ADAPTIVE STEPSIZES MEET MOMENTUM: AN IMPROVED ZEROTH-ORDER METHOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Training deep networks relies heavily on adaptive step sizes and momentum mechanisms. Yet, in gradient-free regimes, their theoretical foundations are still underdeveloped. Since many problems permit only zeroth-order oracle access, classical gradient-based analyses cannot be applied. To address this, we propose ZO-AdaNesterov, a stochastic zeroth-order Nesterov momentum method equipped with adaptive step sizes, and develop a rigorous convergence analysis. We derive oracle complexity bounds for both convex and nonconvex objectives, which are strictly sharper than those of existing zeroth-order adaptive momentum methods. Besides, we provide the *first* last-iterate saddle-avoidance result for zeroth-order adaptive methods: with probability one, ZO-AdaNesterov does not converge to strict saddle points. Numerical experiments further confirm the efficiency of our method.

1 INTRODUCTION

Adaptive stepsizes (Duchi et al. (2011)) and momentum mechanisms (Nesterov (2013)) are two central ingredients behind the empirical success of modern stochastic optimization. Adaptive methods, such as AdaGrad and Adam, reduce the need for careful manual tuning by scaling the stepsize according to the observed history of stochastic gradients, while momentum mechanisms often practically accelerate training models and improve stability in large-scale problems. In the first-order setting, these ideas have been extensively studied and have become standard components of practical optimizers (e.g., Kingma & Ba (2014); Loshchilov & Hutter (2019)).

In many important applications, however, explicit gradient information is unavailable. This occurs in black-box learning (Zhang et al. (2022a)), and several large-model tuning scenarios (Malladi et al. (2023); Zhang et al. (2024)) where only function-value feedback can be queried. In such settings, one must rely on zeroth-order methods, which estimate gradient information from function evaluations. Compared with first-order methods, zeroth-order optimization is substantially more challenging: gradient estimation introduces additional dimension dependence, stochastic approximation error, and bias due to smoothing. These difficulties become even more pronounced when zeroth-order gradient estimators are combined with adaptive stepsizes and momentum.

Recent works have begun to develop adaptive zeroth-order momentum methods, such as ZO-AdaMM (Chen et al. (2019)) and R-AdaZO (Shu et al. (2025)). These algorithms demonstrate that adaptive stepsizes and momentum can be useful in gradient-free optimization. Nevertheless, their theoretical guarantees still have several important limitations. First, their oracle complexities exhibit unfavorable dependence on the ambient dimension d. For example, existing adaptive zeroth-order methods with momentum (Chen et al. (2019); Shu et al. (2025)) typically require unfavorable dimension dependence such as d^2 or d^3 in their complexity bounds. Second, their analyses often rely on restrictive bounded-oracle assumptions, which require the stochastic oracles to be uniformly bounded. Such assumptions can be difficult to justify in stochastic black-box optimization and do not capture many practical settings with unbounded stochastic oracles. Third, existing bounds do not always make the dependence on the noise level explicit, making it unclear how these methods adapt to different stochastic environments.

Beyond first-order convergence, another important theoretical issue is saddle-point avoidance. In nonconvex optimization, convergence to first-order stationary points does not rule out convergence to strict saddle points. For first-order stochastic methods, a growing literature (Mertikopoulos et al. (2020); Liu & Yuan (2024)) has shown that SGD and its momentum variants can avoid strict saddles

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

almost surely under suitable conditions. Such results provide last-iterate second-order stationarity guarantees and offer a deeper description of the limiting behavior of the algorithm. In contrast, for zeroth-order methods with adaptive stepsizes and momentum, last-iterate saddle-avoidance theory is largely missing. Existing zeroth-order adaptive momentum algorithms mainly focus on first-order stationarity and do not establish that their last iterates avoid saddle points almost surely.

One may ask whether the desired guarantees could be obtained by directly refining the analyses of existing adaptive zeroth-order momentum methods such as ZO-AdaMM and R-AdaZO. These methods are important predecessors, but their current theoretical guarantees reveal several structural limitations: their oracle complexities have unfavorable dimension dependence, their analyses rely on bounded-oracle assumptions, and they do not provide last-iterate saddle-avoidance guarantees. Moreover, their Adam-type moving-average normalization couples momentum, stochastic zeroth-order estimation error, and adaptive denominators in a way that makes both unbounded-oracle analysis and saddle-avoidance arguments difficult.

These limitations suggest that obtaining sharper guarantees may require an algorithmic structure that is not only adaptive and momentum-based, but also transparent enough for a refined zeroth-order analysis. At the same time, such a method should not be designed merely for theoretical convenience: it should retain the practical ingredients that make adaptive stepsizes and momentum mechanisms effective. These considerations motivate the following **question**:

Can we design a zeroth-order method with adaptive stepsizes and momentum that improves the oracle complexity of existing adaptive zeroth-order momentum methods, works under unbounded stochastic oracles with explicit noise dependence, enjoys last-iterate saddle-avoidance guarantees, and meanwhile retain the practical performance?

1.1 Summary of Our Contributions

In this paper, we answer this question affirmatively by proposing **ZO-AdaNesterov**, a zeroth-order stochastic Nesterov-type momentum method with adaptive stepsizes. ZO-AdaNesterov combines an AdaGrad-style cumulative normalization with a Nesterov-tyoe momentum update and a carefully controlled zeroth-order gradient estimator. This design preserves the adaptive and momentum-based behavior that is useful in practice, while avoiding some of the technical complications of Adam-type moving-average normalization. As a result, it admits sharper oracle complexity bounds, allows unbounded stochastic oracles with explicit noise dependence, and supports a last-iterate saddle-avoidance analysis. Our main contributions are summarized as follows.

1. **Improved oracle complexities.** We prove expected oracle complexity bounds of $\mathcal{O}(\frac{d}{\delta} + \frac{d\sigma^2}{\delta^2})$ for finding a δ-approximation in fractional moment of convex objectives and $\mathcal{O}(\frac{d}{\delta^2} + \frac{d\sigma^2}{\delta^4})$ for finding δ-stationary point of nonconvex objectives. As summarized in Table 1, these bounds significantly improve the dimension dependence of existing adaptive zeroth-order momentum methods such as ZO-AdaMM and R-AdaZO. Moreover, our analysis allows unbounded oracles and provides explicit dependence on the noise level, whereas the analyses of ZO-AdaMM and R-AdaZO rely on bounded-oracle assumptions and do not provide explicit noise-level dependence.
2. **Last-iterate saddle avoidance.** We prove that the last iterate of ZO-AdaNesterov avoids strict saddle points almost surely. Consequently, the asymptotical behavior of the algorithm satisfies a probability-one guarantee. To the best of our knowledge, this is the first last-iterate saddle-avoidance result for zeroth-order adaptive methods.
3. **Empirical validation.** We demonstrate the practical efficiency of ZO-AdaNesterov through extensive numerical experiments on real-world tasks, showing that the proposed method is competitive with, and often improves upon, existing zeroth-order adaptive momentum baselines.

1.2 Related work

Nesterov Momentum. Nesterov's momentum method (Nesterov (1983)) has long been a cornerstone of first-order optimization, theoretically achieving the optimal convergence rate of $\mathcal{O}(1/T^2)$ for smooth convex functions in deterministic settings. However, extending this momentum to the stochastic regime presents significant theoretical challenges. The intrinsic lower bound established

Algorithms	Complexity	UnbOra	NoiAda	SadAvod	Last Iter.
ZO-AdaMM Chen et al. (2019)	$\tilde{\mathcal{O}}(\frac{d^3}{\delta^2})$(C), $\tilde{\mathcal{O}}(\frac{d^2}{\delta^4})$(NC)	✗	✗	✗	✗
R-AdaZO Shu et al. (2025)	$\tilde{\mathcal{O}}(\frac{d^2}{\delta^4})$(NC)	✗	✗	✗	✗
ZO-AdaNesterov **(This paper)**	$\mathcal{O}(\frac{d}{\delta}+\frac{d\sigma^2}{\delta^2})$(C), $\mathcal{O}(\frac{d}{\delta^2}+\frac{d\sigma^2}{\delta^4})$(NC)	✔	✔	✔	✔

Table 1: Comparison of stochastic zeroth-order methods with adaptive stepsizes and momentum. "**C**" and "**NC**" denote convex and nonconvex settings, respectively. "**UnbOra**" indicates whether the *complexity analysis* allows unbounded oracles. "**NoiAda**" indicates whether the complexity bound is adaptive to the noise level, in the sense that it exhibits explicit dependence on the noise variance σ^2. "**SadAvod**" denotes almost sure saddle avoidance. "**Last Iter.**" denotes last-iterate guarantee. We clarify that the discussion on unbounded oracles is restrictive to complexity analysis, the analysis of last iterate saddle avoidance requires additional assumptions, see Section 3.3 in detail.

by Agarwal et al. (2009) for stochastic first-order oracles shows that Nesterov's momentum cannot generally improve the worst-case complexity of stochastic optimization. Early analyses even revealed that naive applications of Nesterov's momentum method to stochastic objectives often fail to retain the accelerated rate due to the accumulation of gradient noise, which can lead to instability or divergence (Devolder et al. (2014); Jain et al. (2018)). In the more general stochastic setting, Ghadimi & Lan (2012) established the first rigorous convergence guarantees for stochastic convex optimization by carefully tuning stepsize policies. Besides, in the context of non-convex deep learning, Sutskever et al. (2013) reinterpreted Nesterov's momentum method as a momentum correction, sparking extensive empirical adoption. Despite these advances, theoretical understanding remains limited when momentum is coupled with adaptive stepsizes or restricted to zeroth-order oracle access. Current zeroth-order adaptive momentum methods often struggle to balance the aggressive momentum updates with the noise inherent in adaptive estimates, a gap this work aims to address.

Zeroth-Order Optimization. The roots of derivative-free optimization, commonly known as zeroth-order optimization, trace back to the seminal work of Nemirovskij & Yudin (1983). Recent years have witnessed a surge of interest in these techniques across both deterministic and stochastic settings. In the deterministic setting, research has branched based on sampling strategies and function classes. For instance, Lin et al. (2022) analyzed convergence for nonconvex-nonsmooth problems using uniform sampling, while Gaussian sampling has been used for establishing rates in convex settings (Ghadimi & Lan (2013); Nesterov & Spokoiny (2017)). Beyond first-order convergence, significant strides have been made toward identifying δ-second-order stationary points (Vlatakis-Gkaragkounis et al. (2019); Balasubramanian & Ghadimi (2022)), with Ren et al. (2023) recently refining the rate to $\mathcal{O}(d/\delta^{2.5})$. Furthermore, the scope has expanded to include almost sure saddle point avoidance (Vlatakis-Gkaragkounis et al. (2019)) and mini-max optimization (Xu et al. (2023)). A similar evolution is evident in the stochastic setting. Expected convergence rates for convex functions have been established using uniform (Duchi et al. (2015)) and Gaussian sampling (Ghadimi & Lan (2013); Nesterov & Spokoiny (2017)). For challenging nonconvex-nonsmooth scenarios, Lin et al. (2022) derived convergence rates via uniform sampling, though these results originally suffered from a heavy dependence on dimensionality. This dependence was later mitigated by Kornowski & Shamir (2023) through an elegant analysis, alongside other improvements by Huang et al. (2022); Yue et al. (2023). Additional specialized applications include optimization on Riemannian manifolds (Li et al. (2020)), constrained optimization with saddle avoidance (Balasubramanian & Ghadimi (2022)), exploration of orthogonal directions (Rando et al. (2024)), and methods tailored for the finite-sum setting (see Ji et al. (2019) and references therein).

