SILVER STEPSIZE FOR FASTER ZEROTH-ORDER OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study gradient-free minimization of smooth convex functions through Silver stepsizes—a non-monotone, 2-adic schedule that accelerates gradient descent—and show how to compose it with two-point zeroth-order (ZO) estimators on a smoothed objective. We apply Silver's multi-step Lyapunov analysis to smoothed objectives and show that it carries over verbatim when gradients are replaced by unbiased two-point estimators with a tax in the form of a quadratic variance term. We control this term via an orthogonal-on-spikes batching policy that allocates directions proportionally to the Silver steps (with a cap at dimension), achieving budget-optimal variance aggregation. Empirically, we validate our approach through both numerical experiments and MeZO-style forward-pass-only fine-tuning of large language models, incorporating practical considerations such as clipping strategies, and demonstrate its superior performance.

1 INTRODUCTION

Zeroth-order (ZO, derivative-free) optimization addresses the common setting where we can query function values but cannot reliably obtain gradients: the model is a black box, gradients are prohibitively expensive or noisy, or we wish to optimize through a non-differentiable system (e.g., simulators, private APIs). This regime occurs across machine learning and scientific computing: hyperparameter and architecture tuning, black-box adversarial attacks, policy search and evolution strategies in RL, and large-model fine-tuning under tight memory budgets (Larson et al., 2019; Flaxman et al., 2005; Duchi et al., 2015; Shamir, 2017; Salimans et al., 2017; Malladi et al., 2023).

Families of ZO estimators. Modern ZO methods approximate gradients from function values using structured perturbations. (i) *One-point bandit smoothing* forms an unbiased estimator of the gradient of a smoothed objective from a single evaluation (Flaxman et al., 2005). (ii) *Two-point estimators*—our focus—use symmetric differences $f(x+\mu u) - f(x-\mu u)$ along a random direction u, achieving strictly better variance/rates and minimax-optimal guarantees for smooth convex objectives (Duchi et al., 2015; Shamir, 2017; Nesterov & Spokoiny, 2017). (iii) *Coordinate-wise finite differences* estimate partial derivatives one coordinate at a time (often $2d$ queries per gradient) and are widely used in black-box deep learning (e.g., ZOO attacks) (Chen et al., 2017). (iv) *SPSA* perturbs all coordinates simultaneously using Rademacher noise and recovers a two-evaluation gradient proxy with strong SA-style guarantees (Spall, 1992). (v) *Orthogonal batches* sample B mutually orthonormal directions (Stiefel manifold) per iteration; this reduces the estimator variance at fixed budget and unifies several schemes, including spherical smoothing and coordinate descent (Kozak et al., 2023; Feng & Wang, 2023).

Core bottlenecks in ZO. ZO estimators introduce a bias–variance tradeoff via the smoothing radius μ and sampling distribution. Even for smooth convex objectives, the best-known two-point schemes incur a statistical floor that scales with dimension under noisy queries; controlling the *variance accumulation* across iterations is the central algorithmic challenge (Duchi et al., 2015; Shamir, 2017; 2013; Jamieson et al., 2012).

A complementary acceleration lever: stepsize hedging (Silver). Independently of estimator design, recent work shows that carefully structured *stepsizes* alone can accelerate plain gradient descent on smooth convex functions. The *Silver stepsize schedule* is a simple, explicit, fractal

sequence with a 2-adic block structure. It admits a *multi-step Lyapunov certificate* ("Silver identity") which yields a convergence rate of $O\big(\varepsilon^{-\log_\rho 2}\big) = O\big(\varepsilon^{-0.7864}\big)$ iterations for gradient descent, where $\rho = 1 + \sqrt{2}$ is the silver ratio (Altschuler & Parrilo, 2023a;b; 2024). Intuitively, the schedule interleaves small steps with periodic "spikes" whose algebraic cancellation accelerates net progress across blocks.

This paper: composing Silver with two-point ZO on smoothed objectives. We bring these strands together. We run the Silver schedule on a *smoothed* objective $h = f_\mu/L$ (blockwise-constant μ), and replace exact gradients by unbiased *symmetric two-point* estimators for ∇f_μ along *orthonormal* batches of directions. The Silver identity's linear noise terms cancel in expectation, so the entire stochastic tax collapses to an explicit *quadratic variance term*, which we control by aligning batch size with the stepsize spikes ($B_t \propto \alpha_t$, capped at d). This *orthogonal-on-spikes* policy concentrates averaging where it matters most while keeping the total query budget fixed.

Motivation Two-point estimators are unbiased for ∇f_μ (not ∇f), making the smoothed problem f_μ the right analytical object. The Silver identity is robust to *conditionally unbiased* inexact gradients and only pays the quadratic term from the terminal square in the certificate—precisely what batching and blockwise μ can control. Orthogonal directions improve constants without complicating the analysis or the memory footprint (Kozak et al., 2023; Feng & Wang, 2023).

We make the following contributions in this work.

- Silver-on-smoothing with two-point ZO: We adapt the Silver multi-step analysis to $h = f_\mu/L$ with symmetric two-point estimators, showing the identity carries over verbatim with a single *variance aggregation* term $\sum_t \alpha_t^2 \mathbb{E}\|\zeta_t\|^2$ (no linear noise term).
- Variance control via orthogonal-on-spikes batching: Under a fixed query budget per block, we prove that allocating batch sizes proportional to the Silver steps ($B_t \propto \alpha_t$, capped at d) optimally controls $\sum_t \alpha_t^2/B_t$ (Cauchy–Schwarz tightness), and we instantiate this with Stiefel sampling.
- High-probability bounds via Freedman: We give a simple high-probability translation of the Silver identity with martingale differences, yielding dimension-aware tails in terms of the predictable quadratic variation.
- Practical ZO for LLM fine-tuning: We apply the method to MeZO-style forward-only full-parameter fine-tuning and discuss practical details (direction orthogonalization, clipping, memory footprint.) (Malladi et al., 2023; Hu et al., 2021; Dettmers et al., 2023).

Organization. Section 3.1 states the formal setup and notation; Section 3.2 summarizes the Silver schedule and the specific properties we use. Section **??** develops the inexact-gradient Silver identity for two-point ZO on f_μ and the variance control via orthogonal-on-spikes batching. Experiments appear in Section 5.

2 Related Work

Derivative-free / zeroth-order optimization. Classical DFO covers direct-search, model-based trust-region, and interpolation methods; recent surveys unify these with randomized finite-difference estimators used in ML (Larson et al., 2019). For convex ZO with random directions, one-point bandit smoothing dates to Flaxman et al. (2005). Two-point estimators achieve optimal rates in smooth/stochastic and adversarial settings (Duchi et al., 2015; Shamir, 2017). Nesterov & Spokoiny (2017) give a self-contained analysis with explicit smoothing constants. Building on this line of work, MeZO (Malladi et al., 2023) brings two-point, forward-only ZO into LLM fine-tuning, showing that competitive adaptation is possible with inference-level memory (no backprop activations). In this work, we build the analyse with a uniform sphere sampling for slightly tighter dimension-dependent estimation variance at high dimension.

