hessian-vector products. Curvature probes are offline diagnostics and do not enter the ZO updates.

Loss is logged every 10 updates. At each recorded time, Figure 3 shows the arithmetic mean across all ten runs, with a shaded interval of one sample standard deviation (denominator $10 - 1$). Both vertical axes are logarithmic, with means and standard deviations computed on the original scale. No initialization is selected or discarded based on the result, and no temporal smoothing is applied. Rank markers identify the measured times; connecting lines interpolate only for display. Each initialization has its own initial loss and rank, so the initial band reflects variation across models. The shaded bands describe joint variation from model initialization and ZO directions on fixed data; they are neither confidence intervals nor measures of Lanczos estimation uncertainty. The initial effective rank has mean 8.97 and sample standard deviation 5.87.