Adaptive Stepsize. Auer et al. (2002) made significant contributions to the early exploration of adaptive stepsizes in the field of online convex optimization. The pioneering work of AdaGrad (Duchi et al. (2011); McMahan & Streeter (2010)) introduced the mechanism of accumulating the sum of squared gradients to normalize updates, theoretically achieving optimal regret bounds for

Algorithm 1 ZO-AdaNesterov

Input: $d \geq 1$, $\mu_1 > 0$, T, $w_1 = x_1$, $\eta_1 = a/b^{1/2+\epsilon}$, batch size m, and momentum factor θ
for $n = 1$ **to** T **do**
  Sample $\boldsymbol{U_n} = \{u_{i,n}\}_{i=1}^m$ from $\mathrm{Unif}(\mathbb{S}_1(0))$, and $\{\xi_{i,n}\}_{i=1}^m$ independently
  Compute $\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(x_n)$ in (1) and stepsize η_n in (2)
  $w_{n+1} = x_n - \eta_n \hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(x_n)$
  $x_{n+1} = w_{n+1} + \theta(w_{n+1} - w_n)$
  $\mu_{n+1} = \frac{\mu_1}{n+1}$
end for
Ouput: $\bar{w}_T = T^{-1}\sum_{n=1}^T w_n$ in the convex case, or w_R with independent $R \sim \mathrm{Unif}\{1, \ldots, T\}$ in the nonconvex case.

sparse data. However, in non-convex settings such as deep learning, the monotonic accumulation of squared gradients in AdaGrad causes the learning rate to vanish prematurely, halting training. To mitigate this rapid decay, algorithms such as RMSProp (Tieleman (2012)) replaced the global accumulation with an exponential moving average (EMA) of squared gradients. Building upon these ideas, Adam (Kingma & Ba (2014)) emerged as a seminal contribution, combining the adaptive scaling of RMSProp with the momentum-based mechanism of heavy-ball methods. By maintaining EMAs of both the first moment (momentum) and the second moment (uncentered variance) of the gradients, Adam demonstrated superior empirical performance and stability, rapidly becoming the de facto optimizer for training deep neural networks and large-scale language models. To refine the theoretical analysis of Adam, Reddi et al. (2018) proposed AMSGrad, which enforces monotonicity in the denominator of the update rule to guarantee theoretical convergence. Regarding the theoretical analysis of adaptive stepsizes, the researchers mainly focus on AdaGrad and Adam. Due to space limitations, we defer the discussion on the theoretical analysis of AdaGrad and Adam to Appendix A.

2 Setup and Algorithms

2.1 Setup

Throughout this paper, we consider the following generic stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) \ \text{ and } \ f(x) = \mathbb{E}_\xi[F(x,\xi)], \tag{P}$$

where $f(x) : \mathbb{R}^d \to \mathbb{R}$ is a smooth objective function that is possibly non-convex, and ξ is a random variable that encodes uncertainties from the environment. Let $\mathcal{F}_{n-1}$ contain all randomness generated before iteration n. We write $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{n-1}]$.

Throughout this paper, we denote $\tilde{x}^* \in \mathrm{argmin}_{x\in\mathbb{R}^d} f(x)$ and $f^* = \min_{x\in\mathbb{R}^d} f(x) > -\infty$. In this paper we allow the function f have set of saddle points of continuum, say the *strict saddle manifold* (Mertikopoulos et al. (2020)). Precisely, we denote $\mathcal{C} = \{x \in \mathbb{R}^d : \nabla f(x) = 0\}$ the critical set of $f(x)$. A strict saddle manifold $\mathcal{S}$ of $f(x)$ is a compact smooth connected component of $\mathcal{C}$ satisfying:

- Each $x^* \in \mathcal{S}$ is a strict saddle point of f, i.e., $\lambda_{\min}(\nabla^2 f(x^*)) < 0$, and $f(x)$ is C^3 on a neighborhood of $\mathcal{S}$;
- There exist $c_-, c_+ > 0$ such that, for all $x^* \in \mathcal{S}$, all negative eigenvalues of $\nabla^2 f(x^*)$ are bounded from above by $-c_-$, and any positive eigenvalues are bounded from below by c_+.

In this paper, we make the following assumptions on $F(x,\xi)$ regarding smoothness.

Assumption 1. $F(x,\xi)$ is M-smooth, i.e., there exists $M > 0$, for all $x, y \in \mathbb{R}^d$ and any ξ, it satisfies

$$\|\nabla F(x,\xi) - \nabla F(y,\xi)\| \leq M\|x - y\|.$$

Note that Assumption 1 imply that $f(x)$ is M-smooth. Assumption 1 is widely used in zeroth-order optimization (e.g., Huang et al. (2022); Balasubramanian & Ghadimi (2022)).

2.2 Algorithms

Recall that the objective function f is obtained by taking expectation over certain random variable, i.e., $f(x) = \mathbb{E}_\xi[F(x,\xi)]$. Leveraging the random sample at each step, we can construct the stochastic

zeroth-order estimate of gradient in the following manner,

$$g_n := \widehat{\nabla} f_{U_n}^{\mu_n}(x_n) = \frac{1}{m}\sum_{i=1}^{m} q_{\mu_n}(x_n; u_{i,n}, \xi_{i,n}),$$
$$q_\mu(x; u, \xi) := \frac{d}{\mu}\big(F(x+\mu u, \xi) - F(x, \xi)\big)u. \tag{1}$$

where $\boldsymbol{U}_n = \{u_{i,n}\}_{i=1}^m$ are m independently vectors that are all uniformly sampled in the unit sphere $\mathbb{S}_1(0)$ centered at origin, i.e., $u_{i,n} \sim \mathrm{Unif}(\mathbb{S}_1(0))$, and μ is the smoothing parameter. Also, m independent pairs $(u_{i,n}, \xi_{i,n})$ are drawn independently of $\mathcal{F}_{n-1}$. Define the μ-smoothing function of $f(x)$ in unit ball $\mathbb{B}_1(0)$ as $f_\mu(x) = \mathbb{E}_{s\sim\mathrm{Unif}(\mathbb{B}_1(0))}[f(x+\mu s)]$. To control the variance of gradient estimates and make the algorithm converge to a stationary point of the objective function f, we set the smoothing parameter sequence $\{\mu_n\}$ as a decreasing sequence, i.e., $\mu_n = \frac{\mu_1}{n}$. In the following, we use both $\hat{\nabla} f_{\boldsymbol{U}}^{\mu}(x)$ and $\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(x_n)$ to denote the zeroth-order estimate of gradient, depending on whether the iteration needs to be emphasized. Then, we derive ZO-AdaNesterov (cf. Algorithm 1):

$$w_{n+1} = x_n - \eta_n \hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(x_n), \quad x_{n+1} = w_{n+1} + \theta(w_{n+1} - w_n) \tag{ZO-AdaNesterov}$$

where $\theta \in (0,1)$ is the momentum parameter and the adaptive step-size is defined as

$$\eta_n = \frac{a}{\left(b + \sum_{i=1}^{n-1}\left\|\hat{\nabla} f_{\boldsymbol{U}_i}^{\mu_i}(x_i)\right\|^2\right)^{\frac{1}{2}+\epsilon}}, \tag{2}$$

where $a, b > 0$ and $0 \le \epsilon \le \frac{1}{2}$. The output sequence for Algorithm 1 is $\{w_n\}$ (See Appendix B).

Stepsizes Statement. It should be pointed out that the adaptive stepsizes η_n (2) in this paper do not absorb the current gradient oracle g_n. Some papers (e.g., Attia & Koren (2023)) also call stepsizes (2) as delayed adaptive stepsizes. While the standard practice in AdaGrad literature is to set $\epsilon = 0$, we follow the insight of Li & Orabona (2019), who demonstrated that a non-zero ϵ accelerates the step-size decay, which is essential for proving almost sure convergence to critical points. This choice is further supported by Choudhury et al. (2024), whose empirical results demonstrate that step-sizes with $\epsilon = \frac{1}{2}$ outperform the traditional $\epsilon = 0$ setting. This formulation contrasts with the traditional step-size found in Duchi et al. (2011), which absorbs the current gradient:

$$\eta_n = \frac{a}{\left(b + \sum_{i=1}^{n}\left\|\hat{\nabla} f_{\boldsymbol{U}_i}^{\mu_i}(x_i)\right\|^2\right)^{\frac{1}{2}}}. \tag{3}$$

Actually, the choice between (2) and (3) fundamentally alters the theoretical analysis. By applying standard descent inequalities (assuming M-smoothness) for AdaGrad with zeroth-order estimates, $y_{n+1} = y_n - \eta_n \hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)$, we have:

$$\begin{aligned}\mathbb{E}_n[f(y_{n+1})] &\le f(y_n) + \mathbb{E}_n\left[\langle \nabla f(y_n), y_{n+1} - y_n\rangle + \frac{M\|y_{n+1} - y_n\|^2}{2}\right] \\ &= f(y_n) + \underbrace{\mathbb{E}_n[\left\langle \nabla f(y_n), -\eta_n \hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)\right\rangle]}_{\text{First Order}} + \underbrace{\frac{M}{2}\mathbb{E}_n[\eta_n^2\left\|\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)\right\|^2]}_{\text{Second Order}}.\end{aligned}$$

In fact, there exists an analytical trade-off between the stepsizes (2) and (3):

- For stepsizes (3), Lemma 10 in Wang et al. (2023b) simply bounds the summation of "**Second Order Term**" $\sum_{n=1}^{T}\mathbb{E}_n\left[\eta_n^2\left\|\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)\right\|^2\right]$, yet the correlation between η_n and $\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)$ complicates the derivation of the "**First Order Term**".
- Conversely, stepsizes (2) simplify "**First Order Term**" by excluding $\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)$; however, this prevents directly using Lemma 10 in Wang et al. (2023b) to bound the summation of "**Second Order Term**" $\sum_{n=1}^{T}\mathbb{E}_n\left[\eta_n^2\left\|\hat{\nabla} f_{\boldsymbol{U}_n}^{\mu_n}(y_n)\right\|^2\right]$, thereby complicating the complexity analysis.

3 MAIN RESULTS

This section establishes the convergence properties of (ZO-AdaNesterov). Next, we begin by outlining the technical novelty underlying these proofs.

Technical Statement. We outline the main analytical challenges and our technical solutions below:

- **Coupled energy and predictable-compensator analysis.** The zeroth-order estimator interacts with the adaptive denominator through both its nonzero mean and its stochastic fluctuation. Mini-batching reduces the latter, but does not eliminate the mean-square contribution. We retain this contribution in Lemma D.13 and couple the delayed squared-update energy with weighted optimization progress through Lemmas E.1–E.2. To recover unweighted guarantees without uniformly bounded stochastic gradients, we prove the predictable-compensator comparison in Lemma D.16. Combining this comparison with Hölder's inequality yields a common estimate for convex and nonconvex objectives (Lemma E.4). Crucially, retaining the full term $b + 1 + 5M^2$ (M is the smoothness constant in Assumption 1) avoids requiring the initial denominator to dominate the noise budget. Two complementary energy estimates then yield explicit sufficient regions in $h = a/b^{1/2+\epsilon}$ for every $b > 0$. With the stated horizon-dependent batch size m, this establishes Theorems 3.1 and 3.3 and the linear-in-d complexity oracle bounds for convex and nonconvex objectives under their respective stationary criteria.
- **Energy-first almost-sure convergence without a small initial step.** The adaptive stepsizes need not satisfy $\sum_n \eta_n^2 < \infty$, so this condition cannot be imposed as an intermediate assumption. Instead, under samplewise Lipschitz continuity, we first prove the pathwise bound

$$\sum_{n=1}^{\infty} \eta_n^2 \left\| g_n \right\|^2 \leq \frac{a^2}{2\epsilon b^{2\epsilon}} + \frac{d^2 L^2 a^2}{b^{1+2\epsilon}} < \infty, \qquad 0 < \epsilon \leq \tfrac{1}{2}.$$

  At the actual Nesterov coordinates $z_n = w_n + \theta v_n/(1-\theta)$, where $v_n = w_n - w_{n-1}$, a geometric-convolution estimate controls the momentum by this same update energy. Inserting that bound into the descent inequality *before* any energy absorption gives weighted gradient summability without a small-step restriction. Divergence of the total steps and a gradient-variation estimate then establish Theorem 3.6. This separates asymptotic convergence from finite-horizon tuning: any fixed $a, b > 0$ and $0 < \theta < 1$ is admissible, and choosing a, b does not require knowledge of M or L.
- **Deterministic-scale repulsion for biased adaptive dynamics.** A future tail of random adaptive stepsizes cannot be treated as a deterministic quantity, while smoothing generally prevents the total estimation error from being a martingale difference. We handle both obstacles by separating $g_n - \nabla f(x_n) = e_n + \zeta_n$, with $\mathbb{E}_n[\zeta_n] = 0$ (see (48) in detail), and using bounded estimators and excitation of the actual fixed batch to establish eventual deterministic comparisons with constants c_η and C_η:

$$c_\eta n^{-p} \leq \eta_n \leq C_\eta n^{-p}, \qquad p = \tfrac{1}{2} + \epsilon \in [\tfrac{1}{2}, 1].$$

  Lemmas J.1–J.2 then reduce momentum and smoothing to $\mathcal{O}(\alpha_n^2)$ update errors, where $\alpha_n = \eta_n/(1-\theta)$. Building on classical center-stable manifold existence, we prove local distance-repulsion estimates and a scalar nonconvergence lemma on the deterministic scale n^{-p} (Lemmas J.3–J.4). Separate arguments cover $p > 1/2$ and the nonsquare-summable endpoint $p = 1/2$. Localization at every deterministic starting index excludes eventual convergence after arbitrary re-entries, rather than merely proving one exit. Under the stated compact C^3 saddle geometry and spectral gaps, this establishes Theorem 3.8 for every fixed $a, b > 0$, including $\epsilon = 0$ and additional zero Hessian eigenvalues.