Estimator families and variance reduction. Coordinate-wise finite differences (up to $2d$ queries/gradient) are common in black-box deep learning, e.g., ZOO (Chen et al., 2017). SPSA provides a two-evaluation coordinate-free estimator rooted in stochastic approximation (Spall, 1992). Sampling *orthogonal* directions (Stiefel manifold) reduces variance and unifies spherical and coordinate schemes (Kozak et al., 2023); refined bounds appear in Feng & Wang (2023). Variance-reduced

ZO methods (e.g., ZO-SVRG/SPIDER-SZO) are complementary and can be combined with our blockwise policy (Ji et al., 2019; Fang et al., 2018).

Zeroth-order smoothing and two-point estimators. Ball/sphere and Gaussian smoothing with two-point estimators are classical; see Flaxman et al. (2005) (one-point bandit smoothing), Duchi et al. (2015); Shamir (2017) (two-point optimal rates), and Nesterov & Spokoiny (2017) (Gaussian smoothing with explicit moment and bias constants). We emphasize the uniform *ball/sphere* pair, which gives dimension-friendly bias constants and a clean gradient identity.

Stepsize hedging / Silver schedule. The Silver schedule is a simple explicit fractal stepsize sequence that accelerates plain gradient descent in both strongly convex and smooth convex regimes. The analysis hinges on a multi-step descent identity and 2-adic structure; see Altschuler & Parrilo (2023a;b; 2024) for the arXiv and final journal versions. The rate $T^{-\log_\rho 2}$ with $\rho = 1 + \sqrt{2}$ lies between classical $O(\varepsilon^{-1})$ and Nesterov's $O(\varepsilon^{-1/2})$.

Orthogonal directions and variance reduction. Using mutually orthogonal directions (sampling on the Stiefel manifold) provably reduces estimator variance and improves constants compared to i.i.d. directions; see Kozak et al. (2023); Feng & Wang (2023).

3 Preliminaries

3.1 Problem setup and notation

We minimize a convex L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ with minimizer $x^\star$. We adopt the standard uniform-ball smoothing

$$f_\mu(x) \;:=\; \mathbb{E}_{v\sim \mathrm{Unif}(\mathbb{B}^d)} f(x+\mu v), \qquad h(x) \;:=\; f_\mu(x)/L,$$

so that h is 1-smooth and convex. Let $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. The sphere-gradient identity gives

$$\nabla f_\mu(x) \;=\; \frac{d}{\mu}\,\mathbb{E}_u\big[f(x+\mu u)\,u\big],$$

hence both the one-point $\frac{d}{\mu} f(x+\mu u)u$ and symmetric two-point

$$\widehat{g}(x;\mu,u) \;=\; \frac{d}{2\mu}\big(f(x+\mu u) - f(x-\mu u)\big)u$$

are *unbiased* for $\nabla f_\mu(x)$. We use the symmetric two-point estimator because it enjoys sharper variance/rate guarantees in smooth convex problems (Duchi et al., 2015; Shamir, 2017; Nesterov & Spokoiny, 2017).

Iteration, stepsizes, and batching. We run a Silver block of length $N = 2^k - 1$ with stepsizes $\{\alpha_t\}_{t=0}^{N-1}$ (Section 3.2), update

$$x_{t+1} \;=\; x_t - \tfrac{\alpha_t}{L}\,\widehat{g}_t, \qquad \widehat{g}_t \;=\; \frac{d}{2\mu B_t}\sum_{i=1}^{B_t}\big(f(x_t+\mu v_{t,i}) - f(x_t-\mu v_{t,i})\big)v_{t,i},$$

and use *orthogonal-on-spikes* batching $B_t = \min\{d, \lceil c_B\,\alpha_t\rceil\}$ with $V_t = [v_{t,1},\dots,v_{t,B_t}] \in \mathrm{St}(d,B_t)$ drawn via thin QR of a Gaussian matrix (Haar on the Stiefel manifold). Each step costs $2B_t$ function queries. Unless stated otherwise, we assume access to exact function values or conditionally zero-mean value noise so that $\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = \nabla f_\mu(x_t)$ with $\mathcal{F}_t$ the natural filtration.

Bias and variance constants (used later). For L-smooth f,

$$|f_\mu(x) - f(x)| \le \tfrac{L}{2}\mu^2 \cdot \tfrac{d}{d+2}, \qquad \|\nabla f_\mu(x) - \nabla f(x)\| \le \tfrac{L}{2}\,d\,\mu,$$

and for the two-point sphere estimator

$$\mathbb{E}\|\widehat{g}(x;\mu,u) - \nabla f_\mu(x)\|^2 \;\le\; 2d\,\|\nabla f(x)\|^2 \;+\; \tfrac{1}{2}d^2L^2\mu^2.$$

Averaging B_t directions divides the RHS by B_t. We keep μ *fixed within each Silver block* and may geometrically decay it across blocks.

Silver stepsizes (facts used). For a block of length $N = 2^k - 1$, the multi-step identity implies $h(x_N) - h^\star \le r_k \|x_0 - x^\star\|^2$ with explicit $r_k = \Theta(\rho^{-2k})$ and $\rho = 1 + \sqrt{2}$. Moreover, $\sum_{t=0}^{N-1} \alpha_t = \Theta(\rho^k)$. We rely only on these consequences of the identity Altschuler & Parrilo (2023b; 2024).

Uniform-ball smoothing and the sphere gradient identity. Let $v \sim \mathrm{Unif}(\mathbb{B}^d)$ and $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and define $f_\mu(x) := \mathbb{E}_v\, f(x + \mu v)$. Then f_μ is convex and L-smooth and

$$\nabla f_\mu(x) \;=\; \frac{d}{\mu}\, \mathbb{E}_u\big[f(x + \mu u)\, u\big]. \tag{1}$$

By the Descent Lemma,

$$|f_\mu(x) - f(x)| \;\le\; \tfrac{L}{2}\mu^2\, \mathbb{E}\|v\|^2 = \tfrac{L}{2}\mu^2 \cdot \tfrac{d}{d+2}.$$

Moreover (we include a short proof in the appendix),

$$\|\nabla f_\mu(x) - \nabla f(x)\| \;\le\; \tfrac{L}{2}\, d\, \mu. \tag{2}$$

This bound follows from the ball-to-sphere identity $\nabla f_\mu(x) = \frac{d}{\mu}\mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[f(x+\mu u)u]$ and the Descent Lemma; see, e.g., Lemma 4.1 and Proposition 6.5 in the self-contained derivation we follow.[1]

For comparison, under *Gaussian* smoothing, $\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{L}{2}\,(d+3)^{3/2}\mu$ (Nesterov & Spokoiny, 2017, Lemma 3).

