GeoES: Geometry-Aware (1+1)-ES for Black-Box Optimization with Intrinsic Dimension Guarantees

Zixu Wu[1] **Hualin Zhang**[1] **Zhiqiang Xu**[1✉]

[1]Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates

{zixu.wu,hualin.zhang,zhiqiang.xu}@mbzuai.ac.ae

Abstract

The classical $(1+1)$-evolution strategy compares a single random trial with the current point, without estimating gradients. We study how this simple rule can use geometry from known constraints or a chosen parameterization. We introduce **GeoES**, a geometry-aware $(1+1)$-ES that samples Gaussian tangent directions, maps trials back to the manifold with a retraction, and retains improvements. For Gaussian proposals restricted to a prescribed radius, we derive non-asymptotic stationarity bounds for smooth non-convex objectives on compact Riemannian submanifolds. The analysis explains how selection turns symmetric proposals into expected progress and why the quadratic model error scales with the intrinsic tangent dimension d. With problem and truncation constants fixed, a suitable constant step gives an $\mathcal{O}(d\varepsilon^{-2})$ evaluation bound for an expected Riemannian gradient norm at most ε. Smooth sphere PCA shows decreasing weighted gradient norms, and Grassmann PCA maintains similar subspace errors with budgets proportional to d. Sparse PCA achieves comparable objective values to full covariance CMA-ES and faster progress within time budgets. In black-box prompt tuning, GeoES with separate direction and magnitude achieves a higher average score than Euclidean $(1+1)$-ES. Across prompt and adapter tuning, GeoES is competitive with full covariance CMA-ES and zeroth-order baselines under limited query budgets. Code is available at `https://github.com/MLOpt-MBZUAI/GeoES`.

1 Introduction

Black-box optimization uses function evaluations when derivatives are unavailable or costly. Many zeroth-order methods estimate a gradient from these evaluations and use it to construct an update [Nesterov and Spokoiny, 2017, Shamir, 2017]. MeZO follows this route for language model tuning [Malladi et al., 2023]. The classical $(1+1)$-ES uses a direct comparison: it keeps one parent, generates one random trial, and retains the better point. With the current objective value cached, a trial needs only one new evaluation. We ask how this simple rule can use the geometry of the search variables and how that geometry affects its convergence.

This question matters when the variables are high dimensional but have known constraints or symmetries. A sparse PCA loading vector has unit norm. In subspace estimation, orthonormal bases related by a rotation represent the same solution. These structures lead naturally to spheres and Grassmann manifolds. Riemannian optimization uses tangent spaces to describe their local degrees of freedom and retractions to map local steps to feasible points [Edelman et al., 1998, Absil et al., 2008, Boumal, 2023]. For random search, this suggests generating directions in the tangent space and mapping the resulting trials back to the feasible set.

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

Geometry can also be introduced through a search parameterization. Prompt and adapter tuning optimize a small set of variables mapped into a much larger language model [Aghajanyan et al., 2021, Sun et al., 2022b,a, Hu et al., 2022, Li et al., 2025]. Once this reduced search space is fixed, we still need to choose how to perturb its variables. An additive perturbation produces a smaller angular change when the current vector has a larger norm. Separating direction and magnitude allows the angular scale to be set independently of that norm. This choice is motivated by weight normalization and DoRA [Salimans and Kingma, 2016, Liu et al., 2024]. We use this representation to explore how the choice of search variables affects a simple comparison rule.

We introduce **GeoES**, a geometry-aware $(1+1)$-ES. For manifold constraints, it draws an isotropic tangent Gaussian, applies a retraction, and keeps the trial when its objective value is no larger than the parent's. Steps outside a prescribed radius are discarded without evaluation, so the evaluated trials lie where the pullback has a quadratic remainder bound. For language model tuning, we use a normalized tangent proposal to update direction and magnitude jointly.

To understand why GeoES makes progress, pair opposite tangent directions in the local linear model. Their changes cancel before selection. Comparing each trial with the parent preserves decreases and suppresses increases. A Gaussian direction has an expected absolute projection on a unit vector independent of dimension, while its expected squared length is $d = \dim \mathcal{M}$. These moments determine the linear progress and quadratic error in the descent bound. Truncation controls how much directional progress is retained. This explains how one objective comparison produces expected descent and why the tangent dimension enters the stationarity bounds.

We summarize our contributions as follows:

- We introduce GeoES, which combines tangent proposals and retractions with the classical $(1+1)$ selection rule. After the initial evaluation, the Gaussian algorithm uses at most one new objective evaluation per iteration.
- We derive non-asymptotic stationarity bounds for smooth non-convex objectives on compact Riemannian submanifolds. One descent inequality covers general bounded deterministic step-size schedules. A suitable constant step gives an $\mathcal{O}(d\varepsilon^{-2})$ evaluation bound for expected gradient norm at most ε, with problem and truncation constants fixed.
- Grassmann PCA maintains similar subspace errors with budgets proportional to d. Sparse PCA achieves comparable objective values to full covariance CMA-ES and faster progress within time budgets. In BBT, GeoES with separate direction and magnitude exceeds the seven-task average of Euclidean $(1+1)$-ES by 4.1 points. Prompt and adapter results are competitive with the evaluated baselines under limited budgets.

2 Related Work

Zeroth-order optimization and direct search. Randomized finite-difference methods estimate derivatives from function evaluations and admit dimension-dependent complexity bounds [Nesterov and Spokoiny, 2017, Shamir, 2017]. Direct search chooses among evaluated points. The stochastic three points method compares the current point with two symmetric trials and derives convergence from directional progress [Bergou et al., 2020]. GeoES uses one trial and the retained parent. Its analysis combines a symmetry argument with retraction smoothness and radial truncation.

Riemannian optimization. The geometry of orthogonality constraints distinguishes optimization over bases from optimization over subspaces [Edelman et al., 1998]. Tangent spaces and retractions extend local optimization to smooth manifolds [Absil et al., 2008, Boumal, 2023]. Pullback smoothness gives non-asymptotic guarantees for Riemannian gradient and trust-region methods [Boumal et al., 2019]. Riemannian zeroth-order methods construct derivative estimates or stochastic approximation updates from function values [Li et al., 2023, 2024, He et al., 2024]. In particular, Li et al. [2023] obtain complexity bounds in the intrinsic dimension. We use this local modeling framework to analyze elitist selection directly.

Evolution strategies. Analyses of $(1+1)$-ES include global convergence and rates for success-based step-size adaptation on structured objectives [Glasmachers, 2020, Morinaga et al., 2021]. Evolution strategies have also been developed on spheres, nonlinear constraint sets, and quadratic

manifolds [Colutto et al., 2010, Arnold, 2014, Spettel and Beyer, 2022]. Guided ES shapes its search distribution using surrogate gradient directions [Maheswaranathan et al., 2019]. GeoES constructs its proposals from the geometry of the feasible variables. Natural evolution strategies use the Fisher geometry of a search distribution [Wierstra et al., 2014]. Appendix F describes our NES baselines and their Fisher scaling.

Black-box language model tuning. Low-dimensional parameterizations support prompt and adapter tuning [Aghajanyan et al., 2021, Lester et al., 2021, Hu et al., 2022]. BBT uses CMA-ES for prompt search [Sun et al., 2022b], and BBTv2 extends black-box prompt tuning [Sun et al., 2022a]. Uni-LoRA maps low-dimensional vectors to adapter updates [Li et al., 2025]. Weight normalization and DoRA separate magnitude and direction in neural network weights [Salimans and Kingma, 2016, Liu et al., 2024]. Our experiments apply this separation to the latent variables of prompt and adapter search. We also compare with MeZO, HiZOO, and FZOO [Malladi et al., 2023, Zhao et al., 2025, Dang et al., 2026] under a common query budget.

3 Preliminaries

We minimize f over a manifold $\mathcal{M}$ using tangent projections and retractions to generate trials and objective values to select between them. Appendix A.1 gives concrete formulas for the sphere and explains the subspace representation used for Grassmann PCA.

Let $\mathcal{M} \subset \mathbb{R}^n$ be a compact smooth submanifold without boundary, with the metric induced by the ambient inner product. Its intrinsic dimension is $d = \dim \mathcal{M} \geq 1$. We write $T_x\mathcal{M}$ for the tangent space, $\mathrm{proj}(x, \cdot)$ for the orthogonal projection onto it, and $\|\cdot\|$ for the induced norm. The Riemannian gradient is the tangent vector satisfying $\mathrm{D}f(x)[\xi] = \langle \mathrm{grad}\, f(x), \xi \rangle$ for every $\xi \in T_x\mathcal{M}$.

Definition 1 (Retraction [Absil et al., 2008, Boumal, 2023])**.** *A retraction is a smooth map* $\mathrm{Retr} : \mathcal{D} \subset T\mathcal{M} \to \mathcal{M}$*, where* $\mathcal{D}$ *is an open neighborhood of the zero section, such that*

$$\mathrm{Retr}_x(0_x) = x, \qquad \mathrm{D}\,\mathrm{Retr}_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}\,.$$

Here $\mathrm{id}_{T_x\mathcal{M}}$ *is the identity on* $T_x\mathcal{M}$*.*

The first identity keeps a zero step at the current point. The second preserves the proposed tangent direction to first order. For example, on the unit sphere a tangent vector satisfies $\langle x, \xi \rangle = 0$, and normalization gives the retraction $(x + \xi)/\|x + \xi\|$. Compactness gives a uniform radius $R > 0$ such that

$$B_x(R) := \{\xi \in T_x\mathcal{M} : \|\xi\| \leq R\} \subset \mathcal{D}_x \quad \text{for every } x \in \mathcal{M}, \tag{1}$$

where $\mathcal{D}_x = \{\xi : (x, \xi) \in \mathcal{D}\}$. Appendix A.2 proves this fact.

The function $\xi \mapsto f(\mathrm{Retr}_x(\xi))$ expresses the objective in the tangent space at x and is called a pullback. We assume that its linear approximation has a quadratic remainder on the trial region.

Assumption 1 (Pullback smoothness)**.** *The function* $f : \mathcal{M} \to \mathbb{R}$ *is continuously differentiable. For some* $L_g > 0$ *and a radius* R *satisfying Eq.* (1)*,*

$$|f(\mathrm{Retr}_x(\xi)) - f(x) - \langle \mathrm{grad}\, f(x), \xi \rangle| \leq \frac{L_g}{2}\|\xi\|^2 \quad (x \in \mathcal{M},\ \xi \in B_x(R)). \tag{2}$$

The upper bound in this condition is the pullback smoothness assumption used in Riemannian complexity analysis [Boumal et al., 2019]. If f is twice continuously differentiable, a smooth retraction and compactness give the two-sided bound on a sufficiently small tangent ball. Appendix A.3 gives the argument.

Compactness also gives the finite quantities

$$f^\star = \min_{x \in \mathcal{M}} f(x), \qquad G = \max_{x \in \mathcal{M}} \|\mathrm{grad}\, f(x)\|, \qquad \Delta_0 = f(x_0) - f^\star. \tag{3}$$

GeoES uses exact objective comparisons and no gradients. Expected descent uses the upper bound in Eq. (2), and the concentration result in Appendix D uses the two-sided form.

4 GeoES: Geometry-aware $(1+1)$-ES

GeoES uses the tangent space to define local search directions. At x_k, it samples $v_k \sim \mathcal{N}(0, \mathrm{I}_n)$ and sets $u_k = \mathrm{proj}(x_k, v_k)$. In an orthonormal basis of $T_{x_k}\mathcal{M}$, the coordinates of u_k are distributed as $\mathcal{N}(0, \mathrm{I}_d)$. In particular,

$$\mathbb{E}[\|u_k\|^2 \mid x_k] = d. \tag{4}$$

The same distribution can be sampled directly in tangent coordinates. A retraction maps the trial back to the manifold, and selection retains the point with the smaller objective value.

Fix a step-size rule before the run. Let $\mathcal{F}_k$ contain the history before drawing the current direction. The rule selects a positive step size using only this history, with a fixed upper bound:

$$\sigma_k \text{ is } \mathcal{F}_k\text{-measurable}, \qquad 0 < \sigma_k \le \bar{\sigma} < \infty \quad \text{almost surely.} \tag{5}$$

Our default is $\sigma_k = \sigma_0/\sqrt{k+1}$ with $\bar{\sigma} = \sigma_0$. The main theorem treats deterministic schedules. Appendix C.2 gives bounds for rules based on past iterations.

Algorithm 1 GeoES: geometry-aware $(1+1)$-ES

Input: $x_0 \in \mathcal{M}$, a radius R satisfying Assumption 1, and K
Evaluate and cache $f(x_0)$
for $k = 0, \dots, K-1$ **do**
  Select σ_k satisfying Eq. (5)
  Sample $v_k \sim \mathcal{N}(0, \mathrm{I}_n)$ and set $u_k = \mathrm{proj}(x_k, v_k)$
  Set $x_{k+1} = x_k$
  if $\sigma_k \|u_k\| \le R$ **then**
    Set $y_k = \mathrm{Retr}_{x_k}(\sigma_k u_k)$ and evaluate $f(y_k)$
    if $f(y_k) \le f(x_k)$ **then**
      Set $x_{k+1} = y_k$ and cache $f(y_k)$
    end if
  end if
  Retain the cached objective value of x_{k+1}
end for
Output: x_K and its cached objective value

A step longer than R is discarded without evaluating the objective. Every evaluated trial lies in the region of the pullback bound. The radial cutoff preserves the symmetry between opposite tangent directions used in the analysis. After the initial evaluation, each iteration needs at most one new objective evaluation.