3.1 COMPLEXITY ORACLES IN EXPECTATION

We now provide the complexity oracles of ZO-AdaNesterov in expectation for both the convex and non-convex cases. Before presenting the results, we make the following assumption on gradient noise, which is general in the literature Balasubramanian & Ghadimi (2022); Li et al. (2020).

Assumption 2. There exists $\sigma \geq 0$ such that $\mathbb{E}_\xi \left[\|\nabla F(x, \xi) - \nabla f(x)\|^2 \right] \leq \sigma^2$ for all $x \in \mathbb{R}^d$.

Assumption 2 is a mild assumption, which is weaker than bounded gradient assumption and sub-gaussian noise assumption used in Li & Orabona (2019).

We present the expected complexity for the convex case below, whose proof is in Appendix G.

Theorem 3.1. *Suppose that Assumptions 1 and 2 hold and $f(x)$ is convex. Define $\bar{w}_T = \frac{\sum_{n=1}^{T} w_n}{T}$. Let $\mu_1 \leq \frac{1}{d}$. Set $m = \lceil d \max\{1, 2\sigma^2 T\} \rceil$. For any fixed ϵ and θ satisfying $0 \leq \epsilon < \frac{1}{2}$ and $0 < \theta < 1$, we choose the parameters $a > 0$ and $b > 0$ in stepsize (2) such that*

$$b > 0, \quad 0 < a < \max\left\{ \frac{(1-\theta) b^{1/2+\epsilon}}{M\left(4 + \sqrt{16 + 8\theta^2}\right)}, \frac{(1-\theta)^2 b^{1/2+\epsilon}}{2M\left(1 - \theta + \sqrt{(1-\theta)^2 + \theta^4}\right)} \right\}.$$

Then run Algorithm 1, we have

$$\mathbb{E}\left[(f(\bar{w}_T) - f^*)^{\frac{1}{2}-\epsilon} \right] \leq \left(\frac{K_1}{T} \right)^{\frac{1}{2}-\epsilon},$$

where K_1 is a constant independent of m, T, σ, μ_1 and d, the explicit formula of K_1 is given in (25).

Corollary 3.2. *Suppose all conditions of Theorem 3.1 are satisfied. For $0 < \delta \leq 1$, in order to have $\mathbb{E}\left[(f(\bar{w}_T) - f^*)^{\frac{1}{2}-\epsilon} \right] \leq \delta^{\frac{1}{2}-\epsilon}$, we set the number of sampling m and iterations T to be*

$$T = \left\lceil \max\left\{ 1, \frac{K_1}{\delta} \right\} \right\rceil = \mathcal{O}\left(\frac{1}{\delta} \right), \quad m = \lceil d \max\{1, 2\sigma^2 T\} \rceil = \mathcal{O}\left(d + \frac{d\sigma^2}{\delta} \right),$$

thus the complexity of zeroth-order oracles for Algorithm 1 is $\mathcal{O}(\frac{d}{\delta} + \frac{d\sigma^2}{\delta^2})$.

Next, we present the expected complexity oracles of ZO-AdaNesterov for the nonconvex case, whose complete proof is left to Appendix H.

Theorem 3.3. *Suppose that Assumptions 1 and 2 hold. Set the stepsizes as in (2). Let $\mu_1 \leq \frac{1}{d}$. Set $m = \lceil d \max\{1, 2\sigma^2 T\} \rceil$. For any fixed ϵ and θ satisfying $0 \leq \epsilon < \frac{1}{2}$ and $0 < \theta < 1$, we choose the parameters $a > 0$ and $b > 0$ in (2) such that*

$$b > 0, \quad 0 < a < \frac{(1-\theta)^2 b^{1/2+\epsilon}}{2M\left(1 - \theta + \sqrt{(1-\theta)^2 + \theta^4}\right)}.$$

If R is drawn uniformly from $\{1, \ldots, T\}$, then run Algorithm 1, we have

$$\mathbb{E}\left[\min_{1 \leq n \leq T} \|\nabla f(w_n)\|^{1-2\epsilon} \right] \leq \mathbb{E}[\|\nabla f(w_R)\|^{1-2\epsilon}] \leq \left(\frac{K_2}{T} \right)^{\frac{1}{2}-\epsilon},$$

where K_2 is a constant independent of m, T, σ, μ_1 and d, the explicit formula of K_2 is given in (26).

Corollary 3.4. *Suppose that all conditions of Theorem 3.3 are satisfied. For $0 < \delta \leq 1$, in order to have $\mathbb{E}\left[\min_{1 \leq n \leq T} \|\nabla f(w_n)\|^{1-2\epsilon} \right] \leq \mathbb{E}[\|\nabla f(w_R)\|^{1-2\epsilon}] \leq \delta^{1-2\epsilon}$, set the number of sampling m and iterations T to be*

$$T = \left\lceil \max\left\{ 1, \frac{K_2}{\delta^2} \right\} \right\rceil = \mathcal{O}\left(\frac{1}{\delta^2} \right), \quad m = \lceil d \max\{1, 2\sigma^2 T\} \rceil = \mathcal{O}\left(d + \frac{d\sigma^2}{\delta^2} \right),$$

thus the complexity of zeroth-order oracles for Algorithm 1 is $\mathcal{O}(\frac{d}{\delta^2} + \frac{d\sigma^2}{\delta^4})$.

Remark 3.5. The complexity bounds established in Corollaries 3.2 and 3.4 accommodate unbounded oracles, rendering them highly applicable to empirical settings. In contrast, ZO-AdaMM Chen et al. (2019) relies on the assumption of uniformly bounded stochastic gradient oracles (i.e., there exists $C > 0$ such that $\|\nabla F(x, \xi)\|_\infty \leq C$), while R-AdaZO Shu et al. (2025) requires uniformly bounded function values (i.e., there exists $C > 0$ such that $\|F(x, \xi)\| \leq C$). Furthermore, our derived complexity bounds adapt naturally to the noise variance level. Specifically, the rates are $\mathcal{O}(\frac{d}{\delta} + \frac{d\sigma^2}{\delta^2})$ for convex objectives and $\mathcal{O}(\frac{d}{\delta^2} + \frac{d\sigma^2}{\delta^4})$ for non-convex objectives. In the noise-free regime where $\sigma = 0$, these bounds recover the standard rates for zeroth-order gradient descent: $\mathcal{O}(\frac{d}{\delta})$ for convex cases and $\mathcal{O}(\frac{d}{\delta^2})$ for non-convex cases.

3.2 Gradient Norm Vanishes Almost Surely

Unlike convergence in expectation, establishing almost sure convergence presents greater analytical challenges. Consequently, we require an stronger assumption than Assumption 2 below.

Assumption 3. $F(x,\xi)$ is L-Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$, and any ξ, it satisfies

$$\|F(x,\xi) - F(y,\xi)\| \leq L\|x-y\|.$$

This assumption is often used in the existing literature on zeroth-order optimization Ghadimi & Lan (2013); Balasubramanian & Ghadimi (2022). Below, we prove the gradient norm of points generated by (ZO-AdaNesterov) vanishes almost surely, whose proof is deferred to Appendix I.

Theorem 3.6. *Suppose that Assumptions 1 and 3 hold. Set stepsize η_n as (2). Fix any $\epsilon \in (0, \frac{1}{2}]$ and $\theta \in (0,1)$. For all $a, b, \mu_1 > 0$, run Algorithm 1, then we have*

$$\lim_{n\to\infty} \|\nabla f(w_n)\| = 0 \quad \textit{almost surely.}$$

3.3 Almost Sure Saddle Avoidance

Theorem 3.6 only provides a guarantee of gradient norm vanishing for (ZO-AdaNesterov). In the deterministic setting, the gradient descent algorithm with constant stepsizes, vanishing stepsizes and adaptive stepsizes converge to local minimums, i.e., avoid the saddle points (Lee et al. (2019); Panageas et al. (2019); Antonakopoulos et al. (2022)).

In the stochastic setting, stochastic algorithms with vanishing stepsizes avoid the hyperbolic unstable equilibria or saddle points both for the hyperbolicity case (i.e., $\nabla^2 f(x^*)$ is without zero eigenvalue) in Benaim & Hirsch (1995) and more difficult non-hyperbolicity case (i.e., $\nabla^2 f(x^*)$ exists zero eigenvalue) in Mertikopoulos et al. (2020). However, the saddle avoidance of stochastic algorithms with adaptive stepsizes remains open.

In Theorem 3.8, we solve this problem, i.e., proving that ZO-AdaNesterov avoids the saddle points almost surely in the non-hyperbolicity setting (The full proof is in Appendix J). Saddle avoidance is much harder than convergence to critical points. Hence, we need to present an additional assumption:

Assumption 4. For the fixed finite m, there are constants $\chi_m > 0$ and $\mu_{\text{exc}} > 0$ such that, for every $x \in \mathbb{R}^d$, unit vector u, and $0 < \mu \leq \mu_{\text{exc}}$,

$$\mathbb{E}\left[\left\langle \widehat{\nabla} f_U^{\mu}(x) - \nabla f_{\mu}(x), u \right\rangle^{+}\right] \geq \chi_m, \quad \text{where } t^+ = \max\{t, 0\}. \tag{4}$$

Remark 3.7. The noise $\widehat{\nabla} f_U^{\mu}(x) - \nabla f_{\mu}(x)$ stems from both sampling directions $\boldsymbol{U}$ and environmental noise ξ. Assumption 4 is naturally adaptive from the saddle avoidance literature (Mertikopoulos et al. (2020); Benaïm (2006); Liu & Yuan (2024)) and holds in common scenarios, such as finite-sum problems with non-degenerate noise (Mertikopoulos et al. (2020)). Furthermore, To demonstrate that Assumption 4 is practically justified, we provide a detailed example in Appendix C.

Theorem 3.8. *Suppose that Assumptions 1, 3 and 4 hold. Let $\mathcal{S}$ be the compact strict saddle manifold of f. Fix any $a, b, \mu_1 > 0$, $\epsilon \in [0, \frac{1}{2}]$ and $\theta \in (0,1)$. Run (ZO-AdaNesterov), then we have*

$$\mathbb{P}\left(\operatorname{dist}(w_n, \mathcal{S}) \to 0 \text{ as } n \to \infty\right) = 0.$$

Remark 3.9. It is worth noting that the almost sure gradient norm vanishing and saddle avoidance established in Theorems 3.6 and 3.8 holds fully adaptively for any choice of parameters $a, b > 0$. In contrast, the non-asymptotic convergence guarantees in Theorems 3.1 and 3.3 impose mild technical conditions on the hyperparameters a and b. These conditions can often be satisfied, for instance, by taking b sufficiently large or a sufficiently small, and are primarily introduced to control the accumulation of stochastic noise in the adaptive stepsizes. In practice, we can select a and b so as to satisfy the theoretical requirements while avoiding overly conservative stepsizes, striking a balance between theoretical validity and empirical performance.

4 Numerical Experiments

We compare ZO-AdaNesterov with ZO-SGD (Ghadimi & Lan, 2013), ZO-AdaMM (Chen et al., 2019), and R-AdaZO (Shu et al., 2025) on tensor decomposition and full-parameter LLM fine-tuning.

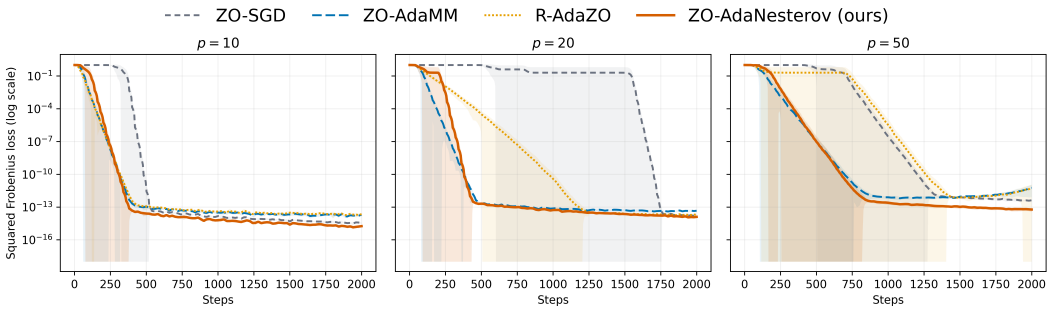

Figure 1: Tensor-decomposition loss over 2,000 steps for $K = 3$.