Remark 3.1 (Default smoothing and unbiasedness). Throughout we define $f_\mu(x) = \mathbb{E}_{v\sim\mathrm{Unif}(\mathbb{B}^d)} f(x + \mu v)$. For this choice,

$$\nabla f_\mu(x) = \frac{d}{\mu}\, \mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[f(x + \mu u)\, u],$$

so both the one-point $\frac{d}{\mu} f(x+\mu u)u$ and the symmetric two-point $\frac{d}{2\mu}(f(x+\mu u) - f(x-\mu u))u$ estimators are *unbiased* for $\nabla f_\mu(x)$. This identity goes back to the divergence-theorem proof used in bandit smoothing (e.g., Flaxman et al. (2005)).[2]

Second moment (uniform sphere, two-point). For L-smooth f and $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$,

$$\mathbb{E}\big\|\widehat{g}(x;\mu,u)\big\|^2 \;\le\; C_1\, d\, \|\nabla f(x)\|^2 \;+\; C_2\, d^2\, L^2\, \mu^2,$$

with explicit constants $C_1 = 2$ and $C_2 = \frac{1}{2}$ via an elementary isotropy argument (details in the theory section). Averaging B_t directions reduces the RHS by $1/B_t$; using orthonormal batches on the Stiefel manifold often improves constants in practice (Kozak et al., 2023; Feng & Wang, 2023).

Remark 3.2 (Ball vs. Gaussian smoothing). We work with the *ball* average and use the sphere-based identity (1) to estimate ∇f_μ from function values. The gradient-bias bound for ball smoothing is $\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{1}{2} L d\, \mu$ (2), whereas for *Gaussian* smoothing it scales as $\frac{1}{2} L (d+3)^{3/2}\mu$ (Nesterov & Spokoiny, 2017, Lemma 3). This explains the better dimension dependence under ball smoothing in our analysis.

Inexact-gradient Silver identity (what we use). Running Silver on $h = f_\mu/L$ with unbiased noise $\zeta_t := (\widehat{g}_t - \nabla f_\mu(x_t))/L$ gives

$$\mathbb{E}[h(x_N) - h^\star] \;\le\; r_k\, \mathbb{E}\|x_0 - x^\star\|^2 \;+\; \sum_{t=0}^{N-1} \alpha_t^2\, \mathbb{E}\|\zeta_t\|^2,$$

i.e., with no linear noise term. The proof mirrors the exact-gradient certificate and uses only tower-property cancellations.

[1] These standard facts (including the two-point second moment below) are proved from first principles with exact constants in a concise reference; we reproduce short proofs in the appendix.

[2] We work with the *ball* definition of f_μ for tighter bias; we only use the *sphere* for the estimator.

3.2 SILVER STEPSIZE PRIMER

The Silver schedule is a deterministic stepsize sequence with a 2-adic, fractal block structure. For a block of length $N = 2^k - 1$ and a 1-smooth convex objective ϕ, the *Silver identity* (a multi-step Lyapunov certificate) implies

$$\phi(x_N) - \phi^\star \;\le\; r_k \, \|x_0 - x^\star\|^2,$$

with explicit $r_k = \Theta(\rho^{-2k})$ for $\rho = 1+\sqrt{2}$ (the silver ratio). Moreover, the steps satisfy $\sum_{t=0}^{N-1} \alpha_t = \Theta(\rho^k)$. Consequently, after $T = \Theta(2^k)$ steps, gradient descent with Silver stepsizes reaches error ε in $O\big(\varepsilon^{-\log_\rho 2}\big) = O(\varepsilon^{-0.7864})$ iterations, strictly improving upon the classical $O(1/\varepsilon)$ rate for smooth convex objectives (Altschuler & Parrilo, 2023a;b; 2024). In our analysis we apply this identity to $h = f_\mu/L$ and rely only on:

1. the block guarantee $\phi(x_N) - \phi^\star \le r_k \|x_0 - x^\star\|^2$;
2. the sum-of-steps property $\sum_{t=0}^{N-1} \alpha_t = \Theta(\rho^k)$;
3. robustness to *conditionally unbiased* inexact gradients, which adds exactly $\sum_t \alpha_t^2 \mathbb{E}\|\zeta_t\|^2$ to the RHS (no linear noise term).

4 ZO-SILVER: ALGORITHM AND THEORETICAL ANALYSIS

ROADMAP OF THIS SECTION

We first state the main guarantees at a glance: (i) an expectation-level *one-block* bound under two-point ZO on the smoothed objective with Silver stepsizes; (ii) its *budget-aligned* specialization under orthogonal-on-spikes batching; (iii) a *multi-block* (restart) bound; (iv) a *high-probability* version via Freedman. We then present the algorithm and the few elementary ingredients (unbiasedness, second moment, inexact-Silver identity, and the variance-optimal batching proposition). Proofs are short and included inline or deferred to the appendix.

4.1 INGREDIENTS (UNBIASEDNESS, SECOND MOMENT, INEXACT SILVER, BATCHING)

Lemma 4.1 (Second moment: uniform sphere, symmetric two-point)**.** *Let $f \in C_L^{1,1}$, $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and $\widehat{g}(x;\mu,u) = \frac{d}{2\mu}\big(f(x+\mu u) - f(x-\mu u)\big)u$. Then*

$$\mathbb{E}\big\|\widehat{g}(x;\mu,u) - \nabla f_\mu(x)\|^2 \;\le\; C_{\mathrm{sig}}\, d\, \|\nabla f(x)\|^2 \;+\; C_{\mathrm{curv}}\, d^2 L^2 \mu^2,$$

with $(C_{\mathrm{sig}}, C_{\mathrm{curv}}) = (2, \frac{1}{2})$. Averaging any $B \ge 1$ unit directions gives a $1/B$ reduction. Using B orthonormal directions (Stiefel sampling) preserves the $1/B$ factor and improves constants in practice (Kozak et al., 2023; Feng & Wang, 2023).

These constants are tight up to lower-order terms for two-point ZO under L-smoothness; see the elementary proof in the appendix and the companion derivation we follow. The proof is bound can be found in the Appendix.

Filtration and conditional unbiasedness. Let $\mathcal{F}_t$ be a filtration. A vector process ζ_t is a martingale difference sequence if ζ_t is $\mathcal{F}_{t+1}$-measurable and $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$.

Let $\mathcal{F}_t$ be the σ-field generated by $\{x_0, \dots, x_t\}$ and all randomness up to the start of iteration t. At iteration t, sample fresh directions V_t independently of $\mathcal{F}_t$ (uniform on $\mathbb{S}^{d-1}$ or Haar on $\mathrm{St}(d, B_t)$), and evaluate f exactly (or with conditionally zero-mean noise) using the same batch for $\pm\mu$ queries. With the symmetric two-point estimator, we then have

$$\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = \nabla f_\mu(x_t) \qquad \text{and hence} \qquad \mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0,$$

by the ball-to-sphere identity and unbiasedness of the estimator for $\nabla f_\mu(x_t)$.