5 Theoretical Analysis

Selection turns a symmetric proposal into expected descent by retaining the parent whenever the trial is worse. We quantify this effect before choosing a step-size schedule. Let $g_k = \mathrm{grad}\, f(x_k)$ and write $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$.

5.1 Expected progress from one trial

For $U \sim \mathcal{N}(0, \mathrm{I}_d)$ and a fixed unit vector e, define

$$\alpha_d(r) = \mathbb{E}[|\langle e, U\rangle| \mathbf{1}\{\|U\| \le r\}], \qquad r > 0. \tag{6}$$

Rotational symmetry makes this quantity independent of e. It measures the absolute directional projection retained after truncation. It is positive, nondecreasing, and tends to $\sqrt{2/\pi}$ as $r \to \infty$.

Lemma 1 (Expected descent)**.** *Under Assumption 1, for every positive $\mathcal{F}_k$-measurable σ_k,*

$$\mathbb{E}_k[f(x_{k+1}) - f(x_k)] \le -\frac{\alpha_d(R/\sigma_k)}{2}\sigma_k\|g_k\| + \frac{L_g d}{4}\sigma_k^2. \tag{7}$$

Proof idea. For a trial that passes truncation, selection takes the smaller of the parent and trial values. It therefore does at least as well as using the trial only when $\langle g_k, u_k \rangle < 0$. On that half of the tangent space, the expected linear term is $-\sigma_k \alpha_d (R/\sigma_k)\|g_k\|/2$. Symmetry bounds the expected quadratic remainder by $(L_g \sigma_k^2/2)(d/2)$. The case $g_k = 0$ follows from monotonicity. Appendix B.1 gives the pointwise argument. □

The bound separates directional progress from the local model error. Selection preserves the favorable component along the gradient, giving a term proportional to $\|g_k\|$. The quadratic remainder depends on the squared length of the proposal and therefore contains d. The bound is strictly negative when $\|g_k\| > L_g d\sigma_k/(2\alpha_d(R/\sigma_k))$. A smaller step lowers this threshold, allowing the bound to quantify progress at smaller gradient norms.

5.2 Step-size schedules and convergence

Theorem 1 (General step-size bound)**.** *Suppose Assumption 1 holds. Let $\{\sigma_k\}$ be a deterministic sequence with $0 < \sigma_k \le \bar{\sigma} < \infty$, and set $\alpha_0 = \alpha_d(R/\bar{\sigma})$. Then the iterates of Algorithm 1, using this sequence, satisfy*

$$\frac{\sum_{k=0}^{K-1} \sigma_k \mathbb{E}\|g_k\|}{\sum_{k=0}^{K-1} \sigma_k} \le \frac{2\Delta_0 + (L_g d/2)\sum_{k=0}^{K-1} \sigma_k^2}{\alpha_0 \sum_{k=0}^{K-1} \sigma_k} \quad (K \ge 1). \tag{8}$$

The proof sums Eq. (7) and uses $f(x_K) \ge f^\star$. The two terms in Eq. (8) have distinct roles. The accumulated step sizes must grow to remove the initial objective gap, while the squared step sizes control the accumulated model error. Thus the weighted average of the expected gradient norm tends to zero whenever

$$\sum_{k=0}^{K-1} \sigma_k \longrightarrow \infty, \qquad \frac{\sum_{k=0}^{K-1} \sigma_k^2}{\sum_{k=0}^{K-1} \sigma_k} \longrightarrow 0.$$

The Robbins–Monro conditions are sufficient for these limits. The default inverse square root schedule also satisfies them, even though its squared step sizes have a divergent sum. Equivalently, sample an iterate with probability proportional to its deterministic step size. The left side of Eq. (8) is its expected gradient norm. Selecting an iterate with these probabilities requires no gradient evaluations.

Corollary 1 (Default schedule)**.** *For $\sigma_k = \sigma_0/\sqrt{k+1}$, take $\bar{\sigma} = \sigma_0$. Then, for every $K \ge 1$,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\| \le \frac{2}{\alpha_0\sqrt{K}}\left(\frac{\Delta_0}{\sigma_0} + \frac{L_g d \sigma_0}{4}(1 + \log K)\right). \tag{9}$$

A constant step $0 < \sigma \le \bar{\sigma}$ instead gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|g_k\| \le \frac{2\Delta_0}{\alpha_0 K \sigma} + \frac{L_g d \sigma}{2\alpha_0}. \tag{10}$$

Corollary 2 (Evaluation complexity)**.** *Under the assumptions of Theorem 1, fix $\bar{\sigma} > 0$ and set $\alpha_0 = \alpha_d(R/\bar{\sigma})$. Suppose $\Delta_0 > 0$ and $\varepsilon > 0$. For an integer*

$$K \ge \max\left\{1, \frac{4\Delta_0}{L_g d \bar{\sigma}^2}, \frac{4 L_g d \Delta_0}{\alpha_0^2 \varepsilon^2}\right\},$$

the constant step $\sigma = 2\sqrt{\Delta_0/(L_g d K)}$ gives $\mathbb{E}\|\operatorname{grad} f(x_I)\| \le \varepsilon$, where I is independent of the run and uniform on $\{0, \dots, K-1\}$. The run uses at most $K+1$ objective evaluations.

The choice of σ balances the initial gap and the quadratic error in Eq. (10). With the problem and truncation constants fixed, the resulting bound is $\mathcal{O}(d\varepsilon^{-2})$. The stochastic three points method has the same ε^{-2} order for an expected gradient norm criterion in Euclidean space and evaluates two symmetric trials per iteration [Bergou et al., 2020]. Appendix C.1 proves Corollary 2 and gives the corresponding squared-gradient bounds.

5.3 Step-size, truncation, and dimension

For a tangent Gaussian, the root mean squared step length is $\sigma_k\sqrt{d}$. Choosing σ_0 proportional to $1/\sqrt{d}$ therefore keeps the initial step length on the same scale across dimensions, as in our Gaussian PCA experiments. The radius must also preserve enough directional progress. Gaussian concentration gives

$$\frac{R}{\bar{\sigma}} \geq \sqrt{d} + 2 \quad \Longrightarrow \quad \alpha_0 \geq \sqrt{2/\pi} - e^{-1} > 0. \tag{11}$$

For a fixed valid radius R, this condition can be met by choosing $\bar{\sigma} \leq R/(\sqrt{d}+2)$. It gives a dimension-independent lower bound on the truncation factor.

The dimension in the tangent Gaussian is the number of local degrees of freedom. For example, a point on $\mathrm{Gr}(2, N)$ can be stored as an $N \times 2$ orthonormal basis and embedded isometrically as $UU^\top/\sqrt{2}$ in the symmetric matrices. This ambient space has dimension $N(N+1)/2$, while the tangent dimension is $2(N-2)$. Carrying the objective and retraction through this isometry preserves the intrinsic search distribution. Appendix E gives bounds on L_g, G, and Δ_0 independent of N for the Grassmann PCA family used below.

Appendix C treats parameters chosen from past iterations and a sufficient decrease rule. Appendix D gives a concentration bound.

6 Experimental Results

The experiments examine three aspects of GeoES. Smooth sphere and Grassmann PCA connect the analysis to stationarity and dimension scaling. Sparse PCA compares search quality under evaluation and time budgets. Language model tuning tests the effect of the search parameterization and compares optimizers under limited evaluation budgets. The appendix also studies the truncation radius and step-size adaptation. Every objective evaluation is counted, including evaluations used by populations and finite differences.

6.1 Stationarity and Dimension Scaling

Stationarity on the sphere. To measure the stationarity quantity in Theorem 1, we use the smooth sphere objective

$$f(x) = -\|A^\top x\|_2^2 + \sum_{i=1}^{n} \sqrt{x_i^2 + \varepsilon}, \qquad x \in \mathbb{S}^{n-1}, \qquad \varepsilon = 10^{-3},$$

where $A \in \mathbb{R}^{n\times 5}$ has independent standard Gaussian entries. This is a smooth surrogate of the sparse PCA objective studied in Section 6.2. With the normalization retraction, it satisfies the pullback condition on the compact sphere. We run GeoES for 1000 iterations with $\sigma_k = \sigma_0/\sqrt{k+1}$, $\sigma_0 = 1.5/\sqrt{d}$, and $R = 2$. Results are averaged over 10 independent runs with fresh draws of A and x_0.

We record the weighted gradient norm $\bar{G}_k = \sum_{j=0}^{k} \sigma_j \|\operatorname{grad} f(x_j)\| / \sum_{j=0}^{k} \sigma_j$. For deterministic step sizes, its expectation is the quantity bounded in Eq. (8) with $K = k+1$. Gradients are used only for this diagnostic, not for the search.

GeoES follows nearly identical objective trajectories on sparse PCA and its smooth surrogate (Figure 1(a)). For the surrogate, $\bar{G}_k$ decreases after an initial rise, with a fitted log–log slope of -0.27 for $k \geq 300$. The rescaling $\sqrt{k}\bar{G}_k/(1+\log k)$, motivated by Eq. (8), levels off after $k \approx 500$. The decreasing stationarity measure shows that objective comparisons alone can guide the search toward smaller gradient norms. Appendix G.4 reports this measure across dimensions.

A controlled dimension study. We use smooth PCA on $\mathrm{Gr}(2, N)$ for $N \in \{25, 50, 100, 200, 500, 1000, 2500, 5000, 10000\}$. A point $[U]$ is the subspace spanned by an orthonormal matrix $U \in \mathbb{R}^{N\times 2}$. Replacing U by UQ for an orthogonal $Q \in \mathbb{R}^{2\times 2}$ leaves the point unchanged [Edelman et al., 1998]. For a fixed target basis $U_\star$, we use

$$f([U]) = -1 - \frac{1}{2}\|U_\star^\top U\|_F^2, \qquad \Delta([U]) = f([U]) + 2 = \frac{1}{2}\|\sin\Theta(U, U_\star)\|_F^2. \tag{12}$$

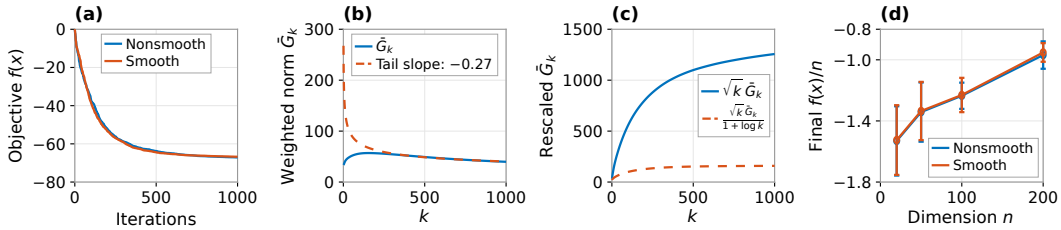


Figure 1: Sparse PCA and its smooth surrogate, both run with $\sigma_0 = 1.5/\sqrt{d}$ and $R = 2$ over 10 runs. **(a)** Mean objective versus iterations at $n = 50$. **(b)** Mean weighted gradient norm $\bar{G}_k$ of the smooth surrogate at $n = 50$, with a least-squares fit on log–log axes for $k \geq 300$. **(c)** Rescaled versions of $\bar{G}_k$. **(d)** Final objective divided by n for $n \in \{20, 50, 100, 200\}$. Error bars show one standard deviation.

Here Θ contains the two principal angles between the subspaces, so Δ measures their mismatch. The objective is deterministic and smooth on a compact manifold of intrinsic dimension $d = 2(N-2)$. Appendix E verifies the assumptions with constants independent of N. The implementation stores an $N \times 2$ basis, which the isometric scaled projector embedding relates to the embedded setting of the theory.

We use ten paired seeds with shared instances and initial points across methods. With 20000 evaluations, the median subspace error rises beyond $N = 500$. With $1 + 20d$ evaluations, it stays between 0.00259 and 0.00398 across all nine dimensions (Figure 2). A budget proportional to d therefore maintains similar accuracy while the objective scale stays fixed. This supports the predicted role of the intrinsic dimension. Moreover, $\|\operatorname{grad} f([U])\|_F^2 \leq 2\Delta([U])$ links small subspace error to small gradient norm, complementing the direct stationarity measurement on the sphere.

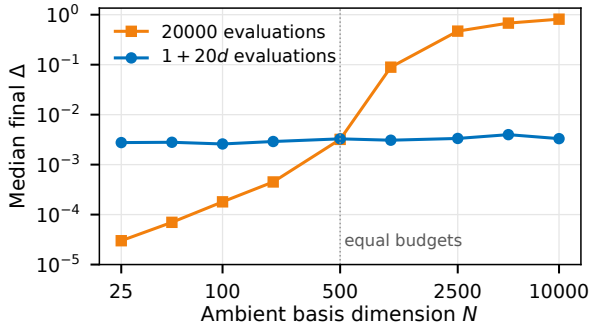


Figure 2: Grassmann PCA: median final subspace error Δ of GeoES over ten paired seeds, on log–log axes. The budgets are 20000 and $1 + 20d$ objective evaluations, with $d = 2(N-2)$. They are nearly equal at $N = 500$.

Comparing search methods. We compare twelve methods, including the CMA variant used in sparse PCA, full covariance CMA-ES [Hansen and Ostermeier, 2001], and sep-CMA-ES [Ros and Hansen, 2008]. Under $1 + 20d$ evaluations, GeoES has the lowest median error at all five dimensions shown in Table 1 and at eight of the nine tested dimensions overall. At $N = 1000$, increasing the budget from 20000 to $1 + 20d$ reduces the median error from 0.08918 to 0.00308 for GeoES and from 0.04552 to 0.00436 for full CMA-ES. At the larger budget, median runtimes are 4.58 s for GeoES, 294.09 s for full CMA-ES, and 323.56 s for Nelder–Mead. GeoES thus reaches the lower error in much less time. Appendix G.1 gives the storage and runtime details.