Table 2: Full-parameter GLUE validation performance (**five-seed** mean $\pm$ standard deviation, %)

Model	Method	MRPC	QNLI	RTE	SST-2	STS-B
RoBERTa -large	ZO-SGD	70.53 ± 4.00	68.20 ± 1.33	52.50 ± 3.38	90.32 ± 1.23	75.40 ± 6.87
	ZO-AdaMM	71.59 ± 6.12	69.56 ± 6.29	51.09 ± 6.19	90.00 ± 0.88	81.99 ± 1.99
	R-AdaZO	74.99 ± 2.48	69.92 ± 1.99	55.47 ± 6.25	91.24 ± 1.26	80.70 ± 4.10
	Ours	$\mathbf{75.15} \pm 1.49$	$\mathbf{74.04} \pm 2.13$	$\mathbf{58.13} \pm 4.44$	$\mathbf{91.67} \pm 0.89$	$\mathbf{83.16} \pm 3.22$
Qwen3.5-2B	ZO-SGD	76.55 ± 0.64	87.68 ± 0.41	82.31 ± 0.92	91.08 ± 0.35	89.09 ± 0.57
	ZO-AdaMM	76.89 ± 1.17	87.80 ± 0.62	81.08 ± 1.29	91.13 ± 0.27	89.41 ± 0.96
	R-AdaZO	77.31 ± 1.17	87.92 ± 0.36	82.09 ± 1.16	91.45 ± 0.44	89.33 ± 1.00
	Ours	$\mathbf{77.50} \pm 1.05$	$\mathbf{88.08} \pm 0.27$	$\mathbf{83.18} \pm 1.02$	$\mathbf{91.78} \pm 0.75$	$\mathbf{90.12} \pm 0.68$

Each method receives the same grid-search budget, and all results average **five seeds** on NVIDIA GeForce RTX 4090 GPUs. Appendix K provides complete protocols, hyperparameters, per-seed statistics, efficiency measurements, and additional classification and black-box attack results.

Tensor decomposition. We use the rank-one K-order tensor decomposition, a standard saddle avoidance benchmark in Ge et al. (2015); Ge & Ma (2017) with $K = 3$ and $p \in \{10, 20, 50\}$, corresponding to $3p$ trainable scalars and p^3 ambient tensor entries. This controlled nonconvex problem lets us examine how the optimizers behave as both the optimization dimension and the size of the reconstructed tensor increase. As shown in Figure 1, ZO-AdaNesterov descends fastest and reaches the lowest final loss for every p. It also obtains the smallest mean log-loss area under the curve at all three widths, indicating that its advantage persists throughout optimization rather than arising only near the final iterate.

Full-parameter LLM fine-tuning. To evaluate scalability in a genuinely high-dimensional black-box setting, we perform prompt-based, full-parameter zeroth-order fine-tuning on five GLUE tasks (Wang et al., 2018): MRPC, QNLI, RTE, SST-2, and STS-B. We optimize every parameter of RoBERTa-large (Liu et al., 2019) (355M parameters) and the text component of Qwen3.5-2B (Qwen Team, 2026) (1.88B parameters), without adding a task-specific classification head or using back-propagated gradients. Each optimizer step uses only function-value evaluations of the prompted mini-batch objective. The resulting validation scores, averaged over five seeds, are reported in Table 2. ZO-AdaNesterov attains the highest mean on all five tasks for both model families. These results indicate that the proposed ZO-AdaNesterov remains effective as the number of optimized parameters increases from hundreds of millions to nearly two billion.

5 Conclusion

In this work, we introduced ZO-AdaNesterov, a novel zeroth-order optimization framework incorporating adaptive step-sizes and Nesterov momentum. Our theoretical results strictly improve the complexity bounds upon prior adaptive methods, such as ZO-AdaMM and R-AdaZO. Furthermore, we provided rigorous guarantees for almost sure saddle point avoidance, and validated the algorithm's efficiency through extensive experiments on real-world datasets. Extending the results of this paper to constrained case is a promising future direction.

AI USE STATEMENT

In this work, we used generative AI tools for checking the spelling errors, grammatical accuracy and reorganizing the paper's structure. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The update rules and zeroth-order gradient estimator of ZO-AdaNesterov are fully specified in the Setup and Algorithms section. All assumptions and theoretical claims are stated in the main paper, with complete proofs provided in the appendix. The experimental protocols, datasets, preprocessing procedures, hyperparameter settings, random seeds, evaluation budgets, and computational resources are described in Section 4 and Appendix K. Anonymous source code, configuration files, and scripts for reproducing the reported tables and figures are provided as supplementary material.

REFERENCES

Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Kimon Antonakopoulos, Panayotis Mertikopoulos, Georgios Piliouras, and Xiao Wang. Adagrad avoids saddle points. In *International Conference on Machine Learning*, pp. 731–771. PMLR, 2022.

Amit Attia and Tomer Koren. Sgd with adagrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, pp. 1147–1171. PMLR, 2023.

Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pp. 1–42, 2022.

Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pp. 1–68. Springer, 2006.

Michel Benaim and Morris W Hirsch. Dynamics of morse-smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6):1005–1030, 1995.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018a.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In *NeurIPS*, 2019.

Zaiyi Chen, Yi Xu, Enhong Chen, and Tianbao Yang. Sadagrad: Strongly adaptive stochastic gradient methods. In *International Conference on Machine Learning*, pp. 913–921. PMLR, 2018b.

Sayantan Choudhury, Nazarii Tupitsa, Nicolas Loizou, Samuel Horváth, Martin Takac, and Eduard Gorbunov. Remove that square root: A new efficient scale-invariant version of adagrad. *Advances in Neural Information Processing Systems*, 37:47400–47431, 2024.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

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.

Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 89–160. PMLR, 2023.

Rong Ge and Tengyu Ma. On the optimization landscape of tensor decompositions. *Advances in neural information processing systems*, 30, 2017.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pp. 797–842. PMLR, 2015.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

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

Yusu Hong and Junhong Lin. On convergence of adam for stochastic optimization under relaxed assumptions. *Advances in Neural Information Processing Systems*, 37:10827–10877, 2024.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pp. 545–604. PMLR, 2018.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980, 2014.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2307.04504*, 2023.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176:311–337, 2019.

Rémi Leluc and François Portier. Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*, 2023.

Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. *Advances in neural information processing systems*, 31, 2018.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.

Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Zeroth-order optimization on riemannian manifolds. *arXiv preprint arXiv:2003.11238*, 3, 2020.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pp. 983–992. PMLR, 2019.

Tianyi Lin, Zeyu Zheng, and Michael Jordan. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35: 26160–26175, 2022.

Jun Liu and Ye Yuan. Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods. *Journal of Machine Learning Research*, 25(271):1–40, 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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 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. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

H.B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, 2010.

Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. In *International conference on machine learning*, pp. 2545–2553. PMLR, 2017.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Y Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:3, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. *Advances in Neural Information Processing Systems*, 32, 2019.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Marco Rando, Cesare Molinari, Silvia Villa, and Lorenzo Rosasco. Stochastic zeroth order descent with structured directions. *Computational Optimization and Applications*, pp. 1–37, 2024.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

Zhaolin Ren, Yujie Tang, and Na Li. Escaping saddle points in zeroth-order optimization: two function evaluations suffice. *ICML*, 2023.

Yao Shu, Qixin Zhang, Kun He, and Zhongxiang Dai. Refining adaptive zeroth-order optimization at ease. In *Forty-second International Conference on Machine Learning*, 2025.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. pmlr, 2013.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in neural information processing systems*, 32, 2019.

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 *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of adam's iteration complexity. *Advances in Neural Information Processing Systems*, 36:39006–39032, 2023a.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023b.

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2960–2969, 2024.

Xiaoxia Wu, Rachel Ward, and Léon Bottou. Wngrad: Learn the learning rate in gradient descent. *arXiv preprint arXiv:1803.02865*, 2018.

Zi Xu, Zi-Qi Wang, Jun-Lin Wang, and Yu-Hong Dai. Zeroth-order alternating gradient descent ascent algorithms for a class of nonconvex-nonconcave minimax problems. *Journal of Machine Learning Research*, 24(313):1–25, 2023.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012.

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

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Yan Zhang, Yi Zhou, Kaiyi Ji, and Michael M Zavlanos. A new one-point residual-feedback oracle for black-box learning and control. *Automatica*, 136:110006, 2022a.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D Lee, Wotao Yin, Mingyi Hong, et al. Revisiting zeroth-order optimization for memory-efficient llm fine-tuning: A benchmark. *arXiv preprint arXiv:2402.11592*, 2024.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35: 28386–28399, 2022b.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11127–11135, 2019.

APPENDIX

A Further Related Work

AdaGrad. The convergence guarantees associated with adaptive Stochastic Gradient Descent (AdaGrad) have received meticulous investigation within the online convex optimization framework (Duchi et al. (2011); Reddi et al. (2019)) and reference therein. Mukkamala & Hein (2017) even derived the logarithmic regret bounds for several variants of AdaGrad, however, their results are restricted to the strongly convex setting. Besides, Yousefian et al. (2012) conducted an analysis of different stepsize variants specifically within the strongly convex setting. Moving beyond convexity, Chen et al. (2018b) investigated the convergence rates (in expectation) for AdaGrad in the nonconvex setting, but their rates rely heavily on the dimension. In refining these rates and alleviating the dependence on dimensions, Zhou et al. (2018) made important contributions. Furthermore, Li & Orabona (2019) proved the almost sure convergence to critical points and the convergence rates (in expectation) for AdaGrad in both convex and nonconvex settings. Leluc & Portier (2023) employed a unified framework to provide the almost sure convergence to critical points for many stochastic first-order algorithms including AdaGrad. Other convergence rates for AdaGrad mostly require stronger assumptions including the norm of gradient not be zero (Wu et al. (2018)) and bounded domain, i.e., the gradient oracle g_n satisfies $\|g_n\| \leq G$ (Levy et al. (2018)). Additionally, Défossez et al. (2020) provided the convergence rates for AdaGrad in the nonconvex setting. Attia & Koren (2023) provided non-asymptotic rates for AdaGrad with high probability. Wang et al. (2023b); Faw et al. (2023) researched convergence results for AdaGrad under relaxed smoothness assumption.

Adam. Early theoretical analyses of Adam were heavily predicated on strong assumptions, such as bounded oracles or bounded gradients (Zaheer et al. (2018); Chen et al. (2018a); Zou et al. (2019); Défossez et al. (2020)). To mitigate these restrictions, recent studies have focused on high-probability guarantees under more realistic conditions. Specifically, Li et al. (2023) established convergence bounds under sub-Gaussian noise and relaxed smoothness, while Hong & Lin (2024) further generalized this to accommodate mild affine noise. Notably, Wang et al. (2023a) proved that Adam achieves an upper bound matching the fundamental lower bound for first-order stochastic optimization. Parallel to these stochastic results, convergence properties have also been explored in the finite-sum setting (Zhang et al. (2022b); Wang et al. (2024)).

B Derivation and properties of ZO-AdaNesterov

In this section, we provide the missing details of deriving ZO-AdaNesterov algorithm. Set

$$q = 1 - \theta$$

with $0 < \theta < 1$. Define

$$v_n = w_n - w_{n-1}, \qquad z_n = w_n + \frac{\theta}{q} v_n. \tag{5}$$

By initialization $v_1 = 0$, we have $z_1 = w_1 = w_0$. From the two actual updates,

$$\begin{aligned} v_{n+1} &= \theta v_n - \eta_n g_n, \qquad & z_{n+1} &= z_n - \frac{\eta_n}{q} g_n, \\ x_n &= w_n + \theta v_n, & z_n - x_n &= \frac{\theta^2}{q} v_n. \end{aligned} \tag{6}$$

Indeed, subtract w_n from the first update for the first identity, and calculate

$$z_{n+1} - z_n = v_{n+1} + \frac{\theta}{q}(v_{n+1} - v_n) = \frac{1}{q} v_{n+1} - \frac{\theta}{q} v_n = -\frac{\eta_n}{q} g_n.$$

The last identity follows from $\theta/q - \theta = \theta^2/q$. All these coordinates are predictable.

Lemma B.1. *Let*

$$P_T = \sum_{n=1}^{T} \eta_n^2 \|g_n\|^2 .$$

For positive stepsizes bounded by h and $v_1 = 0$,

$$\sum_{n=1}^{T} \|v_n\|^2 \le \frac{P_T}{q^2}, \qquad \sum_{n=1}^{T} \eta_n \|v_n\|^2 \le \frac{hP_T}{q^2} \tag{7}$$

pathwise. No restriction on θ beyond $0 < \theta < 1$ is required.