Lemma 4.2 (Martingale square identity)**.** *Let $\{\zeta_t\}_{t=0}^{N-1}$ be a square-integrable vector martingale difference sequence adapted to $\{\mathcal{F}_t\}$, i.e., $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$. Then, for any deterministic scalars $\{\alpha_t\}$,*

$$\mathbb{E}\Big\| \sum_{t=0}^{N-1} \alpha_t \, \zeta_t \Big\|^2 = \sum_{t=0}^{N-1} \alpha_t^2 \, \mathbb{E}\|\zeta_t\|^2.$$

Proof. Expand the square and use $\mathbb{E}\langle\zeta_s, \zeta_t\rangle = 0$ whenever $s \neq t$: for $s < t$, $\mathbb{E}\langle\zeta_s, \zeta_t\rangle = \mathbb{E}[\,\langle\zeta_s, \mathbb{E}[\zeta_t \mid \mathcal{F}_t]\rangle\,] = 0$ by the tower property. □

4.2 Main Results

Lemma 4.3 (Inexact Silver, expectation level)**.** *Let $h = f_\mu/L$ (so h is 1-smooth and convex) and suppose $x_{t+1} = x_t - \alpha_t(\nabla h(x_t) + \zeta_t)$ with $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$. For a Silver block $N = 2^k - 1$,*

$$\mathbb{E}\big[h(x_N) - h^\star\big] \;\le\; r_k\,\mathbb{E}\|x_0 - x^\star\|^2 \;+\; \sum_{t=0}^{N-1} \alpha_t^2\,\mathbb{E}\|\zeta_t\|^2.$$

Proof. See the Appendix for a detailed proof. □

We run the Silver identity on $h = f_\mu/L$, which is 1-smooth and convex. Our update is

$$x_{t+1} = x_t - \alpha_t\big(\nabla h(x_t) + \zeta_t\big), \qquad \zeta_t := \frac{1}{L}\Big(\widehat{g}_t - \nabla f_\mu(x_t)\Big),$$

so $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$ by Lemma A.2.

Variance-optimal batching under a query budget. We motivate the batching policy with the following variance-related observation.

Proposition 4.4 (Optimal allocation of directions under a query budget)**.** *Fix nonnegative weights $\{\alpha_t\}_{t=0}^{N-1}$ and a budget $Q > 0$ of function queries per block. With symmetric two-point queries, $Q = 2\sum_t B_t$. Then, for any $B_t > 0$,*

$$\sum_{t=0}^{N-1} \frac{\alpha_t^2}{B_t} \;\ge\; \frac{\big(\sum_{t=0}^{N-1} \alpha_t\big)^2}{\sum_{t=0}^{N-1} B_t} \;=\; \frac{2\big(\sum_t \alpha_t\big)^2}{Q},$$

with equality iff $B_t \propto \alpha_t$. In particular, the policy $B_t = \min\{d, \lceil c_B \alpha_t \rceil\}$ is (up to the cap and integrality) optimal for a given budget.

One-line proof. By Cauchy–Schwarz, $\Big(\sum \frac{\alpha_t^2}{B_t}\Big)\,(\sum B_t) \ge (\sum \alpha_t)^2$. Substitute $\sum B_t = Q/2$. □

Theorem 4.5 (One block, expectation)**.** *Assume $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth, and fix a Silver block of length $N = 2^k - 1$ with steps $\{\alpha_t\}_{t=0}^{N-1}$. Let f_μ be the uniform-ball smoothing, $h = f_\mu/L$, and define the symmetric two-point estimator averaged over B_t unit directions:*

$$\widehat{g}_t = \frac{d}{2\mu B_t} \sum_{i=1}^{B_t} \Big(f(x_t + \mu v_{t,i}) - f(x_t - \mu v_{t,i})\Big) v_{t,i}, \quad x_{t+1} = x_t - \frac{\alpha_t}{L}\,\widehat{g}_t.$$

Assume the batch $V_t = [v_{t,1}, \ldots, v_{t,B_t}] \in \mathrm{St}(d, B_t)$ is drawn independently of $\mathcal{F}_t$ (thin QR of a Gaussian suffices). Then

$$\mathbb{E}\big[f(x_N) - f^\star\big] \le r_k\,L\,\mathbb{E}\|x_0 - x^\star\|^2 + \sum_{t=0}^{N-1} \frac{\alpha_t^2}{B_t}\left(\frac{1}{2}d^2 L\,\mu^2 + \frac{2d}{L}\,\mathbb{E}\|\nabla f(x_t)\|^2\right) \;+\; \frac{L}{2}\mu^2\frac{d}{d+2}.$$

In particular, if the iterates remain in a ball of radius R around a minimizer (e.g., by projection), then $\|\nabla f(x_t)\| \le LR$ and

$$\mathbb{E}\big[f(x_N) - f^\star\big] \;\le\; r_k LR^2 \;+\; \Big(\tfrac{1}{2}d^2 L\mu^2 + 2d\,LR^2\Big) \sum_{t=0}^{N-1} \frac{\alpha_t^2}{B_t} \;+\; \tfrac{L}{2}\mu^2\frac{d}{d+2}.$$

Proof sketch. We apply the inexact-gradient Silver identity to $h = f_\mu/L$ with $\zeta_t = (\widehat{g}_t - \nabla f_\mu(x_t))/L$ (conditionally unbiased, so the identity has no linear noise term). Use the two-point second-moment bound plus averaging-by-B_t, then convert from h to f using the value-bias of f_μ. (see B.2, Eq. (3.1)) □

Budget-aligned specialization. Let $B_t = \min\{d, \lceil c_B\, \alpha_t \rceil\}$ (orthogonal-on-spikes), and write $\alpha_{\max} = \max_t \alpha_t$. Then

Proposition 4.6 (Variance aggregation under $B_t \propto \alpha_t$)**.**

$$\sum_{t=0}^{N-1} \frac{\alpha_t^2}{B_t} \;\le\; \frac{1}{c_B} \sum_{t=0}^{N-1} \alpha_t \;+\; \frac{\alpha_{\max}}{d} \sum_{t:\, \alpha_t > d/c_B} \alpha_t.$$

In particular, if $d \ge c_B \alpha_{\max}$ (cap inactive), then $\sum_t \alpha_t^2/B_t = (1/c_B)\sum_t \alpha_t = \Theta(\rho^k/c_B)$.

Corollary 4.7 (Per-block calibration of μ and c_B)**.** *If $\frac{\rho^k}{c_B} \cdot \left(\frac{1}{2} d^2 L \mu^2 + 2dLR^2\right) \le \varepsilon\, r_k\, LR^2$, then*

$$\mathbb{E}[f(x_N) - f^\star] \;\le\; (1+\varepsilon)\, r_k\, LR^2 \;+\; \frac{L}{2}\mu^2 \frac{d}{d+2}.$$

A sufficient choice is $c_B \ge \frac{2d\,\rho^k}{\varepsilon r_k}$ and $\mu^2 \le \frac{2\varepsilon r_k}{d^2}\frac{R^2}{\rho^k}$.