6.2 Sparse PCA

We next study sparse PCA, where unit normalization imposes a sphere constraint. The ℓ_1 penalty makes the objective nonsmooth [Absil and Hosseini, 2019].

Table 1: Grassmann PCA with $1 + 20d$ objective evaluations. Entries are median final Δ over ten paired seeds, and bold marks the lowest value in each column. Full CMA-ES and Nelder–Mead were run through $N = 1000$ because of their storage and runtime costs (Appendix G.1).

Method	$N = 25$	$N = 100$	$N = 500$	$N = 1000$	$N = 10000$
GeoES	**0.00276**	**0.00259**	**0.00328**	**0.00308**	**0.00331**
$(1, 1)$-ES	0.81148	0.95499	0.99169	0.99529	0.99949
$(1 + 1)$-NES	0.55002	0.69375	0.87473	0.94403	0.99753
$(1 + 1)$-NES-NGD	0.15825	0.14475	0.10938	0.10947	0.26385
ZO-RGD	0.86598	0.97770	0.99582	0.99676	0.99988
RAZGD	0.82550	0.96256	0.99075	0.99220	0.99979
PSO	0.24320	0.34562	0.47970	0.54194	0.69577
Nelder–Mead	0.42514	0.87926	0.97929	0.98719	–
CMA variant (memory 3)	0.79379	0.96776	0.99295	0.99430	0.99829
Full CMA-ES	0.02454	0.00914	0.00523	0.00436	–
sep-CMA-ES	0.02172	0.00907	0.00464	0.00413	0.00365
DE	0.52672	0.64151	0.84200	0.86020	0.92596

Setup and compared methods. We optimize over $\mathcal{M} = \mathbb{S}^{n-1} \subset \mathbb{R}^n$ with the induced metric and the normalization retraction $\mathrm{Retr}_x(\xi) = (x + \xi)/\|x + \xi\|$. For a Gaussian matrix $A \in \mathbb{R}^{n\times m}$ with independent $\mathcal{N}(0, 1)$ entries, the objective is

$$\min_{x\in\mathbb{S}^{n-1}} f(x) \coloneqq -\|A^\top x\|_2^2 + \sum_{i=1}^{n} |x_i|. \tag{13}$$

We fix $m = 5$ and compare methods at $n \in \{25, 50, 100, 200, 500\}$. Each run draws a fresh A and starts from $x_0 \sim \mathrm{Unif}(\mathbb{S}^{n-1})$. GeoES uses $\sigma_k = \sigma_0/\sqrt{k + 1}$ with $\sigma_0 = 1/\sqrt{d}$, $d = n - 1$, and $R = 1$.

We compare with $(1, 1)$-ES, $(1+1)$-NES (Algorithm 2), its natural gradient variant $(1+1)$-NES-NGD (Algorithm 3), ZO-RGD [Li et al., 2023], RAZGD [He et al., 2024], PSO [Kennedy and Eberhart, 1995], Nelder–Mead [Nelder and Mead, 1965], a CMA variant, and differential evolution [Storn and Price, 1997]. The CMA variant uses reduced covariance with memory 3 and is based on sep-CMA-ES [Ros and Hansen, 2008]. Appendix G.2 gives the settings and implementation details. All methods run in MATLAB on an Apple M4 Pro MacBook Pro with 14 CPU cores and 48 GB of RAM.

Comparisons across dimensions. Several methods reach similar objective values at $n \leq 100$, while GeoES and $(1 + 1)$-NES-NGD reach the lowest values at $n = 200$ and $n = 500$ (Figure 3). GeoES also substantially improves on $(1, 1)$-ES, which accepts every trial that passes truncation. With the same proposals and step-size schedule, retaining the parent therefore improves the search.

Full covariance CMA-ES gives a further comparison in Appendix G.5. At $n = 500$, GeoES and full CMA-ES reach median objectives of -529.41 and -525.36 within 20000 evaluations, and -529.41 and -366.16 within one second. GeoES thus reaches comparable objective quality with faster progress under the time budget. Success-based step sizes further improve its objective at $n = 100$ and $n = 500$ (Appendix G.7). At $n = 500$, fewer than 0.1% of proposals are truncated for any tested radius, so the radius rarely changes which trials are evaluated (Appendix G.6).

6.3 Black-Box Tuning with Geometry-Aware Search

Language model tuning lets us study geometry introduced by the search parameterization. We represent each latent vector by its direction and magnitude and use a normalized tangent proposal to update them jointly.

Setup. We follow the BBT protocol [Sun et al., 2022b] with a frozen $\mathrm{RoBERTa}_{\mathrm{LARGE}}$ backbone [Liu et al., 2019], manual templates and label words, and 16 labeled examples per class. The seven tasks are SST-2, Yelp Polarity, AG's News, DBPedia, MRPC, SNLI, and RTE [Socher et al., 2013, Zhang et al., 2015, Dolan and Brockett, 2005, Bowman et al., 2015, Wang et al., 2019]. The main comparisons use a budget of 8000 objective evaluations. Section 6.3 also evaluates MeZO with larger budgets. All language model experiments run on 4 NVIDIA A100 40GB GPUs.

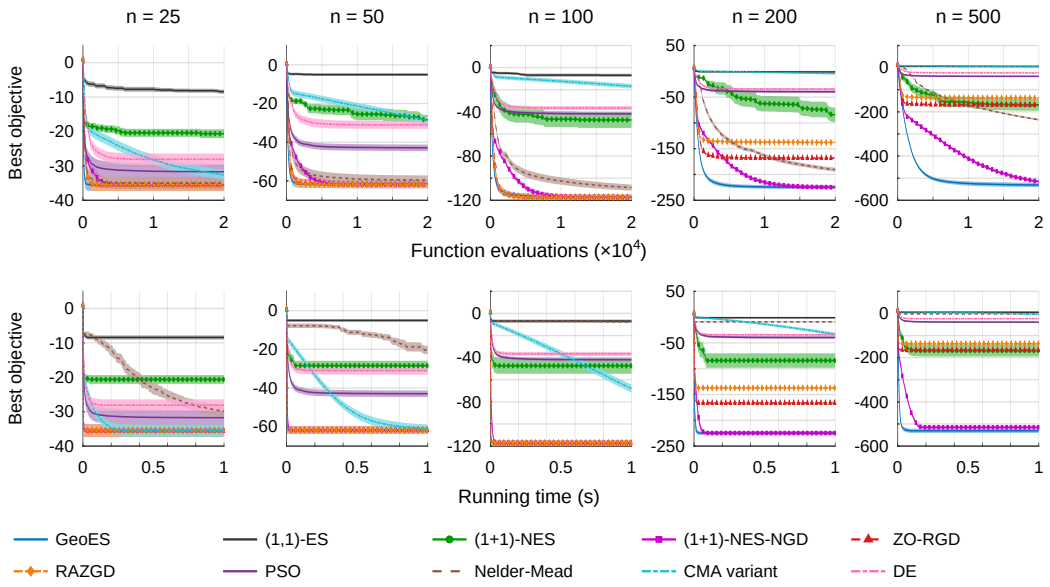


Figure 3: Sparse PCA at $n \in \{25, 50, 100, 200, 500\}$ with $m = 5$. Top: mean best objective evaluated versus function evaluations. Bottom: the same quantity versus wall-clock time. Means are over 10 runs, and shading shows half a standard deviation. Every objective evaluation is counted.

For prompt tuning, a fixed map $p(z) = Az + p_0$ converts a low-dimensional vector into a prompt [Sun et al., 2022b,a]. Uni-LoRA uses fixed random maps to convert such vectors into adapter updates [Li et al., 2025]. We write each nonzero vector as $z = ru$, with $u \in \mathbb{S}^{d-1}$ and $r > 0$, and store $(\log r, u)$. Each proposal draws a Gaussian direction in $\mathbb{R} \times T_u\mathbb{S}^{d-1}$ and normalizes it to length σ_k. The scalar component updates $\log r$, the tangent component updates u by normalization, and both are accepted or rejected together. BBTv2 and Uni-LoRA update blocks in turn. Appendix G.9 gives the implementation details.

Why separate direction and magnitude? For an additive perturbation δz, the first-order change in the direction $z/\|z\|$ is $(\mathrm{I} - uu^\top)\delta z/r$. The same perturbation therefore changes the direction less when r is larger. Proposing a tangent step at the unit vector u removes this dependence on the current magnitude. The scalar update can change the magnitude at the same time. This separation follows the idea used in weight normalization and DoRA [Salimans and Kingma, 2016, Liu et al., 2024].

Effect of the search parameterization. Table 2 compares GeoES with Euclidean $(1 + 1)$-ES in BBT. Both use the same prompt map, evaluation budget, and parent selection rule. Their mean scores differ by less than 0.4 points on SST-2, Yelp Polarity, and AG's News. GeoES obtains higher means on DBPedia, MRPC, SNLI, and RTE, with a seven-task average of 72.91 compared with 68.84. These gains show that the choice of trial generation can materially affect a search that uses the same objective comparison to select trials.

Comparison with CMA-ES. Table 2 compares GeoES with full covariance CMA-ES [Hansen and Ostermeier, 2001] from pycma [Hansen et al., 2019]. Both use the same blocks and update order, with at most 500 search dimensions per block (Appendix G.9).

With 8000 evaluations, the seven-task averages of GeoES and full CMA-ES are 72.91 and 71.59 for BBT, 76.64 and 77.09 for BBTv2, and 77.22 and 77.88 for Uni-LoRA. GeoES therefore achieves similar average scores with one trial per iteration and optimizer state linear in the block dimension. Full CMA-ES maintains a covariance matrix for each block.

Zeroth-order methods and query budgets. We compare Uni-LoRA with GeoES against MeZO + LoRA [Malladi et al., 2023], HiZOO + LoRA [Zhao et al., 2025], and FZOO + LoRA [Dang et al.,

Table 2: Language understanding benchmarks with at most 8000 objective evaluations. All experiments use RoBERTa$_{\text{LARGE}}$ in the 16-shot setting and report mean $\pm$ standard deviation over five seeds. For each task, the best result is highlighted in **bold** and the runner-up is underlined. The Euclidean $(1+1)$-ES row uses the same prompt map and parent selection rule as BBT (GeoES). The Search coords. column counts stored coordinates. BBT (GeoES) stores a unit vector in $\mathbb{R}^{500}$ and one scalar magnitude, giving 500 independent degrees of freedom. * denotes results sourced from prior work.

Method	Search coords.	SST-2 acc	Yelp P. acc	AG's News acc	DBPedia acc	MRPC F1	SNLI acc	RTE acc
Gradient-Free Methods								
Manual Prompt*	0	79.82	89.65	76.96	41.33	67.40	31.11	51.62
BBT (CMA-ES)	500	86.88 ± 2.08	91.43 ± 1.20	82.23 ± 1.49	78.74 ± 2.58	66.45 ± 7.40	45.68 ± 1.80	49.75 ± 2.57
BBT (Euclidean $(1+1)$-ES)	500	85.90 ± 1.36	91.02 ± 0.18	81.17 ± 0.93	62.29 ± 6.61	74.68 ± 4.32	39.22 ± 3.62	47.58 ± 4.58
BBT (GeoES)	501	86.05 ± 3.49	91.09 ± 0.92	81.56 ± 1.02	76.80 ± 2.15	78.22 ± 1.31	43.99 ± 2.85	52.64 ± 1.79
BBTv2 (CMA-ES)	12K	88.72 ± 0.52	91.85 ± 1.08	83.52 ± 0.87	**92.80** ± 1.85	71.64 ± 12.97	54.26 ± 4.80	56.83 ± 6.53
BBTv2 (GeoES)	12K	87.46 ± 1.57	92.06 ± 1.40	83.92 ± 1.35	85.53 ± 1.62	**78.54** ± 2.92	53.81 ± 3.30	55.16 ± 6.34
Uni-LoRA (CMA-ES)	23K	**89.94** ± 0.64	94.33 ± 1.18	83.86 ± 1.00	91.43 ± 1.21	68.83 ± 7.44	**58.10** ± 5.19	58.70 ± 8.21
Uni-LoRA (GeoES)	23K	89.56 ± 1.51	**94.56** ± 0.49	**84.82** ± 1.26	84.58 ± 3.20	72.52 ± 7.88	52.18 ± 4.48	**62.31** ± 5.56

2026] within 8000 calls, and also report MeZO at 100k and 200k calls. Appendix G.9 gives the FZOO settings and evaluation counts.

Table 3: Zeroth-order language model tuning at different query budgets. Entries are mean $\pm$ standard deviation over five seeds. MRPC uses F_1, and the other tasks use accuracy. Bold marks the highest mean among the methods with at most 8000 calls. Avg is the unweighted mean of the seven task scores. The GeoES row is repeated from Table 2.