Proof. Unroll $v_n = -\sum_{i<n} \theta^{n-1-i}\eta_i g_i$. Weighted Cauchy–Schwarz gives

$$\|v_n\|^2 \le \frac{1}{q} \sum_{i<n} \theta^{n-1-i}\eta_i^2 \|g_i\|^2 .$$

Sum over n and interchange the finite sums. The second geometric sum is at most $1/q$. This proves the first inequality, and $\eta_n \le h$ proves the second. □

Lemma B.2. *For every differentiable convex ϕ and every reference point y,*

$$\langle \nabla\phi(x_n), z_n - y\rangle \ge \frac{\phi(w_n) - \theta\phi(w_{n-1}) - q\phi(y)}{q}. \tag{8}$$

Proof. Write $H_n^\phi = \nabla\phi(x_n)$. Since $x_n - w_{n-1} = (1+\theta)v_n$,

$$\begin{aligned}\langle H_n^\phi, z_n - y\rangle &= \langle H_n^\phi, x_n - y\rangle + \frac{\theta^2}{q(1+\theta)}\langle H_n^\phi, x_n - w_{n-1}\rangle \\ &\ge \phi(x_n) - \phi(y) + \frac{\theta^2}{q(1+\theta)}[\phi(x_n) - \phi(w_{n-1})].\end{aligned}$$

Also $w_n = (x_n + \theta w_{n-1})/(1+\theta)$, so $\phi(x_n) \ge (1+\theta)\phi(w_n) - \theta\phi(w_{n-1})$. The coefficient of $\phi(x_n)$ above is $1/[q(1+\theta)] > 0$. Substitution yields coefficients $1/q, -\theta/q, -1$ on the three function values, proving (8). □

C A PRACTICAL EXAMPLE SATISFYING ASSUMPTION 4

Let f be M-smooth and L_f-Lipschitz, and take

$$F(x,\xi) = f(x) + \langle \xi, x\rangle, \qquad \mathbb{E}[\xi] = 0, \quad \|\xi\| \le R_\xi, \quad \mathbb{E}[\xi\xi^{\mathsf{T}}] \succeq \lambda_\xi I_d, \quad \lambda_\xi > 0.$$

Then Assumptions 1 and 3 hold with $L = L_f + R_\xi$. Let $H := dL$, $g = \widehat{\nabla} f_U^\mu(x)$, $\zeta = g - \nabla f_\mu(x)$, and $X = \langle \zeta, u\rangle$ for a unit u. Since $\mathbb{E}[X] = 0$ and $|X| \le 2H$, one has

$$\mathbb{E}[X^+] = \frac{1}{2}\mathbb{E}[|X|] \ge \frac{\mathbb{E}[X^2]}{4H}. \tag{9}$$

Conditional on all directions, the environmental part of the independent-pair average is

$$\frac{d}{m}\sum_{i=1}^{m} \langle \xi_i, u_i\rangle u_i.$$

Independence and centering of the ξ_i show that its variance in direction u is at least $\lambda_\xi d^2 m^{-2}\sum_i \langle u_i, u\rangle^2$. The law of total variance and sphere isotropy therefore give

$$\mathbb{E}[X^2] = \mathrm{Var}(\langle g, u\rangle) \ge \frac{\lambda_\xi d^2}{m^2}\sum_{i=1}^{m} \mathbb{E}[\langle u_i, u\rangle^2] = \frac{\lambda_\xi d}{m}.$$

Together with (9), this verifies Assumption 4 with

$$\chi_m = \frac{\lambda_\xi}{4Lm} > 0. \tag{10}$$

D STATEMENT OF CLASSICAL LEMMAS AND FOR COMPLETENESS

Lemma D.1 (Hölder's inequality)**.** *For nonnegative integrable X, Y and $0 < s < 1$, $\mathbb{E}[X^s Y^{1-s}] \le (\mathbb{E}[X])^s (\mathbb{E}[Y])^{1-s}$.*

Proof. If either expectation is zero the result is immediate. Otherwise normalize X, Y by their expectations and apply $u^s v^{1-s} \le su + (1-s)v$. This scalar inequality follows by maximizing $u^s - su$ after scaling v to one. Taking expectations proves the claim. □

Lemma D.2 (Weighted deterministic convergence)**.** *If $a_n, s_n \ge 0$, $\sum_n a_n = \infty$, $\sum_n a_n s_n < \infty$, and $|s_{n+1} - s_n| \le C a_n$ for some finite C, then $s_n \to 0$.*

Proof. The two series imply $\liminf s_n = 0$. If $C = 0$, the result follows immediately. Otherwise, failure of convergence to zero gives some $\rho > 0$ and infinitely many disjoint downward excursions, starting at $s_k \ge 2\rho$ and ending at the first $j > k$ with $s_j \le \rho$. On each excursion $s_i > \rho$ for $k \le i < j$ and

$$C \sum_{i=k}^{j-1} a_i \ge \sum_{i=k}^{j-1} |s_{i+1} - s_i| \ge \rho.$$

Thus each excursion contributes at least ρ^2/C to $\sum a_n s_n$, a contradiction. No assumption $a_n \to 0$ is needed. □

Lemma D.3. *If $b > 0$, $Y_n \ge 0$, and $e > 1$, then*

$$\sum_{n=1}^{T} \frac{Y_n}{(b + \sum_{i=1}^{n} Y_i)^e} \le \frac{1}{(e-1)b^{e-1}}.$$

Proof. The decreasing function s^{-e} makes each summand no greater than its integral over $[b + \sum_{i<n} Y_i, b + \sum_{i \le n} Y_i]$. Summing and integrating to infinity proves the bound. □

To continue the progress of proofs, we define the μ-smoothing versions of $f(x)$ and $F(x, \xi)$, i.e.,

$$f_\mu(x) = \mathbb{E}_{s \sim \mathrm{Unif}(B_1(0))}[f(x + \mu s)], \qquad F_\mu(x, \xi) = \mathbb{E}_s[F(x + \mu s, \xi)].$$

$f_\mu(x)$ and $F_\mu(x, \xi)$ will be frequently used in the theoretical analysis of our zeroth-order adaptive step algorithm.

Lemma D.4. *Under Assumptions 1–2, $\nabla f(x) = \mathbb{E}_\xi[\nabla F(x, \xi)]$, f is M-smooth, and*

$$\mathbb{E}_u[q_\mu(x; u, \xi)] = \nabla F_\mu(x, \xi), \qquad \mathbb{E}_{u,\xi}[q_\mu(x; u, \xi)] = \nabla f_\mu(x). \tag{11}$$

The same conclusions hold under Assumptions 1 and 3 without Assumption 2.

Proof. At a fixed x, Assumption 2 gives an integrable $\|\nabla F(x, \xi)\|$. Assumption 1 bounds sample gradients on a bounded neighborhood by this random variable plus a deterministic constant. Difference quotients have the same local domination, so dominated convergence interchanges differentiation and environmental expectation. Averaging the sample gradient Lipschitz inequality gives M-smoothness of f. Ball averaging is justified by the same local bound.

Let V_d be the volume of $B_1(0)$, so its boundary has area dV_d. The divergence theorem applied coordinatewise gives

$$\partial_j F_\mu(x, \xi) = \frac{1}{\mu V_d} \int_{S_1(0)} F(x + \mu u, \xi) u_j \, dS(u).$$

The constant $F(x, \xi)$ can be subtracted because the mean of u is zero. Dividing the integral by the sphere area gives the first identity in (11); environmental expectation gives the second. When $d = 1$, the fundamental theorem of calculus on $[-1, 1]$ gives the same formula. Under Assumption 3, $\|\nabla F\| \le L$ replaces the local integrable bound and proves all the same interchanges. □

Lemma D.5. *Under Assumptions 1–2, finite m, deterministic initialization, bounded deterministic radii on a finite horizon, and $\eta_n \le h < \infty$, all iterates have finite second moments. The function values and squared gradients used in the finite-horizon proofs are integrable. The same holds under Assumptions 1 and 3.*

Proof. The fixed-point calculation of Lemma D.13 gives

$$\mathbb{E}_n[\|g_n\|^2] \le C_1 \|\nabla f(x_n)\|^2 + C_2 \le C_3(1 + \|x_n\|^2)$$

for finite deterministic constants on the chosen horizon. The latter step uses $\|\nabla f(x)\| \le \|\nabla f(0)\| + M\|x\|$. Induction in the two linear update formulas yields finite second moments of w_n, x_n and then v_n, z_n. Smoothness implies

$$|f(x)| \le |f(0)| + \|\nabla f(0)\| \|x\| + \frac{M}{2}\|x\|^2,$$

which gives the function-value claim. Under Assumption 3, the direct bound $\|g_n\| \le dL$ suffices. Lemma D.13 is a fixed-point calculation and does not invoke this integrability lemma, so this forward reference is not circular. □

Lemma D.6. *For M-smooth f with $f^* > -\infty$, $\|\nabla f(x)\|^2 \le 2M(f(x) - f^*)$.*

Proof. For $M > 0$, apply smoothness to $x - \nabla f(x)/M$ and compare to f^*. For $M = 0$, the gradient is constant, and lower boundedness on $\mathbb{R}^d$ forces it to be zero. □

Lemma D.7. *Let $0 < p < 1$, $r = 1 - p$, and $x, A, B, D \ge 0$ be finite. If $x \le D^r(A + Bx)^p$, then*

$$x \le \max\{D^r(2A)^p,\ D(2B)^{p/r}\}.$$

Proof. If $Bx \le A$, use $A + Bx \le 2A$. Otherwise $x > 0$, and dividing $x \le D^r(2Bx)^p$ by x^p gives $x^r \le D^r(2B)^p$. The case $B = 0$ belongs to the first branch. □

Lemma D.8. *If $\mu_n = \mu_1/n$, then $\sum_{n=1}^T \mu_n^2 \le 2\mu_1^2$, also for $T = \infty$.*

Proof. Use $\sum_{n\ge1} n^{-2} \le 1 + \int_1^\infty t^{-2}\,dt = 2$. □

Lemma D.9. *For $x, y \ge 0$ and $0 < r \le 1$, $(x + y)^r \le x^r + y^r$.*

Proof. When $x > 0$, put $t = y/x$ and observe that $(1 + t)^r - 1 - t^r$ is nonincreasing for $t > 0$ and vanishes at $t = 0$. The case $x = 0$ is immediate. □

Lemma D.10. *For $x > 0$, $-\log x \ge 1 - x$.*

Proof. The derivative of $x - 1 - \log x$ is $1 - 1/x$; its minimum is zero at $x = 1$. □

Lemma D.11. *If $X \ge 0$ and $\mathbb{E}[X] < \infty$, then $X < \infty$ almost surely.*

Proof. For each $B > 0$, Markov's inequality gives $\mathbb{P}(X \ge B) \le \mathbb{E}[X]/B$. Let $B \to \infty$, we can easily get the stated result. □

Lemma D.12. *For $u \sim \mathrm{Unif}(S_1(0))$ and $w \in \mathbb{R}^d$,*

$$\mathbb{E}[uu^\top] = I_d/d, \qquad \mathbb{E}[\langle w, u\rangle\, u] = w/d, \qquad \mathbb{E}[\langle w, u\rangle^2] = \|w\|^2/d.$$

Proof. Coordinate sign symmetry eliminates off-diagonal entries, coordinate permutation symmetry equates the diagonal entries, and the trace is $\mathbb{E}\|u\|^2 = 1$. The other identities follow by multiplication with w. □

Lemma D.13. *Under Assumptions 1–2 and independent pairs, at every fixed x, μ and finite $m \geq 1$,*

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m} q_\mu(x;u_i,\xi_i)\right\|^2\right] = \frac{1}{m}\mathbb{E}\left[\|q_\mu(x;u,\xi)\|^2\right] + \frac{m-1}{m}\|\nabla f_\mu(x)\|^2. \tag{12}$$

Consequently,

$$\mathbb{E}\left[\left\|\widehat{\nabla} f_U^\mu(x)\right\|^2\right] \leq \left(2+\frac{2d}{m}\right)\|\nabla f(x)\|^2 + \frac{2d\sigma^2}{m} + \left(2+\frac{d^2}{2m}\right)M^2\mu^2. \tag{13}$$

The mean is $\nabla f_\mu(x)$.