Theorem 4.8 (Multi-block restarts)**.** *Run blocks $j = 1, \dots, J$ with lengths $N_j = 2^{k_j} - 1$ and radii μ_j (each fixed within the block), using $B_t = \min\{d, \lceil c_B \alpha_t \rceil\}$. If $\|x_t - x^\star\| \le R$ across the run, then*

$$\mathbb{E}[f(x_{T_J}) - f^\star] \;\le\; LR^2 \sum_{j=1}^{J} r_{k_j} \;+\; \frac{1}{c_B} \sum_{j=1}^{J} \left(\tfrac{1}{2} d^2 L \mu_j^2 + 2d\, LR^2\right) \rho^{k_j} \;+\; \frac{L}{2}\mu_J^2 \frac{d}{d+2}.$$

Lemma 4.9 (High-probability inexact Silver via Freedman)**.** *Let $h = f_\mu/L$ and suppose $x_{t+1} = x_t - \alpha_t(\nabla h(x_t) + \zeta_t)$ with $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$ and $\|\zeta_t\| \le G$ almost surely. Define the predictable quadratic variation*

$$V \;:=\; \sum_{t=0}^{N-1} \alpha_t^2\, \mathbb{E}\big[\zeta_t \zeta_t^\top \mid \mathcal{F}_t\big], \qquad \bar\alpha := \max_{0 \le t \le N-1} \alpha_t.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$h(x_N) - h^\star \;\le\; r_k\, \|x_0 - x^\star\|^2 \;+\; \left(2\sqrt{2\,\lambda_{\max}(V)\, \log\frac{18^d}{\delta}} \;+\; \frac{2}{3}\,\bar\alpha\, G\, \log\frac{18^d}{\delta}\right)^2.$$

In particular, using $(a+b)^2 \le 2a^2 + 2b^2$,

$$h(x_N) - h^\star \;\le\; r_k\, \|x_0 - x^\star\|^2 \;+\; 16\,\lambda_{\max}(V)\, \log\frac{18^d}{\delta} \;+\; \frac{8}{9}\,\bar\alpha^2 G^2\, \log^2\frac{18^d}{\delta}.$$

Proof sketch. For any unit $s \in \mathbb{S}^{d-1}$, apply scalar Freedman to $M_s := \sum_{t=0}^{N-1} \alpha_t \langle \zeta_t, s\rangle$, which has variance proxy $\sigma_s^2 = \sum_t \alpha_t^2 \mathbb{E}[\langle \zeta_t, s\rangle^2 \mid \mathcal{F}_t] = s^\top V s \le \lambda_{\max}(V)$ and bounded increments $|\alpha_t \langle \zeta_t, s\rangle| \le \bar\alpha G$. Freedman gives $|M_s| \le \sqrt{2\sigma_s^2 u} + \frac{\bar\alpha G}{3} u$ with probability $\ge 1 - 2e^{-u}$ Freedman (1975). Take a $1/4$-net $\mathcal{N} \subset \mathbb{S}^{d-1}$ with $|\mathcal{N}| \le 9^d$ and union bound; since $\|z\| \le 2 \max_{s \in \mathcal{N}} \langle z, s\rangle$ (standard net argument; see e.g. Vershynin's notes on sphere nets), we get $\big\|\sum_t \alpha_t \zeta_t\big\| \le 2\sqrt{2\lambda_{\max}(V)\, u} + \frac{2}{3}\bar\alpha G\, u$ for $u = \log(2 \cdot 9^d/\delta)$. Square this and insert into the inexact-Silver identity, which contributes $\big\|\sum_t \alpha_t \zeta_t\big\|^2$ (no extra L's, since ζ_t already contains the $1/L$). □

Remark 4.10 (Matrix-Freedman alternative). One can avoid the sphere net (and the factor $\log(18^d/\delta)$) by applying a matrix Freedman inequality to a self-adjoint dilation of $\sum_t \alpha_t \zeta_t$, giveing a tail with $\log(d/\delta)$. See Tropp (2011). We keep Lemma 4.9 for its elementary proof via scalar Freedman Freedman (1975) and standard net estimates Vershynin (2018).

Remark 4.11 (Empirical status of Silver in first-order GD). As far as we are aware, the original Silver papers and their support material emphasize theoretical certificates, and do not provide systematic first-order empirical benchmarks. Discussions on empirical observations and generalizations (e.g., proximal/projected GD) are included, but a standardized FO benchmark suite on Silver vs. standard schedules has not yet emerged. See Altschuler & Parrilo (2023a;b; 2024); Parrilo (2024); Altschuler & Parrilo (2023c); Bok & Altschuler (2024).

4.3 ALGORITHM

Algorithm 1 ZO-SILVER: block-constant smoothing + orthogonal-on-spikes batching

1: **Input:** block length $N = 2^k - 1$, radius $\mu > 0$, Silver steps $\{\alpha_t\}$, cap d, batching constant $c_B > 0$
2: **for** $t = 0, \dots, N-1$ **do**
3: Stepsize $\eta_t = \alpha_t / L$; Batch $B_t = \min\{d, \lceil c_B\, \alpha_t \rceil\}$
4: Sample $V_t = [v_{t,1}, \dots, v_{t,B_t}] \in \mathrm{St}(d, B_t)$ (orthonormal columns; e.g., thin QR of a Gaussian matrix)
5: $\widehat{g}_t = \dfrac{d}{2\mu B_t} \displaystyle\sum_{i=1}^{B_t} \big(f(x_t + \mu v_{t,i}) - f(x_t - \mu v_{t,i})\big)\, v_{t,i}$
6: $x_{t+1} = x_t - \eta_t\, \widehat{g}_t$

Sampling orthonormal directions. A simple implementation samples $G \in \mathbb{R}^{d \times B_t}$ with i.i.d. $\mathcal{N}(0,1)$ entries and sets V_t to the Q factor of the thin QR decomposition $G = V_t R$, giving $V_t \in \mathrm{St}(d, B_t)$ with Haar-distributed columns.