Method	Evals.	SST-2	Yelp P.	AG's News	DBPedia	MRPC	SNLI	RTE	Avg
MeZO + LoRA	$\leq$8k	72.75 ± 6.31	87.52 ± 1.27	80.46 ± 2.09	70.50 ± 6.44	68.22 ± 9.07	**57.02** ± 1.92	59.28 ± 3.81	70.82
HiZOO + LoRA	$\leq$8k	70.14 ± 8.53	86.62 ± 1.36	76.94 ± 2.41	56.06 ± 7.00	65.52 ± 12.78	55.36 ± 5.35	57.91 ± 5.29	66.94
FZOO + LoRA	$\leq$8k	84.27 ± 0.61	90.91 ± 1.15	76.77 ± 1.57	43.33 ± 0.98	**77.60** ± 1.59	38.84 ± 0.24	53.36 ± 0.54	66.44
Uni-LoRA + GeoES	$\leq$8k	**89.56** ± 1.51	**94.56** ± 0.49	**84.82** ± 1.26	**84.58** ± 3.20	72.52 ± 7.88	52.18 ± 4.48	**62.31** ± 5.56	**77.22**
MeZO + LoRA	100k	81.28 ± 2.83	87.77 ± 1.69	81.10 ± 2.64	93.10 ± 1.42	70.16 ± 7.67	62.92 ± 6.49	62.50 ± 8.84	76.98
MeZO + LoRA	200k	81.58 ± 3.61	86.72 ± 1.39	82.22 ± 1.20	94.18 ± 0.85	71.26 ± 11.24	63.42 ± 1.00	62.50 ± 7.33	77.41

Within 8000 evaluations, Uni-LoRA with GeoES has the highest mean on five tasks, FZOO on MRPC, and MeZO on SNLI. Its largest gains over MeZO at this budget are on SST-2 and DBPedia. The average of 77.22 is also close to MeZO at 100k and 200k evaluations, with averages of 76.98 and 77.41. The larger budgets substantially improve MeZO on DBPedia and SNLI. The results compare each optimizer together with its adapter parameterization: Uni-LoRA for GeoES and LoRA for the other methods.

7 Conclusion and Limitations

We introduced GeoES to use known constraints and chosen parameterizations in black-box optimization while retaining one trial and an objective comparison. The analysis explains how selection turns symmetric tangent proposals into expected progress and why the local model error depends on the intrinsic dimension. Smooth PCA examines stationarity and dimension scaling. Sparse PCA shows competitive objective values and rapid progress within time budgets, while language model tuning demonstrates the benefit of separating direction and magnitude. These results support using geometry to generate trials while keeping selection based on an objective comparison.

Limitations. The theory assumes compactness, smooth pullbacks, Gaussian tangent proposals, and exact comparisons. The truncation factor depends on the radius and dimension. Noisy comparisons, nonsmooth objectives, and adaptive covariance require further analysis. The sparse PCA and language model experiments extend beyond the theoretical setting, and the latter cover one backbone and seven tasks.

Acknowledgments and Disclosure of Funding

This work was supported by MBZUAI internal funding. The authors declare no competing interests.

References

P-A Absil and Seyedehsomayeh Hosseini. A collection of nonsmooth Riemannian optimization problems. In *Nonsmooth Optimization and Its Applications*, pages 1–15. Springer, 2019.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, 2021.

Dirk V Arnold. On the use of evolution strategies for optimization on spherical manifolds. In *International Conference on Parallel Problem Solving from Nature*, pages 882–891. Springer, 2014.

El Houcine Bergou, Eduard Gorbunov, and Peter Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM Journal on Optimization*, 30(4):2726–2749, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(42):1455–1459, 2014. URL `https://www.manopt.org`.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, 2015.

Sebastian Colutto, Florian Fruhauf, Matthias Fuchs, and Otmar Scherzer. The CMA-ES on Riemannian manifolds to reconstruct shapes in 3-D voxel images. *IEEE Transactions on Evolutionary Computation*, 14(2):227–245, 2010.

Sizhe Dang, Yangyang Guo, Yanjun Zhao, Xiaodong Zheng, Guang Dai, Ivor Tsang, and Haishan Ye. FZOO: Fast zeroth-order optimizer for fine-tuning large language models towards Adam-scale speed. In *International Conference on Learning Representations*, 2026.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

Tobias Glasmachers. Global convergence of the (1+1) evolution strategy to a critical point. *Evolutionary Computation*, 28(1):27–53, 2020.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

Nikolaus Hansen, Youhei Akimoto, and Petr Baudis. CMA-ES/pycma on Github. Zenodo, 2019.

Chang He, Zhaoye Pan, Xiao Wang, and Bo Jiang. Riemannian accelerated zeroth-order algorithm: Improved robustness and lower query complexity. In *International Conference on Machine Learning*, pages 17972–18009. PMLR, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95 – International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order Riemannian derivative estimation and optimization. *Mathematics of Operations Research*, 48(2):1183–1211, 2023.

Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Zeroth-order Riemannian averaging stochastic approximation algorithms. *SIAM Journal on Optimization*, 34(4):3314–3341, 2024.

Kaiyang Li, Shaobo Han, Qing Su, Wei Li, Zhipeng Cai, and Shihao Ji. Uni-LoRA: One vector is all you need. In *Advances in Neural Information Processing Systems*, 2025.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 32100–32121, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4264–4273. PMLR, 2019.

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*, volume 36, 2023.

Daiki Morinaga, Kazuto Fukuchi, Jun Sakuma, and Youhei Akimoto. Convergence rate of the (1+1)-evolution strategy with success-based step-size adaptation on convex quadratic functions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1169–1177, 2021.

John A. Nelder and Roger Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965. doi: 10.1093/comjnl/7.4.308.

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Raymond Ros and Nikolaus Hansen. A simple modification in CMA-ES achieving linear time and space complexity. In *Parallel Problem Solving from Nature – PPSN X*, volume 5199 of *Lecture Notes in Computer Science*, pages 296–305. Springer, 2008. doi: 10.1007/978-3-540-87700-4_30.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

Patrick Spettel and Hans-Georg Beyer. On the design of a matrix adaptation evolution strategy for optimization on general quadratic manifolds. *ACM Transactions on Evolutionary Learning*, 2(3): 1–32, 2022.

Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997. doi: 10.1023/A:1008202821328.

Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. BBTv2: Towards a gradient-free future with large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022a.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International Conference on Machine Learning*, 2022b.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(1):949–980, 2014.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.

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

Appendix Contents

A Geometric and Gaussian Facts **15**

A.1 Geometry used by GeoES . . . 15

A.2 Uniform retraction radius . . . 15

A.3 Pullback regularity . . . 15

A.4 Tangent Gaussian moments . . . 16

A.5 Drift retained by truncation . . . 16

B Proofs of the Main Results **16**

B.1 Proof of Lemma 1 . . . 16

B.2 General step sizes and the default schedule . . . 17

C Step Sizes, Evaluation Complexity, and Acceptance **17**

C.1 Constant steps and stationarity targets . . . 17

C.2 Step sizes and radii chosen from past iterations . . . 18

C.3 A sufficient decrease rule . . . 18

D Concentration of the Stationarity Bound **19**

E Examples That Satisfy the Assumptions **20**

F NES Baselines in Tangent Coordinates **20**

F.1 NES updates . . . 20

F.2 Fisher scaling in tangent coordinates . . . 21

G Additional Experimental Results **21**

G.1 Grassmann PCA: Evaluation Protocol . . . 21

G.2 Sparse PCA: Method Settings . . . 23

G.3 Sparse PCA: Single-Method Dimension Scaling . . . 23

G.4 Smooth PCA: Stationarity across Dimensions . . . 24

G.5 Sparse PCA: Full CMA-ES and Larger Dimensions . . . 24

G.6 Sensitivity to the Truncation Radius . . . 25

G.7 Sparse PCA: Step-Size Rules . . . 27

G.8 Language Model Hyperparameters . . . 27

G.9 Language Model Tuning: Implementation Details . . . 27

A Geometric and Gaussian Facts

A.1 Geometry used by GeoES

GeoES needs a way to generate local directions and a way to map them to feasible points. A tangent space supplies the directions, and a retraction performs the mapping. The formulas below illustrate these operations in the settings used in this paper [Absil et al., 2008, Boumal, 2023].

The sphere. For $x \in \mathbb{S}^{n-1}$, tangent vectors are orthogonal to x. The tangent space, orthogonal projection, and normalization retraction are

$$T_x\mathbb{S}^{n-1} = \{\xi : \langle x, \xi\rangle = 0\}, \qquad \mathrm{proj}(x, v) = v - \langle x, v\rangle x, \qquad \mathrm{Retr}_x(\xi) = \frac{x+\xi}{\|x+\xi\|}.$$

Projection removes the radial component of a proposal. Normalization then gives a point on the sphere. Because $\langle x, \xi\rangle = 0$, the denominator satisfies $\|x+\xi\|^2 = 1 + \|\xi\|^2$ and is always positive. This is the geometry used for the sparse PCA search.

Gradients and the local model. The Riemannian gradient describes the first-order change along feasible directions. If f has a smooth ambient extension $\tilde{f}$, then $\mathrm{grad}\, f(x) = \mathrm{proj}(x, \nabla\tilde{f}(x))$. The pullback $\xi \mapsto f(\mathrm{Retr}_x(\xi))$ expresses the objective as a function on the tangent space. Its derivative at zero is $\langle \mathrm{grad}\, f(x), \xi\rangle$. Assumption 1 bounds the error of this linear approximation. A zero Riemannian gradient means that every tangent direction has zero first-order change. GeoES uses none of these derivatives to select a trial.

Subspaces and their bases. A point of $\mathrm{Gr}(2, N)$ is a two-dimensional subspace. An orthonormal matrix $U \in \mathbb{R}^{N\times 2}$ represents that subspace, and UQ represents the same point for any orthogonal 2×2 matrix Q. Tangent directions can be represented by matrices ξ satisfying $U^\top\xi = 0$. This condition excludes changes that only rotate the basis within its current subspace. Appendix E gives the retraction, the connection to the projector embedding, and the objective used in the experiments.

Dimension and computation. The dimension d counts independent tangent directions. It is $n-1$ on the sphere and $2(N-2)$ on $\mathrm{Gr}(2, N)$. The Gaussian moment in Lemma 2 explains why d enters the descent bound. The number of entries used to represent a point is a separate quantity and affects storage and running time. The Grassmann implementation stores an $N \times 2$ basis, while the scaled projector embedding provides the link to the embedded manifold analysis.

A.2 Uniform retraction radius

Proof of Eq. (1). If no uniform radius existed, for each integer $j \geq 1$ we could choose $x_j \in \mathcal{M}$ and $\xi_j \in T_{x_j}\mathcal{M}$ such that $\|\xi_j\| \leq 1/j$ and $(x_j, \xi_j) \notin \mathcal{D}$. Compactness gives a subsequence with $x_j \to x \in \mathcal{M}$. Then $(x_j, \xi_j) \to (x, 0_x)$ in the tangent bundle. Since $\mathcal{D}$ is an open neighborhood of the zero section, $(x_j, \xi_j) \in \mathcal{D}$ for all sufficiently large j, a contradiction. □

A.3 Pullback regularity

For the induced metric, an ambient smooth extension $\tilde{f}$ satisfies $\mathrm{grad}\, f(x) = \mathrm{proj}(x, \nabla\tilde{f}(x))$ [Absil et al., 2008, Boumal, 2023]. The pullback $f_x^\flat(\xi) = f(\mathrm{Retr}_x(\xi))$ satisfies $\mathrm{D}f_x^\flat(0)[\xi] = \langle \mathrm{grad}\, f(x), \xi\rangle$. Assumption 1 controls the remainder of this linear approximation, as in the pullback framework of Boumal et al. [2019].

To check a sufficient condition, suppose f is $\mathcal{C}^2$. Choose R so that the closed disk bundle $\{(x, \xi) : x \in \mathcal{M},\ \|\xi\| \leq R\}$ lies in $\mathcal{D}$. This set is compact. The second derivative of $f_x^\flat$ with respect to ξ is continuous there, so its operator norm has a uniform finite bound L_g. Taylor's formula with integral remainder gives

$$f_x^\flat(\xi) - f_x^\flat(0) - \mathrm{D}f_x^\flat(0)[\xi] = \int_0^1 (1-t)\mathrm{D}^2 f_x^\flat(t\xi)[\xi, \xi]\, dt.$$

Taking absolute values proves Assumption 1. For a merely $\mathcal{C}^1$ objective, the quadratic remainder bound remains an explicit assumption.

A.4 Tangent Gaussian moments

Lemma 2 (Projected Gaussian)**.** *Let P_x be the orthogonal projector onto $T_x\mathcal{M}$, $v \sim \mathcal{N}(0, \mathrm{I}_n)$, and $u = P_x v$. Then u is a standard Gaussian in tangent coordinates. For $g \in T_x\mathcal{M}$,*

$$\mathbb{E}u = 0, \qquad \mathbb{E}\|u\|^2 = d, \qquad \mathbb{E}\|u\|^4 = d(d+2), \qquad \mathbb{E}\langle g, u\rangle^2 = \|g\|^2. \tag{14}$$

Proof. Let $E \in \mathbb{R}^{n\times d}$ have orthonormal columns spanning $T_x\mathcal{M}$. Then $P_x = EE^\top$ and $U = E^\top v \sim \mathcal{N}(0, \mathrm{I}_d)$, with $u = EU$. Thus $\|u\| = \|U\|$ and the second moment is d. Independence of the coordinates gives $\mathbb{E}\|U\|^4 = 3d + d(d-1) = d(d+2)$. Finally, $\langle g, u\rangle$ is a centered Gaussian of variance $\|E^\top g\|^2 = \|g\|^2$. □

A.5 Drift retained by truncation

Lemma 3 (Truncated Gaussian projection)**.** *The function α_d in Eq.* (6) *is positive and nondecreasing on $(0, \infty)$, and $\lim_{r\to\infty} \alpha_d(r) = \sqrt{2/\pi}$. For every $\tau > 0$,*

$$\alpha_d(\sqrt{d} + \tau) \geq \sqrt{2/\pi} - \exp(-\tau^2/4). \tag{15}$$

For the default schedule, the probability p_k that a trial passes truncation satisfies

$$p_k \geq 1 - \exp\Big[-\tfrac{1}{2}\big(R/\sigma_k - \sqrt{d}\big)_+^2\Big], \qquad p_k \longrightarrow 1, \tag{16}$$

where $a_+ = \max\{a, 0\}$.