Proof. For fixed ξ, sample smoothness gives

$$F(x+\mu u,\xi) - F(x,\xi) = \mu\langle \nabla F(x,\xi), u\rangle + e(u,\xi), \qquad |e(u,\xi)| \leq M\mu^2/2.$$

Using $\|u\| = 1$, $(s+t)^2 \leq 2s^2 + 2t^2$, and Lemma D.12,

$$\mathbb{E}_u[\|q_\mu(x;u,\xi)\|^2] \leq 2d\|\nabla F(x,\xi)\|^2 + \frac{d^2M^2\mu^2}{2}.$$

Since $\mathbb{E}_\xi \nabla F = \nabla f$, Assumption 2 implies

$$\mathbb{E}_{u,\xi}[\|q_\mu\|^2] \leq 2d\|\nabla f(x)\|^2 + 2d\sigma^2 + \frac{d^2M^2\mu^2}{2}.$$

The independent q_i have common mean $\nabla f_\mu(x)$ by Lemma D.4. Expanding the averaged squared norm retains $m(m-1)$ cross terms, each equal in expectation to $\|\nabla f_\mu(x)\|^2$, proving (12). Also,

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq M\mu, \qquad \|\nabla f_\mu(x)\|^2 \leq 2\|\nabla f(x)\|^2 + 2M^2\mu^2.$$

The first inequality follows by averaging population gradient smoothness over the ball. Substitution into (12) and $(m-1)/m \leq 1$ prove (13). □

Lemma D.14. *Under either set of assumptions ensuring Lemma D.4,*

$$\|\nabla f_\mu(x) - \nabla f(x)\| \leq M\mu, \qquad \|\nabla f(x)\|^2 \leq 2\|\nabla f_\mu(x)\|^2 + 2M^2\mu^2.$$

Proof. Average $\|\nabla f(x+\mu s) - \nabla f(x)\| \leq M\mu\|s\|$ on the unit ball. The second inequality is the squared triangle inequality with factor two. □

Lemma D.15. *Under the same assumptions, $|f_\mu(x) - f(x)| \leq M\mu^2/2$. If f is convex, so is f_μ, and $f_\mu(x) \geq f(x)$.*

Proof. Average the two-sided smoothness remainder about x. The first-order term has mean zero and $\|s\| \leq 1$. Translations and averages preserve convexity, and Jensen's inequality gives the lower bound. □

Lemma D.16. *Let $Y_n \geq 0$ be adapted and integrable. Set*

$$S_T = b + \sum_{n=1}^{T} Y_n, \qquad A_n = b + \sum_{i=1}^{n}\mathbb{E}[Y_i \mid \mathcal{F}_{i-1}], \qquad A_0 = b > 0.$$

For every $0 < r < 1$,

$$\mathbb{E}[S_T^r] \leq r^{-r}\mathbb{E}[A_T^r]. \tag{14}$$

Proof. Here A_n is $\mathcal{F}_{n-1}$-measurable, and $A_T \geq A_n \geq b$. Hence

$$\begin{aligned}\mathbb{E}\left[\frac{S_T}{A_T^{1-r}}\right] &\leq b^r + \sum_{n=1}^{T} \mathbb{E}\left[\frac{Y_n}{A_n^{1-r}}\right] \\ &= b^r + \mathbb{E}\left[\sum_{n=1}^{T} \frac{A_n - A_{n-1}}{A_n^{1-r}}\right] \\ &\leq b^r + \mathbb{E}\left[\int_b^{A_T} s^{r-1}\, ds\right] \leq \frac{1}{r}\mathbb{E}[A_T^r].\end{aligned}$$

The decreasing integrand justifies the penultimate step. Hölder gives

$$\mathbb{E}[S_T^r] = \mathbb{E}\left[\left(\frac{S_T}{A_T^{1-r}}\right)^r (A_T^r)^{1-r}\right] \leq \left(\mathbb{E}\left[\frac{S_T}{A_T^{1-r}}\right]\right)^r (\mathbb{E}[A_T^r])^{1-r},$$

proving (14). All terms are finite since $A_T \geq b > 0$ and Y_n are integrable. Only the predictable A_n, never the terminal A_T, passes through conditional expectation. □

E Energy and moment lemmas for general parameters

Before formally presenting the proofs of all theorems, we define some important quantities below.

Let

$$p = \tfrac{1}{2} + \epsilon, \qquad h = \frac{a}{b^p}, \qquad a, b > 0. \tag{15}$$

For $0 \leq \epsilon \leq 1/2$, define

$$r := 1 - p = 1/2 - \epsilon.$$

Define the two increasing absorption coefficients

$$\kappa_{\mathrm{C}}(h) = \frac{8Mh}{q} + \frac{8M^2\theta^2h^2}{q^2}, \qquad \kappa_{\mathrm{N}}(h) = \frac{4Mh}{q} + \frac{4M^2\theta^4h^2}{q^4}. \tag{16}$$

For $M > 0$, their positive unit-level roots are

$$h_{\mathrm{C}}^{\mathrm{dir}} = \frac{q}{M\left(4 + \sqrt{16 + 8\theta^2}\right)}, \qquad h_{\mathrm{N}} = \frac{q^2}{2M\left(q + \sqrt{q^2 + \theta^4}\right)}. \tag{17}$$

Set $h_{\mathrm{C}} = \max\{h_{\mathrm{C}}^{\mathrm{dir}}, h_{\mathrm{N}}\}$.

Let

$$N_\star = 1 + 5M^2, \qquad c_v = \frac{2M^2\theta^2}{q^2}. \tag{18}$$

For finite nonnegative k, Q, W, define

$$\mathcal{L}_b(Q) = r^{-r}\left[(b + N_\star)^r + (4c_vQ)^r\right], \quad \mathcal{J}(k) = r^{-r}(4k)^r, \tag{19}$$

$$\mathcal{K}_{a,b}(k, Q, W) = \left[\max\left\{(W/a)^r(2\mathcal{L}_b(Q))^p, (W/a)(2\mathcal{J}(k))^{p/r}\right\}\right]^{1/r}. \tag{20}$$

Note that (19) and (20) are used only for $0 \leq \epsilon < 1/2$. In particular $W = 0$ gives $\mathcal{K}_{a,b} = 0$, including when $k = 0$. The formula is valid for every $b > 0$.

For $\Delta_0 := f(w_1) - f^*$, define the deterministic progress constants

$$A_1(h) = \frac{qR_0^2}{2} + \frac{h(M + \theta\Delta_0)}{q}, \qquad \text{where } R_0 = \|w_1 - \tilde{x}^*\|, \tag{21}$$

$$A_2(h) = 2q\Delta_0 + 2M^2h, \qquad B_2(h) = \frac{M}{q} + \frac{M^2\theta^4h}{q^4}. \tag{22}$$

The two energy bounds, each defined only when its denominator is positive, are

$$Q_{\mathrm{C}}^{\mathrm{dir}}(h) = \frac{16MhA_1(h) + h^2N_\star}{1 - \kappa_{\mathrm{C}}(h)} \qquad \text{if } \kappa_{\mathrm{C}}(h) < 1, \tag{23}$$

$$Q_{\mathrm{N}}(h) = \frac{4hA_2(h) + h^2N_\star}{1 - \kappa_{\mathrm{N}}(h)} \qquad \text{if } \kappa_{\mathrm{N}}(h) < 1. \tag{24}$$

For a convex objective and $h < h_{\mathrm{C}}$, define $Q_1(h)$ to be the minimum of the *available* bounds in (23)–(24). This is a minimum of one or two finite numbers, never an expression with a nonpositive denominator. Set

$$W_1(h) = A_1(h) + \frac{Q_1(h)}{2q}, \qquad K_1 = \mathcal{K}_{a,b}(4M, Q_1(h), W_1(h)). \tag{25}$$

For a general lower-bounded smooth objective and $h < h_{\mathrm{N}}$, set

$$\begin{aligned} Q_2(h) = Q_{\mathrm{N}}(h), \qquad W_2(h) &= 2\big(A_2(h) + B_2(h)Q_2(h)\big) + c_v h Q_2(h), \\ K_2 &= \mathcal{K}_{a,b}(2, Q_2(h), W_2(h)). \end{aligned} \tag{26}$$

Note that (25) and (26) are used only for $0 \le \epsilon < 1/2$. For fixed admissible a, b, these constants depend only on $M, \theta, \epsilon, a, b$ and the stated initialization quantities, not on d, m, T, σ or on $\mu_1 \le 1/d$.

For a finite run, we define,

$$\nu^2 = \frac{2d\sigma^2}{m}, \qquad \rho_n = \nu^2 + \left(2 + \frac{d^2}{2m}\right) M^2\mu_n^2, \qquad N_T = \sum_{n=1}^{T} \rho_n. \tag{27}$$

The batch rule and radius bound imply

$$N_T \le \frac{2d\sigma^2 T}{m} + M^2\left(\frac{4}{d^2} + \frac{1}{m}\right) \le 1 + 5M^2 = N_\star. \tag{28}$$

There is no comparison of $N_\star$ with b. Let

$$P_T = \sum_{n=1}^{T} \eta_n^2 \left\|g_n\right\|^2, \qquad Q_T = \mathbb{E}[P_T], \qquad G_T = \mathbb{E}[\sum_{n=1}^{T} \eta_n \left\|\nabla f(x_n)\right\|^2]. \tag{29}$$

Under Assumptions 1–2 and the choice of m, Lemma D.13 yields

$$\mathbb{E}_n[g_n] = \nabla f_{\mu_n}(x_n), \qquad \mathbb{E}_n[\left\|g_n\right\|^2] \le 4\left\|\nabla f(x_n)\right\|^2 + \rho_n. \tag{30}$$

Lemma E.1. *Under these finite-horizon assumptions,*

$$Q_T \le 4hG_T + h^2N_T \le 4hG_T + h^2N_\star. \tag{31}$$

Proof. By predictability and (30),

$$Q_T = \mathbb{E}\sum_n \eta_n^2 \mathbb{E}_n[\left\|g_n\right\|^2] \le 4\mathbb{E}\left[\sum_n \eta_n^2 \left\|\nabla f(x_n)\right\|^2\right] + h^2N_T.$$

Using $\eta_n^2 \le h\eta_n$ and (28), we can get the stated result. □

Lemma E.2 (Weighted descent before any absorption)**.** *Under Assumptions 1–2, or instead Assumptions 1 and 3, let $f^* > -\infty$, $a, b > 0$, and $\mu_n = \mu_1/n$. For every finite T,*

$$G_T \le 2q\Delta_0 + 2M^2h\mu_1^2 + B_2(h)Q_T. \tag{32}$$

In the finite-horizon regime $\mu_1 \le 1/d \le 1$, this implies

$$G_T \le A_2(h) + B_2(h)Q_T. \tag{33}$$

No small-step assumption is needed for this lemma itself.

Proof. For vectors A, B, e,

$$\langle A+B, A+e\rangle \geq \tfrac{1}{2}\|A\|^2 - \tfrac{1}{2}\|B\|^2 - \tfrac{1}{2}\|e\|^2, \tag{34}$$

because the difference equals $\|A+B+e\|^2/2$. Use

$$A = \nabla f(x_n), \quad B = \nabla f(z_n) - \nabla f(x_n), \quad e_n = \nabla f_{\mu_n}(x_n) - \nabla f(x_n).$$

Lemma D.14 and (6) give $\|B\| \leq M\theta^2\|v_n\|/q$ and $\|e_n\| \leq M\mu_n$. Smoothness along $z_{n+1} = z_n - \eta_n g_n/q$, followed by conditional expectation, yields

$$\mathbb{E}_n[f(z_{n+1})] \leq f(z_n) - \frac{\eta_n}{2q}\|\nabla f(x_n)\|^2 + \frac{M^2\theta^4\eta_n}{2q^3}\|v_n\|^2 + \frac{M^2\eta_n\mu_n^2}{2q} + \frac{M\eta_n^2}{2q^2}\mathbb{E}_n[\|g_n\|^2]. \tag{35}$$

All expectations are finite by Lemma D.5. Sum, use $f(z_{T+1}) \geq f^*$ and $z_1 = w_1$, multiply by $2q$, and apply Lemmas B.1 and D.8. This gives

$$G_T \leq 2q\Delta_0 + 2M^2 h\mu_1^2 + \left(\frac{M}{q} + \frac{M^2\theta^4 h}{q^4}\right) Q_T,$$

as required. In particular the smoothing bias has not been discarded. □

Lemma E.3. *Under Assumption 3, $\|g_n\| \leq H := dL$ almost surely, for every finite $m \geq 1$.*

Proof. Each summand has norm at most $(d/\mu_n)L\mu_n = dL$, and the norm of the average is no greater than the average of the norms. □

Recall that $\eta_n = a/\left(b + \sum_{i=1}^{n-1}\|g_i\|^2\right)^p$. We can get the following lemma.