Corollary 4.12 (Per-block calibration of μ and batching)**.** *Fix a block of length $N = 2^k - 1$ with $B_t = \min\{d, \lceil c_B \alpha_t \rceil\}$. Assume $\|\nabla f(x_t)\| \le L R_t$ along the block (e.g., by projection or local boundedness). If we choose μ and c_B to satisfy $\frac{\rho^k}{c_B} \cdot \left(\frac{1}{2} d^2 L \mu^2 + 2 d L R_t^2\right) \le \varepsilon \cdot r_k L R_t^2$, then the block guarantee becomes*

$$\mathbb{E}[f(x_N) - f^\star] \;\le\; (1+\varepsilon)\, r_k L R_t^2 \;+\; \tfrac{L}{2}\mu^2 \frac{d}{d+2}.$$

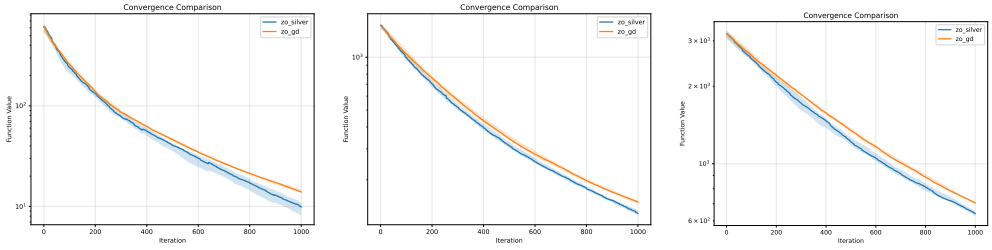
Equivalently, one sufficient choice is $c_B \ge \frac{2d\,\rho^k}{\varepsilon\, r_k}$ *and* $\mu^2 \le \frac{2\varepsilon\, r_k}{d^2} \cdot \frac{R_t^2}{\rho^k}$.

Remark 4.13 (Short blocks + geometric μ). Variance aggregation scales like ρ^k / c_B. Using *short* blocks (e.g., fix a small $k_\circ$) and *geometrically decaying* radii across blocks (e.g., $\mu \leftarrow \mu / 1 + \sqrt{2}$) keeps the variance tax controlled while the Silver term $r_{k_\circ}$ still provides a consistent per-block drop. This matches the ZO folklore: you must either (i) grow the batch, or (ii) reduce smoothing, or (iii) keep blocks short and restart. See also the two-point rates and floors in (Duchi et al., 2015; Shamir, 2017).

5 EXPERIMENTS

5.1 NUMERICAL EXPERIMENTS

Goals: (1) Verify that running Silver on the smoothed objective with *block-constant* μ preserves the deterministic decay down to the ZO floor; (2) test our *orthogonal-on-spikes* batching against flat/i.i.d. batching. **Tasks and oracles:** *Synthetic smooth convex*: ridge-regularized quadratic $f(x) = \frac{1}{2} x^\top H x + b^\top x$ with $\kappa(H)$ adjustable; *logistic regression* with ℓ_2 regularization. We evaluate both *noiseless* function values and *noisy* values $f(x) + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$. **Baselines:** (i) Two-point ZO with Gaussian smoothing and tuned constant stepsize (Nesterov & Spokoiny, 2017); **Protocols:** We run horizons aligning with Silver blocks ($T = 2^k - 1$). Within a block, μ is constant; across blocks we use $\mu_{j+1} = \mu_j / 1 + \sqrt{2}$ (or report a sweep). Batching follows $B_t = \min\{d, \lceil c_B \alpha_t \rceil\}$ with a fixed c_B; comparison baselines use (a) $B_t \equiv B$ and (b) $B_t \equiv 1$. **Budgets and metrics:** We report #queries and $f(x_t) - f^\star$ against queries and wall-clock. Each ZO step uses $2B_t$ function calls. We use medians over 10 seeds with 95% bootstrap CIs.



(a) $d = 200$ (b) $d = 500$ (c) $d = 1000$

Figure 1: Function value vs. iterations for ridge-regularized quadratic problem.

5.2 Experiments on Zeroth-Order Fine-Tuning LLMs

MeZO (Malladi et al., 2023) showed that large language models can be fine-tuned using *only forward passes* via two-point zeroth-order steps, achieving competitive accuracy with substantially lower memory than backprop. In the subsection, We instantiate our ZO-Silver scheme in the same 16-shot setting as in MeZO and conduct experiments on RoBERTa-large 350M (Liu et al., 2019).

For silver learning rate scheduler, we apply a simple clipping strategy $\alpha_t = \min\{\alpha_t, \alpha_{\max}\}$, where $\alpha_{\max}$ is the largest learning rate multiplier and searched in $\{16.0, 32.0, 64.0, 128.0\}$. For MeZO, we adopt a constant learning rate η_0, which is the same as the one used in silver learning rate scheduler and a large constant learning rate by taking the average over all the silver stepsizes. Specifically, after clipping, the limit of mean value of silver sizes is $\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \min\{\alpha_t, \alpha_{\max}\} = (\rho/2)^{J+1} + 2^{-(J+1)(\alpha_{\max}-1)}$, where $J = \lfloor 1 + \log_\rho(\alpha_{\max} - 1) \rfloor$. We plot the average sivler stepsizes across the training steps in Figure.2 with base learning rate $1e-7$ and largest learning rate multiplier 128.

As shown in Figure.3, MeZO-Silver achieves more stable training and lower validation loss under the same query budget, demonstrating the benefits of structured stepsize scheduling in practical LLM fine-tuning.

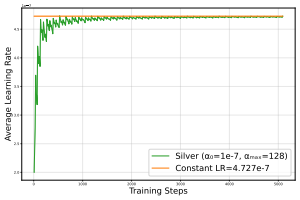
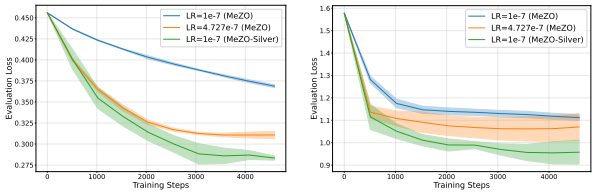


Figure 2: Clipped Silver schedule vs. constant LR baseline. We plot the running average learning rate induced by the Silver stepsizes with base learning rate $1e-7$ and clipping value 128.



(a) SST-2 (b) RTE

Figure 3: Evaluation loss vs. training steps for RoBERTa-large fine-tuning on (a) SST-2 and (b) RTE. We compare standard MeZO with constant learning rates to MeZO-Silver using a clipped Silver schedule.

6 Conclusion

The theoretical analysis and subsequent experimental work show that, under the studied assumptions for smooth objectives, composing two-point zeroth-order estimation with the Silver stepsize schedule yields a clean inexact-gradient certificate whose stochastic cost collapses to a single quadratic variance term. More broadly, our results provide the first step toward bringing stepsize hedging into the ZO regime: the Silver schedule—originally proved to accelerate plain gradient descent for smooth convex optimization—translates verbatim to the smoothed problem with conditionally unbiased ZO gradients, preserving its deterministic multi-step progress up to standard ZO floors.

REFERENCES

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging i: Multi-step descent and the silver stepsize schedule. *arXiv preprint*, 2023a. URL https://arxiv.org/abs/2309.07879.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization. *arXiv preprint*, 2023b. URL https://arxiv.org/abs/2309.16530.

Jason M. Altschuler and Pablo A. Parrilo. Support material for Acceleration by Stepsize Hedging. https://jasonaltschuler.github.io/AccelerationByStepsizeHedging/, 2023c.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization. *Mathematical Programming*, 199:1179–1219, 2024. doi: 10.1007/s10107-024-02164-2. URL https://link.springer.com/article/10.1007/s10107-024-02164-2. Final journal version.