Proof. Write $Z = \langle e, U\rangle$. Positivity follows from the positive Gaussian density on any ball of positive radius. Monotonicity is immediate, and dominated convergence gives the limit $\mathbb{E}|Z| = \sqrt{2/\pi}$. Cauchy–Schwarz gives

$$\alpha_d(r) = \sqrt{2/\pi} - \mathbb{E}[|Z|\mathbf{1}\{\|U\| > r\}] \geq \sqrt{2/\pi} - \sqrt{\mathbb{P}(\|U\| > r)}.$$

The Gaussian concentration inequality for the Euclidean norm, together with $\mathbb{E}\|U\| \leq \sqrt{d}$, gives

$$\mathbb{P}(\|U\| > \sqrt{d} + \tau) \leq e^{-\tau^2/2} \quad (\tau \geq 0)$$

[Boucheron et al., 2013]. This proves Eqs. (15) and (16). The last limit uses $\sigma_k \to 0$. □

B Proofs of the Main Results

B.1 Proof of Lemma 1

Proof. Condition on $\mathcal{F}_k$, so x_k, g_k, and σ_k are fixed. Define $t_k = \mathbf{1}\{\sigma_k\|u_k\| \leq R\}$. On the event $t_k = 1$, elitist selection gives

$$f(x_{k+1}) - f(x_k) = \min\{0, f(y_k) - f(x_k)\}.$$

For any real number a and any indicator h, $\min\{0, a\} \leq ha$. Taking $h = \mathbf{1}\{\langle g_k, u_k\rangle < 0\}$ and using the upper pullback bound, we obtain

$$f(x_{k+1}) - f(x_k) \leq t_k\mathbf{1}\{\langle g_k, u_k\rangle < 0\}\left(\sigma_k\langle g_k, u_k\rangle + \frac{L_g}{2}\sigma_k^2\|u_k\|^2\right). \tag{17}$$

When $t_k = 0$, both sides are zero, so this inequality holds for every trial. No evaluation of an infeasible retraction is used in this argument.

If $g_k \neq 0$, symmetry under $u_k \mapsto -u_k$ gives

$$\mathbb{E}_k[t_k\mathbf{1}\{\langle g_k, u_k\rangle < 0\}\langle g_k, u_k\rangle] = -\frac{1}{2}\mathbb{E}_k[t_k|\langle g_k, u_k\rangle|] = -\frac{1}{2}\alpha_d(R/\sigma_k)\|g_k\|,$$

$$\mathbb{E}_k[t_k\mathbf{1}\{\langle g_k, u_k\rangle < 0\}\|u_k\|^2] = \frac{1}{2}\mathbb{E}_k[t_k\|u_k\|^2] \leq \frac{d}{2}.$$

Taking expectations in Eq. (17) proves the result. If $g_k = 0$, selection gives $f(x_{k+1}) - f(x_k) \leq 0$, which also proves Eq. (7). □

The halfspace groups directions by the sign of their linear objective change. The algorithm uses objective comparisons to decide which trials to retain. Related randomized direct-search analyses use symmetric directions to control the expected linear term [Bergou et al., 2020]. Here a single trial and the retained parent suffice for the bound.

B.2 General step sizes and the default schedule

Proof of Theorem 1. Since $\sigma_k \le \bar{\sigma}$, $\alpha_d(R/\sigma_k) \ge \alpha_0$. Take total expectations in Eq. (7), sum over $0 \le k < K$, and use $\mathbb{E} f(x_K) \ge f^\star$ to obtain

$$\frac{\alpha_0}{2} \sum_{k=0}^{K-1} \sigma_k \mathbb{E}\|g_k\| \le \Delta_0 + \frac{L_g d}{4} \sum_{k=0}^{K-1} \sigma_k^2. \tag{18}$$

Division by $\alpha_0 \sum_k \sigma_k/2$ proves Eq. (8). Also, $\|g_k\|^2 \le G\|g_k\|$ gives a bound on the weighted average of $\mathbb{E}\|g_k\|^2$ by multiplying the right side of Eq. (8) by G. □

For a deterministic sequence, let I_K be independent of the run, with $\mathbb{P}(I_K = k) = \sigma_k / \sum_{j=0}^{K-1} \sigma_j$. The left side of Eq. (8) equals $\mathbb{E}\| \operatorname{grad} f(x_{I_K})\|$. This sampling interpretation requires no gradient evaluations. If $\sum_k \sigma_k = \infty$ and $\sum_{k<K} \sigma_k^2 / \sum_{k<K} \sigma_k \to 0$, the bound tends to zero. The usual Robbins–Monro conditions $\sum_k \sigma_k = \infty$ and $\sum_k \sigma_k^2 < \infty$ are one sufficient choice [Robbins and Monro, 1951].

Proof of Corollary 1. For $0 \le k < K$, $\sigma_k \ge \sigma_0/\sqrt{K}$, and $\sum_{k<K} \sigma_k^2 \le \sigma_0^2(1+\log K)$. Substituting these inequalities in Eq. (18) gives

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\|g_k\| \le \frac{2}{\alpha_0 \sqrt{K}} \left(\frac{\Delta_0}{\sigma_0} + \frac{L_g d \sigma_0}{4} (1 + \log K) \right).$$

This proves Eq. (9). The squared-gradient bound follows from $\|g_k\|^2 \le G\|g_k\|$. □

C Step Sizes, Evaluation Complexity, and Acceptance

C.1 Constant steps and stationarity targets

Setting $\sigma_k = \sigma$ in Eq. (18) proves Eq. (10). For a fixed positive step, this gives

$$\limsup_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\|g_k\| \le \frac{L_g d \sigma}{2\alpha_0}.$$

This is an upper bound on the average stationarity measure.

Fix $\bar{\sigma} > 0$ independently of K and set $\alpha_0 = \alpha_d(R/\bar{\sigma})$. Suppose $\Delta_0 > 0$ and $K \ge 4\Delta_0/(L_g d \bar{\sigma}^2)$. The choice

$$\sigma = 2\sqrt{\frac{\Delta_0}{L_g d K}}$$

then satisfies $\sigma \le \bar{\sigma}$. For an independent uniform index $I \in \{0, \dots, K-1\}$, substitution gives

$$\mathbb{E}\| \operatorname{grad} f(x_I)\| \le \frac{2}{\alpha_0} \sqrt{\frac{L_g d \Delta_0}{K}}. \tag{19}$$

Thus it suffices to take an integer

$$K \ge \max\left\{1, \frac{4\Delta_0}{L_g d \bar{\sigma}^2}, \frac{4 L_g d \Delta_0}{\alpha_0^2 \varepsilon^2}\right\} \tag{20}$$

to ensure $\mathbb{E}\| \operatorname{grad} f(x_I)\| \le \varepsilon$. The total number of objective evaluations is at most $K + 1$. The balanced choice describes a theoretical horizon-dependent tuning rule. The default schedule does not require L_g or Δ_0. If $\Delta_0 = 0$, the initial point is a global minimizer and all retained points are stationary, so the stationarity conclusions hold directly.

For comparison with squared-gradient criteria, use $\mathbb{E}\| \operatorname{grad} f(x_I)\|^2 \le G\mathbb{E}\| \operatorname{grad} f(x_I)\|$. For the same constant-step construction, the last term in Eq. (20) becomes

$$\frac{4G^2 L_g d \Delta_0}{\alpha_0^2 \varepsilon^2} \quad \text{for } \mathbb{E}\| \operatorname{grad} f(x_I)\|^2 \le \varepsilon,$$

and

$$\frac{4G^2 L_g d\Delta_0}{\alpha_0^2 \varepsilon^4} \quad \text{for } \mathbb{E}\| \operatorname{grad} f(x_I)\|^2 \le \varepsilon^2.$$

For the target $\mathbb{E}\| \operatorname{grad} f(x_I)\|^2 \le \varepsilon^2$, the default schedule gives $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ through Corollary 1, with problem constants held fixed. These stationarity bounds concern the randomly selected iterate x_I. The algorithm returns x_K, whose objective value is the smallest observed during the run.

C.2 Step sizes and radii chosen from past iterations

Let σ_k and R_k be chosen from $\mathcal{F}_k$ before drawing the current tangent Gaussian. Suppose, almost surely,

$$0 < \sigma_k \le \bar{\sigma}, \qquad 0 < R_k \le R, \qquad R_k/\sigma_k \ge \rho > 0.$$

Apply truncation with R_k. The proof of Lemma 1 gives the same conditional bound with $\alpha_d(R_k/\sigma_k) \ge \alpha_d(\rho)$. Taking expectations and summing yields

$$\frac{\alpha_d(\rho)}{2} \sum_{k=0}^{K-1} \mathbb{E}[\sigma_k \|g_k\|] \le \Delta_0 + \frac{L_g d}{4} \sum_{k=0}^{K-1} \mathbb{E}[\sigma_k^2]. \tag{21}$$

Both σ_k and g_k depend on the past, so their product remains inside the expectation.

This bound permits adaptation based on past successes. A concrete way to preserve the default rate is to constrain the resulting step size to

$$\frac{\sigma_0}{2\sqrt{k+1}} \le \sigma_k \le \frac{2\sigma_0}{\sqrt{k+1}}.$$

With a fixed radius R, take $\rho = R/(2\sigma_0)$. Then Eq. (21) gives

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\|g_k\|^2 \le \frac{4G}{\alpha_d(\rho)\sigma_0\sqrt{K}} \left(\Delta_0 + L_g d\sigma_0^2(1+\log K)\right).$$

Other adaptation rules require control of their accumulated step sizes. The tangent sampling in this argument remains isotropic.

For the radius, one admissible rule is $R_k = \min\{R, \sigma_k(\sqrt{d}+2)\}$. It gives $R_k/\sigma_k \ge \min\{R/\bar{\sigma}, \sqrt{d}+2\} > 0$. Thus the radius may track the trial scale while respecting the region where the retraction and pullback bound hold. If both are valid for every tangent step, truncation can be removed. In that case $\alpha_d(\infty) = \sqrt{2/\pi}$.

C.3 A sufficient decrease rule

For a trial that passes truncation, consider the computable test

$$f(y_k) \le f(x_k) - c\sigma_k^2 \|u_k\|^2, \qquad c > 0. \tag{22}$$

This compares the achieved decrease with the squared trial length. It uses the same objective evaluations as the original rule and does not require knowledge of L_g.

Proposition 1 (Sufficient decrease)**.** *Under Assumption 1, the rule in Eq.* (22) *satisfies*

$$\mathbb{E}_k[f(x_{k+1}) - f(x_k)] \le -\frac{\alpha_d(R/\sigma_k)}{2}\sigma_k\|g_k\| + \frac{L_g + 2c}{4} d\sigma_k^2. \tag{23}$$

Consequently, the expectation bounds above hold with L_g replaced by $L_g + 2c$. If h_k indicates that the trial passes truncation and is accepted, then

$$c \sum_{k=0}^{K-1} h_k \sigma_k^2 \|u_k\|^2 \le \Delta_0 \tag{24}$$

for every realized run.

Proof. On $t_k = 1$, write $D = f(y_k) - f(x_k)$ and $q = c\sigma_k^2\|u_k\|^2$. The accepted increment is $D\mathbf{1}\{D + q \le 0\}$, and

$$D\mathbf{1}\{D + q \le 0\} \le \min\{0, D + q\}.$$

Use the halfspace argument from Eq. (17), now with

$$D + q \le \sigma_k\langle g_k, u_k\rangle + \left(\frac{L_g}{2} + c\right)\sigma_k^2\|u_k\|^2.$$

Symmetry gives Eq. (23). Rejected and truncated trials have zero increment, and $g_k = 0$ is again covered by monotonicity. Finally, sum the accepted inequalities $f(x_k) - f(x_{k+1}) \ge ch_k\sigma_k^2\|u_k\|^2$ to obtain Eq. (24). □

The additional guarantee controls the total squared length of accepted steps. The stationarity order is unchanged, while increasing c makes the acceptance test stricter at a given state and trial.