Lemma E.4. *Let $0 < p < 1$, $r = 1-p$, $a, b > 0$, and let the estimator have integrable squared norms and satisfy (30) with $\sum_{n=1}^{T}\rho_n \leq N_\star$. Let $\ell_n \geq 0$ be predictable, $D_T = \sum_{n=1}^{T}\ell_n$, and suppose that finite deterministic $k, Q, W \geq 0$ satisfy*

$$\mathbb{E}[P_T] \leq Q, \qquad \mathbb{E}\left[\sum_{n=1}^{T}\eta_n\ell_n\right] \leq W, \qquad \sum_{n=1}^{T}\|\nabla f(x_n)\|^2 \leq kD_T + c_v P_T \tag{36}$$

pathwise for the last inequality. Then

$$\left(\mathbb{E}[D_T^r]\right)^{1/r} \leq \mathcal{K}_{a,b}(k, Q, W). \tag{37}$$

The function in (20) is used with the same $N_\star, c_v$.

Proof. Let $S_T = b + \sum_{n\leq T}\|g_n\|^2$, $A_T = b + \sum_{n\leq T}\mathbb{E}_n[\|g_n\|^2]$, and $X = \mathbb{E}[D_T^r]$. Because $\eta_n \geq a/S_T^p$,

$$\mathbb{E}[D_T/S_T^p] \leq W/a.$$

Also $\mathbb{E}[S_T] < \infty$, so $\mathbb{E}[S_T^r] < \infty$. Hölder first gives

$$X = \mathbb{E}[(D_T/S_T^p)^r (S_T^r)^p] \leq (W/a)^r(\mathbb{E}[S_T^r])^p < \infty.$$

Thus no rearrangement below involves an unknown infinite quantity. If $W = 0$, positivity of the steps gives $D_T = 0$ almost surely and the conclusion is immediate.

Using the conditional moment and the pathwise comparison,

$$A_T \leq b + N_\star + 4kD_T + 4c_v P_T.$$

Lemma D.16, subadditivity, and $\mathbb{E}[P_T^r] \leq (\mathbb{E}[P_T])^r$ therefore imply

$$\mathbb{E}[S_T^r] \leq r^{-r}\left[(b+N_\star)^r + (4k)^r X + (4c_v Q)^r\right] = \mathcal{L}_b(Q) + \mathcal{J}(k)X. \tag{38}$$

Combine the two inequalities to obtain

$$X \leq (W/a)^r\left(\mathcal{L}_b(Q) + \mathcal{J}(k)X\right)^p.$$

Lemma D.7, with $D = W/a$, proves

$$X \leq \max\{(W/a)^r(2\mathcal{L}_b(Q))^p, (W/a)(2\mathcal{J}(k))^{p/r}\}.$$

Taking the power $1/r$ gives (37). The exact term $b + N_\star$ is retained, rather than being replaced by $2b$ without a budget comparison. This is the change that permits arbitrary $b > 0$. □

F Proof of Section 3

G Proof of Theorem 3.1 and Corollary 3.2

Proof of Theorem 3.1. For the convex case, define

$$\delta_n = f(w_n) - f^*, \quad \delta_0 = \Delta_0, \quad \Delta_T = \sum_{n=1}^{T} \delta_n, \quad W_T^{\mathrm{C}} = \mathbb{E}\left[\sum_{n=1}^{T} \eta_n \delta_n\right].$$

All gaps are nonnegative and predictable.

Step 1: weighted progress for arbitrary h. Apply Lemma B.2 to $\phi = f_{\mu_n}$ and $y = \widetilde{x}^*$. Convex smoothing implies

$$f_{\mu_n}(w_n) - \theta f_{\mu_n}(w_{n-1}) - q f_{\mu_n}(\widetilde{x}^*) \geq \delta_n - \theta \delta_{n-1} - M\mu_n^2/2.$$

The two negative terms cost at most $(\theta + q)M\mu_n^2/2 = M\mu_n^2/2$. The squared-distance update and the conditional mean give

$$\begin{aligned} \mathbb{E}_n\left[\|z_{n+1} - \widetilde{x}^*\|^2\right] \leq\ & \|z_n - \widetilde{x}^*\|^2 - \frac{2\eta_n}{q^2}(\delta_n - \theta\delta_{n-1}) \\ & + \frac{M\eta_n\mu_n^2}{q^2} + \frac{\eta_n^2}{q^2}\mathbb{E}_n[\|g_n\|^2]. \end{aligned} \tag{39}$$

By nonincreasing steps and nonnegative gaps,

$$\begin{aligned} \sum_{n=1}^{T} \eta_n(\delta_n - \theta\delta_{n-1}) &= \sum_{n=1}^{T-1}(\eta_n - \theta\eta_{n+1})\delta_n + \eta_T\delta_T - \theta h\Delta_0 \\ &\geq q\sum_{n=1}^{T}\eta_n\delta_n - \theta h\Delta_0. \end{aligned} \tag{40}$$

Sum (39), drop the terminal squared distance, use $z_1 = w_1$, and use $\sum \mu_n^2 \leq 2\mu_1^2 \leq 2$. This gives

$$W_T^{\mathrm{C}} \leq \frac{qR_0^2}{2} + \frac{Mh}{q} + \frac{\theta h\Delta_0}{q} + \frac{Q_T}{2q} = A_1(h) + \frac{Q_T}{2q}. \tag{41}$$

The initialization term is retained. No condition on κ_{C} or κ_{N} has yet been used.

Step 2a: the direct convex energy certificate. Since $x_n - w_n = \theta v_n$, smoothness and self-bounding imply

$$\|\nabla f(x_n)\|^2 \leq 2\|\nabla f(w_n)\|^2 + 2M^2\theta^2\|v_n\|^2 \leq 4M\delta_n + 2M^2\theta^2\|v_n\|^2. \tag{42}$$

Weighting and using Lemma B.1 gives

$$G_T \leq 4MW_T^{\mathrm{C}} + c_v h Q_T.$$

Together with Lemma E.1 and (41),

$$Q_T \leq 16MhA_1(h) + \kappa_{\mathrm{C}}(h)Q_T + h^2N_\star.$$

When $\kappa_{\mathrm{C}}(h) < 1$, all terms are already finite, so division by $1 - \kappa_{\mathrm{C}}(h)$ proves $Q_T \leq Q_{\mathrm{C}}^{\mathrm{dir}}(h)$.

Step 2b: a second certificate valid also for convex objectives. Convexity implies $f^* > -\infty$ under the present minimizer assumption, so the independently proved Lemma E.2 applies. Lemmas E.1–E.2 give

$$Q_T \leq 4h[A_2(h) + B_2(h)Q_T] + h^2N_\star = 4hA_2(h) + \kappa_{\mathrm{N}}(h)Q_T + h^2N_\star.$$

When $\kappa_{\mathrm{N}}(h) < 1$, division by its positive margin gives $Q_T \leq Q_{\mathrm{N}}(h)$.

Because $h < h_{\mathrm{C}}$, at least one certificate is available. Taking the smaller of the available bounds gives

$$Q_T \leq Q_1(h), \qquad W_T^{\mathrm{C}} \leq W_1(h). \tag{43}$$

Step 3: unweighted output Summing (42) without weights gives

$$\sum_{n=1}^{T} \|\nabla f(x_n)\|^2 \le 4M\Delta_T + c_v P_T.$$

Thus Lemma E.4 applies with $D_T = \Delta_T$, $k = 4M$, $Q = Q_1(h)$, and $W = W_1(h)$. It yields

$$\mathbb{E}[(\Delta_T)^r] \le (K_1)^r.$$

Finally, convexity gives $f(\bar{w}_T) - f^* \le \Delta_T/T$. Take the rth power and expectation to complete the proof of Theorem 3.1. □

Proof of Corollary 3.2. Theorem 3.1 gives

$$\left(\mathbb{E}[(f(\bar{w}_T) - f^*)^r]\right)^{1/r} \le K_1/T \le \delta.$$

The choice of T is not circular: K_1 is independent of m, T, σ. For $0 < \delta \le 1$,

$$T \le 2 + \frac{K_1}{\delta} \le \frac{2 + K_1}{\delta}.$$

The ceiling in the batch size satisfies

$$m \le 1 + d\max\{1, 2\sigma^2 T\} \le 2d + 2d\sigma^2 T.$$

Every direction uses its own environmental sample and its own baseline, so

$$N_{\mathrm{C}} = 2mT \le 4dT + 4d\sigma^2 T^2 = \mathcal{O}(d/\delta + d\sigma^2/\delta^2).$$

This completes the proof. □

H Proof of Theorem 3.3 and Corollary 3.4

Proof of Theorem 3.3. For the nonconvex case, define

$$\widetilde{\Delta}_T = \sum_{n=1}^{T} \|\nabla f(w_n)\|^2, \qquad W_T^{\mathrm{NC}} = \mathbb{E}\left[\sum_{n=1}^{T} \eta_n \|\nabla f(w_n)\|^2\right].$$

Step 1: close the update energy. Lemma E.2 gives $G_T \le A_2(h) + B_2(h)Q_T$. Substitution into Lemma E.1 yields

$$[1 - \kappa_{\mathrm{N}}(h)]Q_T \le 4hA_2(h) + h^2 N_\star.$$

Since $h < h_{\mathrm{N}}$, this margin is strictly positive. Hence

$$Q_T \le Q_2(h) = Q_{\mathrm{N}}(h), \qquad G_T \le A_2(h) + B_2(h)Q_2(h). \tag{44}$$

No convexity is invoked.

Step 2: transfer weighted progress to the reported iterates. Smoothness in the direction from x_n to w_n gives

$$\|\nabla f(w_n)\|^2 \le 2\|\nabla f(x_n)\|^2 + 2M^2\theta^2\|v_n\|^2.$$

After multiplication by η_n, summation, and expectation,

$$W_T^{\mathrm{NC}} \le 2G_T + c_v h Q_T \le W_2(h).$$

For the reverse comparison needed by moment closure, apply smoothness afresh:

$$\|\nabla f(x_n)\|^2 \le 2\|\nabla f(w_n)\|^2 + 2M^2\theta^2\|v_n\|^2.$$

Therefore $\sum_n \|\nabla f(x_n)\|^2 \le 2\widetilde{\Delta}_T + c_v P_T$ pathwise. Lemma E.4 with $k = 2$, $Q = Q_2(h)$, $W = W_2(h)$ gives

$$\mathbb{E}[(\widetilde{\Delta}_T)^r] \le (K_2)^r.$$

Dividing by T^r proves $\frac{\mathbb{E}[(\widetilde{\Delta}_T)^r]}{T^r} \le \frac{(K_2)^r}{T^r}$.

Step 3: random and best-iterate outputs. Condition on the entire trajectory and draw independent uniform R. Concavity of s^r implies

$$\mathbb{E}[\|\nabla f(w_R)\|^{2r} \mid w_1, \dots, w_T] = \frac{1}{T}\sum_{n=1}^{T}(\|\nabla f(w_n)\|^2)^r \le \left(\frac{\widetilde{\Delta}_T}{T}\right)^r.$$

Take expectations. The minimum of the powers is no greater than their average, proving the second inequality

$$\mathbb{E}[\min_{1\le n\le T}\|\nabla f(w_n)\|^{2r}] \le \left(\frac{K_2}{T}\right)^r$$

as well. Uniform random output requires no gradient evaluations. This completes Theorem 3.3. □

Proof of Corollary 3.4. Theorem 3.3 gives $K_2/T \le \delta^2$, so

$$\mathbb{E}[\min_{1\le n\le T}\|\nabla f(w_n)\|^{2r}] \le \mathbb{E}[\|\nabla f(w_R)\|^{2r}] \le \delta^{2r}.$$

Since K_2 is independent of T and m, this tuning is not circular. Moreover,

$$T \le (2+K_2)/\delta^2, \qquad m \le 2d + 2d\sigma^2 T.$$

Consequently

$$N_{\mathrm{NC}} = 2mT \le 4dT + 4d\sigma^2T^2 = \mathcal{O}(d/\delta^2 + d\sigma^2/\delta^4).$$

This completes the proof. □

I Proof of Theorem 3.6

Proof of Theorem 3.6. Write $H := dL$ and recall $h = a/b^p$.