Jinho Bok and Jason M. Altschuler. Accelerating proximal gradient descent via silver stepsizes. https://optimization-online.org/wp-content/uploads/2024/12/proxgd_stepsize_accel.pdf, 2024. arXiv:2412.05497.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks. In *Proceedings of the 10th Workshop on Artificial Intelligence and Security*, 2017.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.14314.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL https://proceedings.neurips.cc/paper/2018/file/ba9a56ce0a9bfa26e8ed9e10b2cc8f46-Paper.pdf. ZO variants discussed in follow-ups.

Yasong Feng and Tianyu Wang. Stochastic zeroth order gradient and hessian estimators: Variance reduction and refined bias bounds. *Information and Inference: A Journal of the IMA*, 12(3):1514–1545, 2023. doi: 10.1093/imaiai/iaad014. URL https://academic.oup.com/imaiai/article/12/3/1514/7150743.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th ACM–SIAM Symposium on Discrete Algorithms (SODA)*, pp. 385–394. SIAM, 2005. URL https://arxiv.org/abs/cs/0408007.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Kevin G. Jamieson, Robert Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems 25 (NeurIPS)*, pp. 2672–2680, 2012. URL https://proceedings.neurips.cc/paper/2012/file/e6d8545daa42d5ced125a4bf747b3688-Paper.pdf.

Kaiyi Ji, Zhe Wang, Yingbin Zhou, Yingbin Liang, and Ji Liu Wang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. URL https://proceedings.mlr.press/v97/ji19a/ji19a.pdf.

David Kozak, Cesare Molinari, Lorenzo Rosasco, Luis Tenorio, and Silvia Villa. Zeroth-order optimization with orthogonal random directions. *Mathematical Programming*, 2023. URL https://link.springer.com/content/pdf/10.1007/s10107-022-01866-9.pdf.

Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. doi: 10.1017/S0962492919000060. URL https://www.cambridge.org/core/journals/acta-numerica/article/abs/derivativefree-optimization-methods/84479E2B03A9BFFE0F9CD46CF9FCD289.

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.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.17333. Oral.

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

Pablo A. Parrilo. Accelerating gradient descent by stepsize hedging. https://aaforml.com/slides/SilverStepsizes.pdf, 2024. Applied Algorithms for ML, Paris, June 2024.

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

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, volume 30 of *Proceedings of Machine Learning Research*, pp. 3–24. PMLR, 2013. URL https://proceedings.mlr.press/v30/Shamir13.pdf.

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. URL https://jmlr.org/papers/volume18/16-632/16-632.pdf.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992. doi: 10.1109/9.119632.

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

A Uniform-ball bias bounds and the sphere gradient identity

Let $v \sim \mathrm{Unif}(\mathbb{B}^d)$ and $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. For $f_\mu(x) = \mathbb{E}_v f(x + \mu v)$, the Descent Lemma gives

$$-\tfrac{L}{2}\mu^2 \mathbb{E}\|v\|^2 \le f_\mu(x) - f(x) \le \tfrac{L}{2}\mu^2 \mathbb{E}\|v\|^2, \qquad \mathbb{E}\|v\|^2 = \frac{d}{d+2}.$$

Moreover $\|\nabla f_\mu(x) - \nabla f(x)\| \le L\mu\frac{d}{2}$. To prove (1), apply the divergence theorem to $\int_{\mathbb{B}^d} \nabla f(x + \mu z)\, dz$.

Lemma A.1 (Ball-to-sphere gradient identity, with constants)**.** *Let $f_\mu(x) = \mathbb{E}_{v\sim\mathrm{Unif}(\mathbb{B}^d)} f(x + \mu v)$ and $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Then*

$$\nabla f_\mu(x) = \frac{d}{\mu}\, \mathbb{E}_u\big[f(x + \mu u)\, u\big].$$

Lemma A.2 (Unbiasedness for ∇f_μ)**.** *With $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\widehat{g}(x; \mu, u) = \frac{d}{2\mu}(f(x + \mu u) - f(x - \mu u))u$, we have $\mathbb{E}_u[\widehat{g}(x; \mu, u)] = \nabla f_\mu(x)$.*

Lemma A.3 (Bias of f_μ and ∇f_μ)**.** *Assume $f \in C_L^{1,1}$.*

(a) (Ball value bias) For $v \sim \mathrm{Unif}(\mathbb{B}^d)$,

$$|f_\mu(x) - f(x)| \;\le\; \tfrac{L}{2}\,\mu^2\,\mathbb{E}\|v\|^2 \;=\; \tfrac{L}{2}\,\mu^2\,\frac{d}{d+2}.$$

(b) (Ball gradient bias) For $v \sim \mathrm{Unif}(\mathbb{B}^d)$,

$$\|\nabla f_\mu(x) - \nabla f(x)\| \;\le\; \tfrac{L}{2}\,d\,\mu.$$

Proof of (b). By (1), $\nabla f_\mu(x) - \nabla f(x) = \frac{d}{\mu}\,\mathbb{E}_u\big[(f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u\rangle)u\big]$. By the Descent Lemma along the line $x + \tau\mu u$ and $\|u\| = 1$, $|f(x + \mu u) - f(x) - \langle \nabla f(x), \mu u\rangle| \le \frac{L}{2}\mu^2$. Taking norms and expectations gives $\|\nabla f_\mu(x) - \nabla f(x)\| \le \frac{d}{\mu} \cdot \frac{L}{2}\mu^2 \mathbb{E}\|u\| = \frac{L}{2} d\mu$, since $\|u\| = 1$ a.s. □

B Spherical two-point second moment

Let $A_{d-1} = \mathrm{surf}(\mathbb{S}^{d-1})$ and $V_d = \mathrm{vol}(\mathbb{B}^d)$; recall $A_{d-1} = dV_d$. We use the standard sphere isotropy $\mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{d-1})}[u] = 0$, $\mathbb{E}[uu^\top] = \frac{1}{d}I$, and $\|u\| = 1$ a.s.