D Concentration of the Stationarity Bound

We retain the default schedule and write

$$Y_k = f(x_{k+1}) - f(x_k), \qquad D_k = Y_k - \mathbb{E}_k Y_k, \qquad M_K = \sum_{k=0}^{K-1} D_k.$$

Monotonicity gives $-\Delta_0 \le Y_k \le 0$, so $|D_k| \le \Delta_0$. The two-sided pullback bound and the Gaussian moments give

$$\mathbb{E}_k Y_k^2 \le 2\sigma_k^2\|g_k\|^2 + \frac{L_g^2}{2}\sigma_k^4 d(d+2). \tag{25}$$

Indeed, the increment is zero after rejection or truncation. For an accepted trial its absolute value is at most $\sigma_k|\langle g_k, u_k\rangle| + (L_g/2)\sigma_k^2\|u_k\|^2$. Squaring and using $(a+b)^2 \le 2a^2 + 2b^2$ proves the bound. Consequently,

$$\sum_{k=0}^{K-1} \mathbb{E}_k D_k^2 \le \bar{V}_K, \qquad \bar{V}_K = 2G^2\sigma_0^2(1 + \log K) + \frac{L_g^2\pi^2}{12}\sigma_0^4 d(d+2). \tag{26}$$

Theorem 2 (Concentration)**.** *Under Assumption 1, use the default schedule and $\alpha_0 = \alpha_d(R/\sigma_0)$. For each fixed $K \ge 1$ and $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|g_k\|^2 \le \frac{2G}{\alpha_0\sigma_0\sqrt{K}}\left(\Delta_0 + \frac{L_g d\sigma_0^2}{4}(1 + \log K) + \sqrt{2\bar{V}_K\log(1/\delta)} + \frac{2\Delta_0}{3}\log(1/\delta)\right). \tag{27}$$

Proof. The case $\Delta_0 = 0$ is immediate, so assume $\Delta_0 > 0$. The sequence M_K is a martingale with differences bounded above by Δ_0 and predictable variance bounded by Eq. (26). The scalar Freedman inequality [Freedman, 1975], in the form stated by Tropp [2011, Theorem 1.1], yields

$$\mathbb{P}(M_K \ge t) \le \exp\left(-\frac{t^2}{2(\bar{V}_K + \Delta_0 t/3)}\right).$$

It follows that, with probability at least $1 - \delta$,

$$M_K \le \sqrt{2\bar{V}_K\log(1/\delta)} + \frac{2\Delta_0}{3}\log(1/\delta).$$

Since $Y_k = \mathbb{E}_k Y_k + D_k$, the descent lemma gives pathwise

$$\frac{\alpha_0}{2}\sum_{k=0}^{K-1}\sigma_k\|g_k\| \le \Delta_0 + \frac{L_g d}{4}\sum_{k=0}^{K-1}\sigma_k^2 + M_K.$$

Apply $\sigma_k \ge \sigma_0/\sqrt{K}$, $\|g_k\|^2 \le G\|g_k\|$, and $\sum_{k<K}\sigma_k^2 \le \sigma_0^2(1 + \log K)$. □

The variance bound grows only logarithmically with K. The resulting fluctuation term therefore preserves the $\widetilde{\mathcal{O}}(K^{-1/2})$ stationarity rate.

E Examples That Satisfy the Assumptions

Smooth objectives on the sphere. On $\mathcal{M} = \mathbb{S}^{n-1}$, use $\mathrm{Retr}_x(\xi) = (x+\xi)/\|x+\xi\|$ for $\langle x, \xi\rangle = 0$. Since $\|x+\xi\|^2 = 1 + \|\xi\|^2$, the retraction is smooth for every tangent vector. A $\mathcal{C}^2$ objective has the required uniform pullback bound on any fixed finite tangent ball by Appendix A.3. This covers the smooth sparse PCA surrogate used for the stationarity diagnostic. The objective with an ℓ_1 penalty is studied in the sparse PCA application.

Grassmann PCA. Represent a point of $\mathrm{Gr}(2, N)$ by an orthonormal matrix $U \in \mathbb{R}^{N\times 2}$. For a fixed orthonormal $U_\star$, let

$$f([U]) = -1 - \frac{1}{2}\|U_\star^\top U\|_F^2, \qquad f^\star = -2.$$

The function is invariant under $U \mapsto UQ$ for orthogonal Q and is smooth on the compact Grassmann manifold. A horizontal tangent vector satisfies $U^\top \xi = 0$. The polar retraction is represented by

$$\mathrm{Retr}_{[U]}(\xi) = \left[(U+\xi)(\mathrm{I}_2 + \xi^\top\xi)^{-1/2}\right].$$

For the standard Grassmann metric, $[U] \mapsto UU^\top/\sqrt{2}$ is an isometric embedding into the symmetric matrices [Boumal, 2023]. Its differential has squared Frobenius norm $\|\xi\|_F^2$. The intrinsic dimension is $d = 2(N-2)$, and the symmetric-matrix space has dimension $N(N+1)/2$.

Proposition 2 (Uniform constants for Grassmann PCA)**.** *For this objective and retraction,* $0 \le f([U]) - f^\star \le 1$ *and* $\|\operatorname{grad} f([U])\|_F \le 1/\sqrt{2}$*. The two-sided pullback bound holds for every horizontal tangent vector with* $L_g = 2$*, independently of* N*.*

Proof. Write $P_\star = U_\star U_\star^\top$. The horizontal gradient is $-(\mathrm{I}_N - UU^\top)P_\star U$. If θ_1, θ_2 are the principal angles between the two subspaces, then

$$f([U]) - f^\star = \frac{1}{2}\sum_{i=1}^{2}\sin^2\theta_i, \qquad \|\operatorname{grad} f([U])\|_F^2 = \sum_{i=1}^{2}\sin^2\theta_i\cos^2\theta_i \le \frac{1}{2}.$$

These identities give the first two bounds.

For the pullback remainder, set $M = \xi^\top\xi$, $S = (\mathrm{I}_2 + M)^{-1}$, and $Q = \mathrm{I}_2 - S$. Also write $B = U^\top P_\star U$, $H = U^\top P_\star \xi$, and $D = \xi^\top P_\star \xi$. Direct substitution in the objective gives

$$f(\mathrm{Retr}_{[U]}(\xi)) - f([U]) - \langle \operatorname{grad} f([U]), \xi\rangle = \frac{1}{2}\operatorname{tr}(BQ - DS) + \operatorname{tr}(HQ).$$

Here $0 \preceq B \preceq \mathrm{I}_2$, $0 \preceq D \preceq M$, and $Q = MS$. Consequently, both $\operatorname{tr}(BQ)$ and $\operatorname{tr}(DS)$ lie in $[0, \operatorname{tr} M]$. Moreover, $Q^2 \preceq M/4$, since $m^2/(1+m)^2 \le m/4$ for every $m \ge 0$. Thus $\|Q\|_F \le \|\xi\|_F/2$ and $\|H\|_F \le \|\xi\|_F$. The absolute remainder is at most

$$\frac{1}{2}\operatorname{tr} M + \|H\|_F\|Q\|_F \le \|\xi\|_F^2.$$

This is Assumption 1 with $L_g = 2$ for every tangent step. □

The reported geometric error is $\Delta([U]) = f([U]) - f^\star = \frac{1}{2}\|\sin\Theta(U, U_\star)\|_F^2$. It satisfies $\|\operatorname{grad} f([U])\|_F^2 \le 2\Delta([U])$, so a small geometric error implies a small gradient norm. The converse need not hold, since stationary subspaces need not be optimal. The retraction and pullback bound are global in this example, so truncation can also be removed. With finite truncation, choosing $R/\bar\sigma \ge \sqrt{d} + 2$ keeps $\alpha_d(R/\bar\sigma)$ bounded away from zero uniformly in N.

F NES Baselines in Tangent Coordinates

F.1 NES updates

The two NES baselines use the same tangent geometry as Algorithm 1, but update the mean and scale using the objective value of the trial. Their update formulas are motivated by isotropic Gaussian

NES [Wierstra et al., 2014]. They use radial clipping of the trial vector and of the mean update. Algorithm 1 instead discards trials outside the radius and retains the parent unless the objective decreases.

For $U \sim \mathcal{N}(0, \mathrm{I}_d)$ in $T_x\mathcal{M}$, the clipped offspring is

$$Y = \mathrm{Retr}_x\left(\sigma U \min\left\{1, \frac{R}{\sigma\|U\|}\right\}\right),$$

with tangent step zero when $U = 0$. This defines the average offspring objective

$$\widetilde{f}_\sigma(x) = \mathbb{E}[f(Y)]. \tag{28}$$

The implementations maintain a moving baseline b_k and an exponentially weighted average ν_k^2 of squared objective deviations. They use this average to normalize the objective difference:

$$\delta_k = f(y_k) - b_k, \qquad w_k = \frac{\delta_k}{\sqrt{\nu_{k+1}^2 + 10^{-12}}},$$

where ν_{k+1}^2 includes the current squared deviation, as specified in Algorithm 2. The plain update directions are

$$g_{\mu,k} = \frac{w_k}{\sigma_k} u_k, \qquad g_{\ell,k} = w_k(\|u_k\|^2 - d), \qquad \ell_k = \log \sigma_k. \tag{29}$$

These update directions use Gaussian score formulas weighted by the normalized objective difference. Clipping, normalization by the current sample, and the dependence of the retraction on its base point distinguish them from exact score estimators of the offspring distribution. The Fisher calculation below describes their scaling in fixed tangent coordinates.

F.2 Fisher scaling in tangent coordinates

For an unclipped isotropic Gaussian location-scale family in a fixed tangent coordinate system, the Fisher blocks for the mean and log scale are [Wierstra et al., 2014]

$$\mathcal{F}_{\mu,\mu} = \sigma^{-2}\mathrm{I}_d, \qquad \mathcal{F}_{\ell,\ell} = 2d, \qquad \mathcal{F}_{\mu,\ell} = 0. \tag{30}$$

To obtain these blocks, differentiate the Gaussian log density with respect to the mean and $\ell = \log \sigma$. The scores are U/σ and $\|U\|^2 - d$, where $U \sim \mathcal{N}(0, \mathrm{I}_d)$. Their second moments follow from

$$\mathbb{E}[UU^\top] = \mathrm{I}_d, \qquad \mathbb{E}[(\|U\|^2 - d)^2] = 2d, \qquad \mathbb{E}[U(\|U\|^2 - d)] = 0.$$

The last identity follows from symmetry. These moments give the Fisher blocks in Eq. (30). Applying their inverses to the directions in Eq. (29) gives

$$\widetilde{g}_{\mu,k} = \sigma_k w_k u_k, \qquad \widetilde{g}_{\ell,k} = \frac{w_k}{2d}(\|u_k\|^2 - d). \tag{31}$$

For a given sample and normalized utility, the plain mean direction scales as $1/\sigma_k$, while the Fisher-scaled direction scales as σ_k. The log scale direction is also divided by $2d$. These changes explain the different update scales of the two baselines before clipping. Algorithm 3 uses the Fisher-scaled expressions with the same clipping and utility normalization as plain NES. The Fisher matrix here belongs to the Gaussian family in fixed tangent coordinates. Section 5 analyzes the parent selection rule in Algorithm 1.

G Additional Experimental Results

G.1 Grassmann PCA: Evaluation Protocol

We use the objective in Eq. (12) and report the smallest objective value evaluated within the budget as the subspace error $\Delta = f + 2$. GeoES uses $\sigma_k = \sigma_0/\sqrt{k+1}$ with $\sigma_0 = 1/\sqrt{d}$, $d = 2(N-2)$, and $R = 2$, where k counts proposals from zero. The budgets are 20000 and $1 + 20d$ objective evaluations, including the initial evaluation. Methods share the problem instance and initial subspace within each seed and use separately fixed algorithmic random streams. We report medians over ten paired seeds.

All methods use one global hyperparameter rule at every dimension, without per-dimension tuning. ZO-RGD uses $q = 5$ random directions. RAZGD follows Algorithm 1 of He et al. [2024] with Option I and $\epsilon = 10^{-3}$, $l = 4$, $\rho = 1$, and $b = 1$. Full CMA-ES runs pycma on the $2N$ basis coordinates and evaluates each candidate after a polar repair onto orthonormal bases.

Algorithm 2 Riemannian $(1+1)$-NES (isotropic smoothing)

Input: initial point $x_0 \in \mathcal{M}$, initial step-size $\sigma_0 > 0$, clipping radius $R > 0$, learning rates $\eta_\mu > 0$, $\eta_\sigma > 0$, baseline parameter $\beta \in (0,1)$, iteration budget K
Set $d \leftarrow \dim(\mathcal{M})$, $b_0 \leftarrow f(x_0)$, $\nu_0^2 \leftarrow 1$
for $k = 0, 1, \ldots, K-1$ **do**
  Sample $v_k \sim \mathcal{N}(0, \mathrm{I}_n)$ and set $u_k \leftarrow \mathrm{proj}(x_k, v_k) \in T_{x_k}\mathcal{M}$
  if $\|u_k\| = 0$ **then**
    $x_{k+1} \leftarrow x_k$, $\sigma_{k+1} \leftarrow \sigma_k$, $b_{k+1} \leftarrow b_k$, $\nu_{k+1}^2 \leftarrow \nu_k^2$
    continue
  end if
  Offspring generation: $\xi_k \leftarrow \sigma_k u_k$
  if $\|\xi_k\| > R$ **then**
    $\xi_k \leftarrow \dfrac{R}{\|\xi_k\|}\,\xi_k$
  end if
  $y_k \leftarrow \mathrm{Retr}_{x_k}(\xi_k)$
  Evaluate $f(y_k)$
  Baseline and normalization update:
  $\delta_k \leftarrow f(y_k) - b_k$
  $b_{k+1} \leftarrow (1-\beta) b_k + \beta f(y_k)$
  $\nu_{k+1}^2 \leftarrow (1-\beta)\nu_k^2 + \beta \delta_k^2$
  $w_k \leftarrow \dfrac{\delta_k}{\sqrt{\nu_{k+1}^2 + 10^{-12}}}$
  Mean update: $\Delta\mu_k \leftarrow -\dfrac{\eta_\mu}{\sigma_k} w_k u_k$
  if $\|\Delta\mu_k\| > R$ **then**
    $\Delta\mu_k \leftarrow \dfrac{R}{\|\Delta\mu_k\|}\,\Delta\mu_k$
  end if
  $x_{k+1} \leftarrow \mathrm{Retr}_{x_k}(\Delta\mu_k)$
  Scale update: $\sigma_{k+1} \leftarrow \sigma_k \exp\!\big(-\eta_\sigma w_k (\|u_k\|^2 - d)\big)$
end for

Storage and running time. Full CMA-ES and Nelder–Mead operate in $D = 2N$ coordinates. Full CMA-ES stores dense $D \times D$ matrices and periodically computes an eigendecomposition. Nelder–Mead stores $D + 1$ points, each with D coordinates. Their storage therefore grows quadratically with N. At $N = 10000$, one dense matrix or the simplex alone occupies about 2.98 GiB in double precision, before auxiliary arrays. At $N = 1000$, their median runtimes already reach 294.09 s and 323.56 s under the $1 + 20d$ budget (Table 4). We therefore limit these two methods to $N \leq 1000$ and use sep-CMA-ES, with linear optimizer storage, at every tested dimension. Dashes in Table 1 indicate runs not performed.

Table 4: Grassmann PCA at $N = 1000$ with a budget of $1 + 20d = 39921$ objective evaluations. Entries are medians over the same ten paired seeds.

Method	Final subspace error Δ	Running time (s)
GeoES	**0.00308**	**4.58**
Full CMA-ES	0.00436	294.09
Nelder–Mead	0.98719	323.56

GeoES uses about $1/64$ of the full CMA-ES time and $1/71$ of the Nelder–Mead time at this budget, with lower median subspace error. It stores an $N \times 2$ basis and generates one tangent trial at a time. Tangent projection and retraction work with this thin basis, without a dense covariance matrix or a simplex. This keeps the optimizer operations small when objective evaluations are inexpensive.

Algorithm 3 Riemannian $(1+1)$-NES with natural gradient descent

Input: initial point $x_0 \in \mathcal{M}$, initial step-size $\sigma_0 > 0$, clipping radius $R > 0$, NGD learning rates $\eta_\mu^{\mathrm{nat}} > 0$, $\eta_\sigma^{\mathrm{nat}} > 0$, baseline parameter $\beta \in (0,1)$, iteration budget K
Set $d \leftarrow \dim(\mathcal{M})$, $b_0 \leftarrow f(x_0)$, $\nu_0^2 \leftarrow 1$, $\ell_0 \leftarrow \log \sigma_0$
for $k = 0, 1, \ldots, K-1$ **do**
  $\sigma_k \leftarrow \exp(\ell_k)$
  Sample $v_k \sim \mathcal{N}(0, \mathrm{I}_n)$ and set $u_k \leftarrow \mathrm{proj}(x_k, v_k) \in T_{x_k}\mathcal{M}$
  if $\|u_k\| = 0$ **then**
    $x_{k+1} \leftarrow x_k$, $\ell_{k+1} \leftarrow \ell_k$, $\sigma_{k+1} \leftarrow \sigma_k$
    $b_{k+1} \leftarrow b_k$, $\nu_{k+1}^2 \leftarrow \nu_k^2$
    continue
  end if
  Offspring generation: $\xi_k \leftarrow \sigma_k u_k$
  if $\|\xi_k\| > R$ **then**
    $\xi_k \leftarrow \dfrac{R}{\|\xi_k\|}\,\xi_k$
  end if
  $y_k \leftarrow \mathrm{Retr}_{x_k}(\xi_k)$
  Evaluate $f(y_k)$
  Baseline and normalization update:
  $\delta_k \leftarrow f(y_k) - b_k$
  $b_{k+1} \leftarrow (1-\beta) b_k + \beta f(y_k)$
  $\nu_{k+1}^2 \leftarrow (1-\beta)\nu_k^2 + \beta \delta_k^2$
  $w_k \leftarrow \dfrac{\delta_k}{\sqrt{\nu_{k+1}^2 + 10^{-12}}}$
  Plain NES directions:
  $g_{\mu,k} \leftarrow \dfrac{w_k}{\sigma_k} u_k$
  $g_{\ell,k} \leftarrow w_k(\|u_k\|^2 - d)$
  Fisher scaling:
  $\widetilde{g}_{\mu,k} \leftarrow \sigma_k^2 g_{\mu,k} = \sigma_k w_k u_k$
  $\widetilde{g}_{\ell,k} \leftarrow \dfrac{1}{2d} g_{\ell,k} = \dfrac{w_k}{2d}(\|u_k\|^2 - d)$
  Mean update: $\Delta\mu_k \leftarrow -\eta_\mu^{\mathrm{nat}} \widetilde{g}_{\mu,k}$
  if $\|\Delta\mu_k\| > R$ **then**
    $\Delta\mu_k \leftarrow \dfrac{R}{\|\Delta\mu_k\|}\,\Delta\mu_k$
  end if
  $x_{k+1} \leftarrow \mathrm{Retr}_{x_k}(\Delta\mu_k)$
  Log scale update: $\ell_{k+1} \leftarrow \ell_k - \eta_\sigma^{\mathrm{nat}} \widetilde{g}_{\ell,k}$
  $\sigma_{k+1} \leftarrow \exp(\ell_{k+1})$
end for

G.2 Sparse PCA: Method Settings

Table 5 lists the settings used for Figure 3. Each method uses one fixed setting across all seeds and dimensions. Every run uses at most 20000 objective evaluations, and curves report the best value found so far. PSO and Nelder–Mead use the Manopt implementations [Boumal et al., 2014] with their default settings. These experiments run in MATLAB on an Apple M4 Pro MacBook Pro with 14 CPU cores and 48 GB of RAM. Time curves use a common wall-clock axis and remain constant after a method stops or exhausts its budget.

G.3 Sparse PCA: Single-Method Dimension Scaling

We examine how the sparse PCA objective varies with dimension for GeoES (Algorithm 1). We fix $m = 5$ and use $n \in \{20, 50, 100, 200\}$, corresponding to intrinsic dimensions $d = n - 1$. For each n, we generate a fresh matrix A, run 1000 iterations, and repeat over 10 independent runs. These

Table 5: Sparse PCA method settings. Here $d = n - 1$ is the intrinsic dimension of $\mathbb{S}^{n-1}$.

Method	Settings
GeoES	$\sigma_k = \sigma_0/\sqrt{k+1}$, $\sigma_0 = 1/\sqrt{d}$, $R = 1$
$(1, 1)$-ES	same schedule, σ_0, and R as GeoES
$(1 + 1)$-NES	$(\eta_\mu, \eta_\sigma, \beta) = (0.02, 0.01, 0.1)$
$(1 + 1)$-NES-NGD	$(\eta_\mu, \eta_\sigma, \beta) = (0.05, 0.01, 0.1)$
ZO-RGD	smoothing $\mu = 2/\sqrt{n}$, step $\eta = 0.03$, $q = 5$ directions
RAZGD	as ZO-RGD, with $\beta = 0.2$
PSO	Manopt `pso`, default settings
Nelder–Mead	Manopt `neldermead`, default settings
CMA variant	$\lambda = 20+\lfloor 3 \log d \rfloor$, $\mu = \lfloor \lambda/2 \rfloor$, memory 3, initial step 0.2 in $[10^{-3}, 1]$, $c_\sigma = 0.3$, $c_c = c_d = 0.2$
DE	population 50, $F = 0.8$, $k = 0.5$, $P_F = 0.7$

runs use the sparse PCA settings $\sigma_0 = 1/\sqrt{d}$ and $R = 1$, whereas Figure 1(d) uses $\sigma_0 = 1.5/\sqrt{d}$ and $R = 2$, so the two figures report slightly different values.

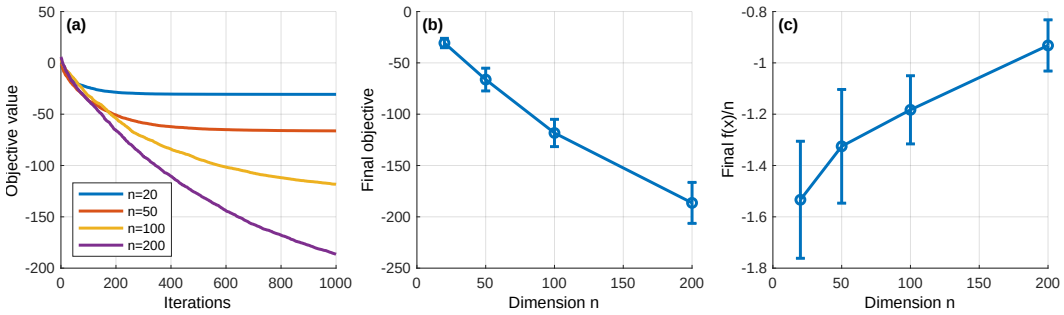


Figure 4: Sparse PCA on $\mathbb{S}^{n-1}$: single-method dimension-scaling results for GeoES. **(a)** Mean objective value $f(x_k)$ versus iterations for $n \in \{20, 50, 100, 200\}$, averaged over 10 runs. **(b)** Mean final objective value after 1000 iterations versus n; error bars denote one standard deviation. **(c)** Mean final objective per dimension $f(x)/n$ versus n; error bars denote one standard deviation.

GeoES makes rapid initial progress followed by smaller improvements (Figure 4(a)). Increasing n also changes the scale of the quadratic term in Eq. (13). The more negative final values in panel (b) reflect this change, and division by n reduces the variation in panel (c). The smooth surrogate in Section 6.1 complements these objective values with a stationarity measurement.

G.4 Smooth PCA: Stationarity across Dimensions

We use the smooth objective and experimental settings from Section 6.1. Figure 5 reports the weighted gradient norm divided by $\sqrt{d}$ at the reported terminal index, with 10 independent runs at each dimension.

The weighted gradient norm divided by $\sqrt{d}$ increases from about 2.8 at $n = 20$ to about 13 at $n = 200$, with a fitted log–log slope of 0.66. This family changes the objective scale together with the dimension: for the sampled matrices, $\mathbb{E}[A^\top A] = n\mathrm{I}_5$. Theorem 1 describes this dependence through d, L_g, Δ_0, and the truncation factor. The Grassmann family in Section 6.1 keeps the objective scale fixed and has uniform bounds on L_g, G, and Δ_0 (Appendix E). The two studies distinguish scaling the data model from increasing the search dimension with these quantities controlled.

G.5 Sparse PCA: Full CMA-ES and Larger Dimensions

We add full covariance CMA-ES and $n = 1000$ to the sparse PCA comparison. All methods here run in Python, with 10 instances per dimension ($m = 5$), independent of those in Figure 3, and at most 20000 calls. GeoES uses $\sigma_0 = 1/\sqrt{d}$ and $R = 1$. Full CMA-ES (pycma 4.4.0 [Hansen et al., 2019], $\sigma_0 = 0.2$, default population size) searches in $\mathbb{R}^n$ and normalizes each candidate onto the sphere before evaluation.

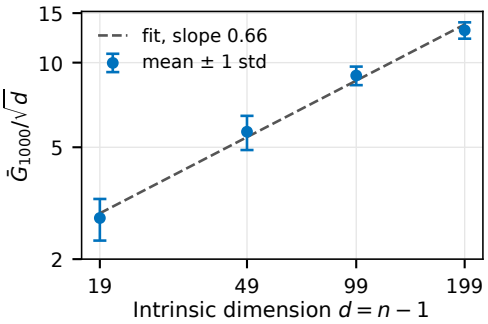


Figure 5: Normalized stationarity diagnostic $\bar{G}_{1000}/\sqrt{d}$ on the smooth PCA objective versus the intrinsic dimension, on log–log axes. Error bars show one standard deviation over 10 runs. The dashed line is a least-squares fit of the logarithm of the mean on $\log d$, with slope 0.66.

Table 6: Sparse PCA with full CMA-ES. Median best objective over 10 instances within 20000 evaluations and within 1 second. Lower is better.

	20000 Evals.			1 second		
n	GeoES	CMA variant	Full CMA-ES	GeoES	CMA variant	Full CMA-ES
25	-38.67	-37.31	-38.67	-38.67	-36.68	-38.67
50	-74.15	-38.06	-74.15	-74.15	-36.37	-74.15
100	-113.37	-10.92	-113.43	-113.37	-10.08	-113.42
200	-234.01	-6.48	-234.41	-234.01	-5.41	-230.02
500	-529.41	7.02	-525.36	-529.41	7.84	-366.16
1000	-1015.80	13.09	-1018.99	-1015.08	14.18	-236.80

Under 20000 calls, GeoES and full CMA-ES reach comparable objectives at every dimension, for example -529.41 and -525.36 at $n = 500$ and -1015.80 and -1018.99 at $n = 1000$ (Table 6). Within one second, GeoES stays within one unit of its value after 20000 evaluations, whereas full CMA-ES reaches -366.16 at $n = 500$ and -236.80 at $n = 1000$. The CMA variant remains well above both methods.

G.6 Sensitivity to the Truncation Radius

We first study the radius on symmetric positive definite (SPD) matrices, where the additive retraction $\mathrm{Retr}_X(\xi) = X + \xi$ is valid only when $X + \xi \succ 0$. We minimize $f(X) = \frac{1}{2}\|X - X_\star\|_F^2$ over 20×20 SPD matrices, so $d = 210$. The target $X_\star$ has a random orthogonal eigenbasis and eigenvalues spaced logarithmically from 0.05 to 1, and the initial point has eigenvalues in $[0.05, 1]$. GeoES uses $\sigma_k = \sigma_0/\sqrt{k+1}$ with $\sigma_0 = 1/\sqrt{d}$ and $R = c\,\sigma_0\sqrt{d}$ for $c \in \{0.01, 0.03, 0.1, 0.3, \infty\}$, where $c = \infty$ disables truncation. A proposal is truncated if $\sigma_k\|U_k\| > R$ and invalid if $X_k + \sigma_k U_k$ is not positive definite. Neither is evaluated. Each setting runs 20000 iterations, averaged over 10 runs. On this noncompact manifold, the experiment measures how truncation affects feasibility and objective progress.