Lemma B.1 (Inexact Silver, expectation level)**.** *Let $h = f_\mu/L$ (so h is 1-smooth and convex) and suppose $x_{t+1} = x_t - \alpha_t(\nabla h(x_t) + \zeta_t)$ with $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$. For a Silver block $N = 2^k - 1$,*

$$\mathbb{E}\big[h(x_N) - h^\star\big] \;\le\; r_k\,\mathbb{E}\|x_0 - x^\star\|^2 \;+\; \sum_{t=0}^{N-1} \alpha_t^2\,\mathbb{E}\|\zeta_t\|^2.$$

Proof. Let $\{\lambda_{ij}\}$ be the Silver multipliers such that for any 1-smooth convex ϕ,

$$\sum_{i\neq j} \lambda_{ij} Q_{ij}[\phi] = \|x_0 - x^\star\|^2 - \|x_N - c_k\nabla\phi(x_N) - x^\star\|^2 + \frac{\phi(x^\star) - \phi(x_N)}{r_k}. \tag{$\star$}$$

Apply $(\star)$ with $\phi = h$. In the Silver derivation, the only places where the update rule enters are: (i) linear telescopings $x_a - x_b = -\sum_{s=b}^{a-1} \alpha_s \nabla h(x_s)$ and (ii) the terminal square $\|x_N - c_k\nabla h(x_N) - x^\star\|^2$. With inexact updates we have $x_a - x_b = -\sum \alpha_s \nabla h(x_s) - \sum \alpha_s\zeta_s$. Every such *linear* ζ-term appears inside an inner product with an $\mathcal{F}_s$-measurable vector, hence its expectation is 0 by $\mathbb{E}[\zeta_s \mid \mathcal{F}_s] = 0$. For the terminal square,

$$x_N - c_k\nabla h(x_N) - x^\star = A - \sum_{s=0}^{N-1} \alpha_s\zeta_s, \qquad A := x_0 - x^\star - \sum_{s=0}^{N-1} \alpha_s \nabla h(x_s) - c_k\nabla h(x_N).$$

Therefore

$$\mathbb{E}\|x_N - c_k\nabla h(x_N) - x^\star\|^2 = \mathbb{E}\|A\|^2 + \mathbb{E}\Big\|\sum_{s=0}^{N-1}\alpha_s\zeta_s\Big\|^2$$

(the cross term vanishes in expectation as above). By Lemma 4.2, this last term equals $\sum_{t=0}^{N-1}\alpha_t^2\,\mathbb{E}\|\zeta_t\|^2$. Taking expectations in $(\star)$, dropping the nonnegative left-hand side $\sum\lambda_{ij}Q_{ij}[h]$, and rearranging gives the claim.

Lemma B.2 (Second moment: uniform sphere, symmetric two-point)**.** *Let $f \in C_L^{1,1}$, $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, and $\widehat{g}(x;\mu,u) = \frac{d}{2\mu}\big(f(x+\mu u) - f(x-\mu u)\big)u$. Then*

$$\mathbb{E}\big\|\widehat{g}(x;\mu,u) - \nabla f_\mu(x)\|^2 \;\le\; C_{\mathrm{sig}}\,d\,\|\nabla f(x)\|^2 \;+\; C_{\mathrm{curv}}\,d^2L^2\mu^2,$$

with $(C_{\mathrm{sig}}, C_{\mathrm{curv}}) = (2, \frac{1}{2})$. Averaging any $B \ge 1$ unit directions gives a $1/B$ reduction. Using B orthonormal directions (Stiefel sampling) preserves the $1/B$ factor and improves constants in practice (Kozak et al., 2023; Feng & Wang, 2023).

These constants are tight up to lower-order terms for two-point ZO under L-smoothness; see the elementary proof in the appendix and the companion derivation we follow.

Proof of Lemma B.2 (Proof sketch). Apply the Descent Lemma along $\pm\mu u$ to get $\big|f(x+\mu u) - f(x-\mu u) - 2\mu\langle\nabla f(x), u\rangle\big| \le L\mu^2$, then expand $\|\widehat{g}\|^2$, use $(a+b)^2 \le 2a^2 + 2b^2$ and isotropy $\mathbb{E}[uu^\top] = I/d$. The full derivation is in Appendix B.

Finally, since $\mathbb{E}_u\,\widehat{g}(x;\mu,u) = \nabla f_\mu(x)$ (Lemma A.1), we have

$$\mathbb{E}\|\widehat{g} - \nabla f_\mu(x)\|^2 = \mathbb{E}\|\widehat{g}\|^2 - \|\nabla f_\mu(x)\|^2 \;\le\; \mathbb{E}\|\widehat{g}\|^2 \overset{(2)}{\le}\; 2d\,\|\nabla f(x)\|^2 + \tfrac{1}{2}d^2L^2\mu^2.$$

For a batch $\{v_i\}_{i=1}^B$ of unit vectors (not necessarily independent),

$$\left\|\frac{1}{B}\sum_{i=1}^{B}(Y_i - \mathbb{E}Y)\right\|^2 \;\le\; \frac{1}{B}\sum_{i=1}^{B}\|Y_i - \mathbb{E}Y\|^2$$

by convexity of $\|\cdot\|^2$; taking expectations gives the stated $1/B$ reduction. □

□

Proof of Lemma B.2. Fix $x \in \mathbb{R}^d$ and $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Since $f \in C_L^{1,1}$, write the Descent Lemma at x in the two directions $\pm\mu u$:

$$f(x \pm \mu u) \;=\; f(x) \;\pm\; \mu\langle\nabla f(x), u\rangle \;+\; r_\pm(x,u), \qquad |r_\pm(x,u)| \le \frac{L}{2}\mu^2.$$

Subtract to get the symmetric difference

$$\Delta(x,u) \;:=\; f(x+\mu u) - f(x-\mu u) \;=\; 2\mu\langle\nabla f(x), u\rangle \;+\; (r_+ - r_-),$$

with $|r_+ - r_-| \le L\mu^2$. Hence

$$\widehat{g}(x;\mu,u) \;=\; d\,\langle\nabla f(x), u\rangle\,u \;+\; \frac{d}{2\mu}\,(r_+ - r_-)\,u.$$

Using $\mathbb{E}[uu^\top] = I/d$ for $u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and $\|u\| = 1$,

$$\mathbb{E}\big\|\widehat{g}(x;\mu,u)\big\|^2 \;\le\; 2\,d\,\|\nabla f(x)\|^2 \;+\; \frac{d^2}{2\mu^2}\,\mathbb{E}\big[(r_+ - r_-)^2\big] \;\le\; 2\,d\,\|\nabla f(x)\|^2 \;+\; \frac{1}{2}\,d^2L^2\mu^2.$$

For the centered version, note that $\mathbb{E}\,\widehat{g}(x;\mu,u) = \nabla f_\mu(x)$ by the ball-to-sphere identity, so $\mathbb{E}\|\widehat{g} - \nabla f_\mu(x)\|^2 = \mathbb{E}\|\widehat{g}\|^2 - \|\nabla f_\mu(x)\|^2 \le \mathbb{E}\|\widehat{g}\|^2$, which yields the same bound. Finally, for an average over B unit directions $\widehat{g}_B = \frac{1}{B}\sum_{i=1}^B\widehat{g}(x;\mu,u_i)$ (independent or not), convexity of $\|\cdot\|^2$ gives $\mathbb{E}\|\widehat{g}_B - \mathbb{E}\widehat{g}_B\|^2 \le \frac{1}{B}\sum_{i=1}^B\mathbb{E}\|\widehat{g}(x;\mu,u_i) - \mathbb{E}\widehat{g}\|^2$, so both second-moment bounds divide by B. □