In Figure 6(a), the objective stays near its initial value for about 9000 iterations at $c = 0.01$ and about 1000 at $c = 0.03$. Against evaluated trials in panel (b), both settings improve from the start, so the delay comes mainly from proposals that are never evaluated. For $c \in \{0.1, 0.3\}$, the curves are close to $R = \infty$.

About half of all proposals are truncated at $c = 0.01$, about 6% at $c = 0.03$, and almost none for $c \geq 0.1$ (Figure 7). Invalid proposals are rare. Short steps are accepted more often: the accepted fraction is about 0.22 at $c = 0.01$ and about 0.08 for $c \geq 0.1$.

For small c, the truncation frequency in Figure 8 starts near one and falls as σ_k decreases, because $\mathbb{P}(\sigma_k\|U_k\| \leq R)$ increases. Acceptance at $c = 0.01$ rises only after truncation stops. For $c \geq 0.1$, invalid trials occur mainly in the first 2000 iterations, while $\lambda_{\min}(X_k)$ approaches the boundary of the SPD cone before settling at 0.05, the smallest eigenvalue of $X_\star$.

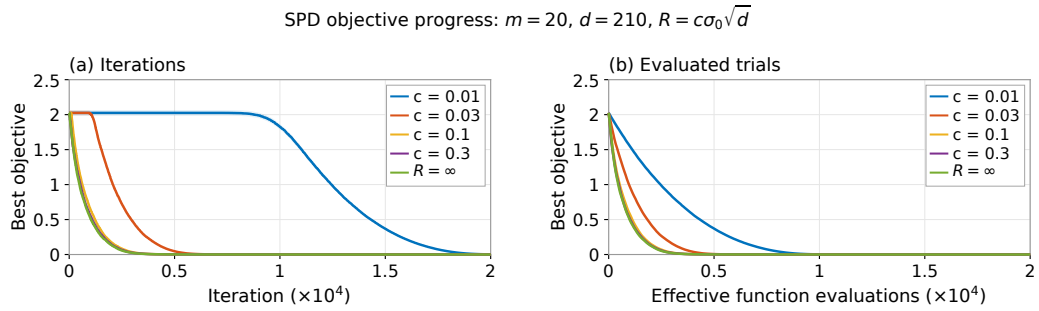


Figure 6: Truncation radius study on the SPD problem with $R = c\sigma_0\sqrt{d}$. Best objective value found so far, averaged over 10 runs, versus **(a)** iterations and **(b)** evaluated trials.

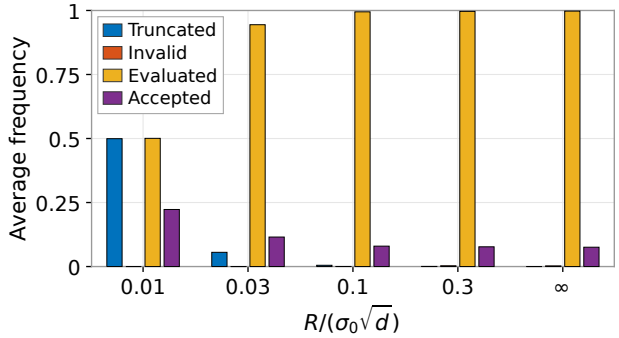


Figure 7: Fractions of truncated, invalid, evaluated, and accepted proposals over the full run on the SPD problem, averaged over 10 runs.

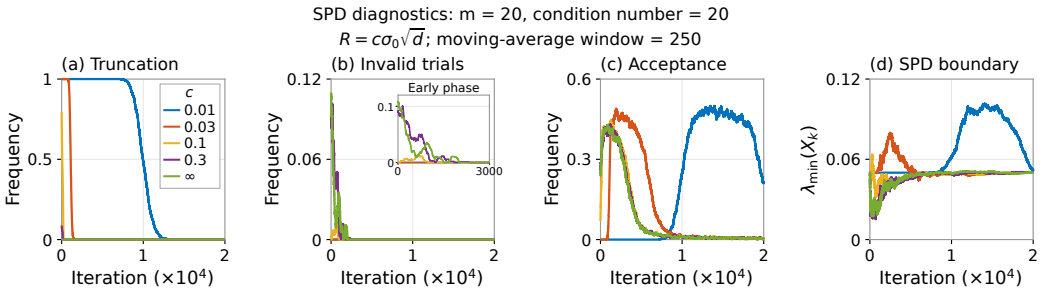


Figure 8: Diagnostics for the SPD problem, averaged over 10 runs. Panels (a)–(c) show moving averages over 250 iterations of **(a)** the truncation frequency, **(b)** the invalid-trial frequency, with the first 3000 iterations enlarged in the inset, and **(c)** the acceptance frequency. **(d)** Smallest eigenvalue $\lambda_{\min}(X_k)$ of the current iterate.

On sparse PCA at $n = 500$, with the setup and instances of Appendix G.5, we test $R \in \{0.25, 0.5, 1, 2, \infty\}$. Fewer than 0.1% of proposals are truncated at every radius, the median final objectives lie between -533.96 and -529.41, and every paired bootstrap 95% interval for the difference from $R = 1$ includes zero.

On SPD matrices, radii between $0.1\,\sigma_0\sqrt{d}$ and $0.3\,\sigma_0\sqrt{d}$ preserve progress close to that without truncation, while smaller radii reject many early proposals. On sparse PCA, truncation is rare throughout the tested range. These observations explain why the radius has a larger effect on the SPD search: it changes how many proposals reach evaluation.

G.7 Sparse PCA: Step-Size Rules

We compare four scalar step-size rules while keeping the Gaussian proposal, elitist selection, initial point, objective instance, seed, and evaluation budget fixed. The rules are the default $\sigma_k = \sigma_0/\sqrt{k+1}$, a constant step $\sigma_k = \sigma_0$, a $1/5$ success rule, and a cumulative success rule. Every 20 proposals, the $1/5$ rule multiplies σ by 1.5 when the success frequency exceeds $1/5$, divides it by 1.5 when the frequency is below $1/5$, and leaves it unchanged otherwise. The cumulative rule averages the success indicator exponentially with weight 0.1 and adjusts $\log\sigma$ using a target success rate of 0.2 and damping 2. Both adaptive rules constrain σ to $[10^{-4}, 10^{2}]\sigma_0$ and retain isotropic tangent proposals.

For the two success-based rules, the median final objective is lower than under the default schedule at $n = 100$ and $n = 500$, and the paired bootstrap 95% intervals for the differences lie below zero. The reductions are about 0.02 at $n = 100$ and between 2.2 and 5.3 at $n = 500$. At $n = 1000$, both intervals include zero. These comparisons show that adapting a scalar step can improve GeoES without changing its tangent directions or parent selection. Appendix C.2 gives the corresponding descent bound, with convergence determined by the accumulated step sizes.

G.8 Language Model Hyperparameters

Table 7 reports the hyperparameters selected by grid search for the results in Table 2.

Table 7: Selected hyperparameters for the language model experiments. Each cell reports $(\sigma_0, \log r_0)$.

Method	SST-2	Yelp P.	AG's News	DBPedia	MRPC	SNLI	RTE
BBT (GeoES)	$(1.0, 4.5)$	$(1.0, 4.0)$	$(1.0, 4.0)$	$(1.0, 4.0)$	$(1.0, 4.0)$	$(1.0, 3.0)$	$(1.0, 4.0)$
BBTv2 (GeoES)	$(1.0, 2.5)$	$(2.0, 3.0)$	$(1.0, 3.0)$	$(2.0, 3.0)$	$(1.5, 2.0)$	$(2.0, 3.0)$	$(1.0, 2.5)$
Uni-LoRA (GeoES)	$(1.2, 3.3)$	$(1.2, 3.5)$	$(1.0, 3.3)$	$(1.2, 3.5)$	$(1.2, 3.5)$	$(1.2, 3.5)$	$(1.0, 3.5)$

G.9 Language Model Tuning: Implementation Details

GeoES. Each block is stored as $(\log r, u)$ with $u \in \mathbb{S}^{d-1}$. We draw a standard Gaussian vector in $\mathbb{R}^{d+1}$, project its last d coordinates onto $T_u\mathbb{S}^{d-1}$, and normalize the resulting joint vector to length σ_k. The first component is added to $\log r$. The remaining components form a tangent step at u, which is retracted by normalization. The two components are accepted or rejected together. The joint tangent space has dimension d, so the proposed increment in $\log r$ has standard deviation $\sigma_k/\sqrt{d}$ before selection. The value $\log r_0$ in Table 7 specifies the initial magnitude.

CMA-ES. We use `cma.CMAEvolutionStrategy` from pycma with its default full covariance update. BBT optimizes one 500-dimensional vector. BBTv2 assigns one 500-dimensional vector to each of the 24 layers of RoBERTa$_{\text{LARGE}}$, and Uni-LoRA assigns one 480-dimensional vector to each of the 48 query and value modules. Each block has its own CMA-ES instance with population size 20, and the blocks are updated in turn. The initial step size is $\sigma_0 = 0.2$ for BBTv2 and Uni-LoRA. The budget of 8000 evaluations is divided evenly across blocks in whole generations, so CMA-ES uses at most 8000 calls. GeoES uses the same blocks and the same update order within the same budget.

Zeroth-order adapter tuning. For FZOO + LoRA, we use rank 8, LoRA scaling $\alpha = 16$, eight perturbation directions, learning rate 10^{-4}, perturbation scale 10^{-4}, and a constant schedule. One update evaluates the unperturbed objective and eight perturbed objectives, giving nine objective

evaluations per update and 888 updates within the budget of 8000 evaluations. The evaluation uses the same seven tasks and five seeds as the comparison with at most 8000 calls in Table 3.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction describe GeoES, state the smooth compact manifold setting of the stationarity results, and qualify the intrinsic dimension dependence by the remaining problem and truncation constants. Experimental claims identify the tested tasks, comparisons, and query budgets.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The Conclusion and Limitations section states the compactness, smoothness, and exact comparison assumptions and describes directions beyond them. Section 5 explains the role of the truncation constant and distinguishes evaluation complexity from representation costs. Section 6.3 and the conclusion describe the scope of the language model experiments.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The assumptions are stated in Section 3, and Section 5 gives the descent lemma and stationarity bounds. The appendix provides the proofs, extensions to step sizes and acceptance rules, and the concentration result. Appendix F specifies the NES baselines and derives their Gaussian Fisher scaling.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper states the objectives, datasets, metrics, number of runs, evaluation budgets, model backbone, search parameterizations, and compared methods. Appendix G.2 lists the settings of all sparse PCA methods, and Appendices G.8 and G.9 report the language model hyperparameters and the CMA-ES and FZOO settings.

   Guidelines:

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

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

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

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

   Answer: [Yes]

   Justification: Code and configuration files for sparse PCA, smooth surrogate diagnostics, radius diagnostics, and black-box tuning are available at `https://github.com/MLOpt-MBZUAI/GeoES`. The experiments use publicly available datasets and pretrained models cited in the paper. No new dataset or pretrained model is released.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Sections 6.1, 6.2, and 6.3 describe the objectives, data, metrics, evaluation budgets, and compared methods. Appendices G.1, G.2, G.8, and G.9 provide the experimental protocols, method settings, and implementation details.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: The main sparse PCA curves show means over 10 runs with bands of half a standard deviation. The language model tables report means and standard deviations over five seeds. The Grassmann PCA figure and table report medians over ten paired seeds. The dimension diagnostics specify their error bars in the figure captions.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper reports the hardware used for the sparse PCA experiments, including a MacBook Pro with an Apple M4 Pro processor and 48 GB RAM, and the hardware used for black-box tuning experiments, namely a cluster with 4 NVIDIA A100 40GB GPUs. The function-evaluation and query budgets are specified for the corresponding experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is methodological and experimental, uses standard public benchmark datasets and pretrained models, and does not involve human-subject experiments, sensitive personal data collection, or deployment of models. To the best of our knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The work may have positive impacts by improving memory- and query-efficient black-box optimization methods, which can make derivative-free tuning more accessible when gradients are unavailable. Potential negative impacts are indirect: improved black-box tuning could also make it easier to adapt language models for undesirable uses, although this paper does not release new pretrained models, datasets, or deployment systems.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release a new pretrained language model, image generator, scraped dataset, or other high-risk asset. The experiments use existing benchmark datasets and pretrained models only for evaluating optimization algorithms.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We credit the original papers and public repositories for all existing assets used in the experiments, including RoBERTa$_{\text{LARGE}}$, SST-2, MRPC, RTE, SNLI, Yelp Polarity, AG's News, DBPedia, BBT/BBTv2, MeZO, LoRA, Uni-LoRA, and derivative-free optimization baselines. We use these assets only for evaluation and do not redistribute modified datasets or pretrained model checkpoints.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: Code and configuration files are released at `https://github.com/MLOpt-MBZUAI/GeoES`. No new dataset or pretrained model is released.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or new research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or new research with human subjects, so IRB approval is not applicable.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The paper uses a frozen $\text{RoBERTa}_{\text{LARGE}}$ model as the black-box objective in the tuning experiments in Section 6.3. The proposed optimization method itself does not rely on an LLM as a methodological component, no new LLM is trained or released, and no LLM is used to generate labels or evaluate outputs. Any use of language models is limited to the experimental benchmark setting described in the paper.

    Guidelines:

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