Step 1: pathwise finite squared-update energy. By Lemma E.3, $Y_n := \|g_n\|^2 \le H^2$. For every finite T,

$$\begin{aligned}\sum_{n=1}^{T}\eta_n^2 Y_n &= \sum_{n=1}^{T}\eta_{n+1}^2 Y_n + \sum_{n=1}^{T}(\eta_n^2 - \eta_{n+1}^2)Y_n \\ &\le \frac{a^2}{2\epsilon b^{2\epsilon}} + H^2(\eta_1^2 - \eta_{T+1}^2) \\ &\le \frac{a^2}{2\epsilon b^{2\epsilon}} + H^2h^2 =: \overline{Q} < \infty.\end{aligned}$$

The first sum is bounded by Lemma D.3 with exponent $2p = 1+2\epsilon > 1$. The second uses nonnegative stepsize differences and telescopes. Taking monotone limits and applying Lemma B.1 yields

$$\sum_{n\ge1}\eta_n^2\|g_n\|^2 \le \overline{Q}, \qquad \sum_{n\ge1}\|v_n\|^2 \le \frac{\overline{Q}}{q^2}, \qquad \sum_{n\ge1}\eta_n\|v_n\|^2 \le \frac{h\overline{Q}}{q^2}. \tag{45}$$

In particular $v_n \to 0$. These are bounds on update energy, not an assertion that $\sum_n \eta_n^2$ is always finite.

Step 2: weighted exact-gradient summability. Lemma D.4 applies by bounded differentiation under Assumption 3. At every finite horizon, bounded estimators and $\eta_n \le h$ even give deterministic bounds on the iterates, so all conditional expectations in the descent calculation are justified. Use the general-radius form (32), before imposing $\mu_1 \le 1$:

$$\mathbb{E}\left[\sum_{n=1}^{T}\eta_n\|\nabla f(x_n)\|^2\right] \le 2q\Delta_0 + 2M^2h\mu_1^2 + B_2(h)\overline{Q}. \tag{46}$$

The right-hand side is finite for *every* $a, b > 0$ and independent of T. There is no absorption step and no division by $1-\kappa_{\mathrm{N}}(h)$ or any potentially nonpositive quantity.

Also,

$$\|\nabla f(z_n)\|^2 \le 2\|\nabla f(x_n)\|^2 + \frac{2M^2\theta^4}{q^2}\|v_n\|^2.$$

Using (45), expected partial sums of $\sum_n \eta_n \|\nabla f(z_n)\|^2$ are uniformly bounded. Monotone convergence and Lemma D.11 give

$$\sum_{n=1}^{\infty} \eta_n \|\nabla f(z_n)\|^2 < \infty \quad \text{almost surely.} \tag{47}$$

Step 3: divergence of total steps and the last-iterate conclusion. For every path,

$$\eta_n \ge \frac{a}{(b+(n-1)H^2)^p}.$$

Since $p \le 1$, the sum of these lower bounds diverges. If $H = 0$, all sample and population gradients vanish and the conclusion is immediate. Otherwise $\sum_n \eta_n = \infty$.

Assumption 3 implies $\|\nabla f(x)\| \le L$, so smoothness and (6) imply

$$\begin{aligned}\left|\|\nabla f(z_{n+1})\|^2 - \|\nabla f(z_n)\|^2\right| &\le 2LM\|z_{n+1} - z_n\| \\ &\le \frac{2LMH}{q}\eta_n.\end{aligned}$$

On the probability-one event in (47), apply Lemma D.2 with $a_n = \eta_n$ and $s_n = \|\nabla f(z_n)\|^2$. Hence $\|\nabla f(z_n)\| \to 0$. Finally,

$$\|\nabla f(w_n)\| \le \|\nabla f(z_n)\| + \frac{M\theta}{q}\|v_n\| \to 0,$$

and the corresponding comparison with x_n uses $M\theta^2\|v_n\|/q$. This proves Theorem 3.6. □

J Proof of Theorem 3.8

Before presenting the formal proof, we state the following noise decomposition for $g_n := \widehat{\nabla} f_{U_n}^{\mu_n}(x_n)$ and get the corresponding properties via Lemma D.14 and direct computation:

$$\begin{aligned} &g_n = \nabla f(x_n) + e_n + \zeta_n, \qquad e_n := \nabla f_{\mu_n}(x_n) - \nabla f(x_n), \\ &\zeta_n := g_n - \nabla f_{\mu_n}(x_n), \qquad \mathbb{E}_n[\zeta_n] = 0, \qquad \|e_n\| \le M\mu_n. \end{aligned} \tag{48}$$

J.1 The adaptive clock and the momentum error

Lemma J.1. *Under Assumptions 1, 3, and 4 with fixed a, b, m and $p \in [1/2, 1]$, set $H = dL$ and $\gamma_n = n^{-p}$. Then $H > 0$, and there are deterministic positive constants*

$$c_\eta = \frac{a}{(b+H^2)^p}, \qquad C_\eta = a\left(\frac{4}{\chi_m^2}\right)^p \tag{49}$$

such that, almost surely, for all sufficiently large n,

$$c_\eta\gamma_n \le \eta_n \le C_\eta\gamma_n. \tag{50}$$

In addition, $\eta_{n+1}/\eta_n \to 1$ and $\sum_n \eta_n = \infty$ almost surely. The squared steps are summable for $p > 1/2$ and nonsummable for $p = 1/2$. The random index after which (50) holds is not treated as a deterministic index in later conditional expectations.

Proof. Recall the algorithm's accumulator $S_n = b + \sum_{i=1}^n \|g_i\|^2$; S_n is not a repulsion energy. Since $\mu_n \to 0$, Assumption 4 applies after a deterministic index. For any fixed unit vector u, conditional Jensen gives

$$\mathbb{E}_n[\|\zeta_n\|^2] \ge \mathbb{E}_n\left[(\langle\zeta_n, u\rangle^+)^2\right] \ge \left(\mathbb{E}_n[\langle\zeta_n, u\rangle^+]\right)^2 \ge \chi_m^2.$$

The exact conditional centering in (48) yields

$$\mathbb{E}_n[\|g_n\|^2] = \|\nabla f_{\mu_n}(x_n)\|^2 + \mathbb{E}_n[\|\zeta_n\|^2] \geq \chi_m^2. \tag{51}$$

Thus excitation rules out $H = 0$. Let

$$D_n^{\mathrm{clk}} = \|g_n\|^2 - \mathbb{E}_n[\|g_n\|^2].$$

These are martingale differences with $|D_n^{\mathrm{clk}}| \leq H^2$. For each $t > 0$, the exponential martingale bound gives

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} D_i^{\mathrm{clk}}\right| \geq tn\right) \leq 2\exp\left(-\frac{t^2 n}{2H^4}\right).$$

For completeness, if $|D| \leq H^2$ and $\mathbb{E}_n D = 0$, convexity of the exponential on $[-H^2, H^2]$ gives $\mathbb{E}_n[e^{sD}] \leq \cosh(sH^2) \leq e^{s^2 H^4/2}$. The latter inequality follows by integrating $\tanh t \leq t$ for $t \geq 0$ and using evenness. Iterate this conditional estimate and optimize the exponential Markov bound at $s = t/H^4$. Applying Borel–Cantelli for positive rational t proves $n^{-1}\sum_{i=1}^{n} D_i^{\mathrm{clk}} \to 0$ almost surely. Together with (51), it follows that

$$\liminf_{n\to\infty} \frac{S_n}{n} \geq \chi_m^2, \qquad S_n \leq b + nH^2.$$

In particular $S_n \geq \chi_m^2 n/2$ eventually. For $n \geq 2$, substituting these bounds into $\eta_n = a/S_{n-1}^p$ gives (50) with (49). The lower bound in fact holds at every $n \geq 1$. Also,

$$\frac{\eta_{n+1}}{\eta_n} = \left(1 + \frac{\|g_n\|^2}{S_{n-1}}\right)^{-p} \longrightarrow 1.$$

The series conclusions follow from (50) and $1/2 \leq p \leq 1$. □

Lemma J.2. *Under Lemma J.1, define*

$$C_v = 1 + \frac{3^p C_\eta H}{q}, \qquad B_{\mathrm{r}} = \frac{M\theta^2 C_v}{q} + M\mu_1, \qquad C_{\mathrm{r}} = \frac{qB_{\mathrm{r}}}{c_\eta}. \tag{52}$$

Almost surely, eventually $\|v_n\| \leq C_v\gamma_n$. With $\alpha_n = \eta_n/q$, the exact recursion is

$$z_{n+1} = z_n + \alpha_n\big(-\nabla f(z_n) - \zeta_n + \mathfrak{r}_n\big), \qquad \mathfrak{r}_n = \nabla f(z_n) - \nabla f(x_n) - e_n, \tag{53}$$

where $\mathfrak{r}_n$ is predictable and, eventually,

$$\|\mathfrak{r}_n\| \leq B_{\mathrm{r}}\gamma_n \leq C_{\mathrm{r}}\alpha_n, \qquad \|z_{n+1} - z_n\| \leq H\alpha_n. \tag{54}$$

Thus the actual drift error in the update has order α_n^2, even at $p = 1$.

Proof. The correct momentum recursion gives

$$v_n = -\sum_{i<n} \theta^{n-1-i}\eta_i g_i.$$

Put $k = \lfloor n/2 \rfloor$ and $h = a/b^p$. The contribution from $i < k$ is bounded by $hH\theta^{n-k}/q$, which is at most γ_n for all sufficiently large deterministic n. Once k exceeds the random index in Lemma J.1, the contribution from $i \geq k$ is bounded, using $k \geq n/3$, by $3^p C_\eta H\gamma_n/q$. This proves the claimed eventual bound with a deterministic C_v. In particular $v_n \to 0$.

Substitute (48) into $z_{n+1} = z_n - \eta_n g_n/q$ to obtain (53). Since $z_n - x_n = \theta^2 v_n/q$,

$$\|\mathfrak{r}_n\| \leq \frac{M\theta^2}{q}\|v_n\| + M\mu_n.$$

For $p \leq 1$, $\mu_n = \mu_1/n \leq \mu_1\gamma_n$. This proves the first bound in (54); the second comparison uses $\alpha_n \geq c_\eta\gamma_n/q$. Finally, the exact update and $\|g_n\| \leq H$ imply its increment bound. No division by $1 - 2\theta^2$ or lower bound on b was used. □

J.2 A LOCAL REPULSION ENERGY FOR GRADIENT DYNAMICS

Lemma J.3. *For compact strict saddle manifold $\mathcal{S}$, there is a C^2 center-stable manifold Γ near $\mathcal{S}$, with $\mathcal{S} \subset \Gamma$, and an open neighborhood $\mathcal{O}$ of $\mathcal{S}$ on which the orthogonal tubular projection Π onto Γ is defined. For a sufficiently small compact neighborhood $\mathcal{U}$ of $\mathcal{S}$ with $\mathcal{U} \subset \mathcal{O}$, the function*

$$\mathcal{V}(z) = \|z - \Pi(z)\| = \operatorname{dist}(z, \Gamma) \tag{55}$$

has the following properties, with uniform constants $\lambda > 0$, $C_{\mathcal{V}} < \infty$, and $\rho > 0$:

$$D^+\mathcal{V}(z)[-\nabla f(z)] \geq \lambda \mathcal{V}(z), \tag{56}$$

$$\mathcal{V}(z+t) \geq \mathcal{V}(z) + D^+\mathcal{V}(z)[t] - C_{\mathcal{V}} \|t\|^2 \quad (\|t\| \leq \rho), \tag{57}$$

$$|\mathcal{V}(z') - \mathcal{V}(z)| \leq \|z' - z\| \quad \textit{for the nearby points used below}. \tag{58}$$

Here D^+ denotes the one-sided directional derivative. Off Γ, $\|\nabla \mathcal{V}(z)\| = 1$. On Γ,

$$D^+\mathcal{V}(z)[t] = \|P_{N_z\Gamma} t\|, \tag{59}$$

where $P_{N_z\Gamma}$ is orthogonal projection onto the nonzero normal space. At all such z, $D^+\mathcal{V}(z)$ is convex, positively homogeneous, and 1-Lipschitz as a function of its direction argument. On Γ it is invariant under adding a tangent vector.

Proof. **Classical input and its hypotheses.** The classical existence input, specialized to the present setting, is the following: let a C^2 vector field B vanish on a compact smooth equilibrium submanifold $\mathcal{S}$. Suppose there is a constant-rank invariant splitting

$$T_{\mathcal{S}}\mathbb{R}^d = E^{\mathrm{cs}} \oplus E^{\mathrm{u}}, \qquad T_s\mathcal{S} \subset E^{\mathrm{cs}}_s, \qquad \dim E^{\mathrm{u}}_s \geq 1,$$

with nonexpanding forward linearized dynamics on E^{cs} and uniformly expanding forward dynamics on E^{u}. Then there is a local C^2 center-stable submanifold Γ containing $\mathcal{S}$, tangent to E^{cs} on $\mathcal{S}$, and locally invariant under B. At a single equilibrium one may also choose its local center-stable manifold to contain all forward orbits that remain in a sufficiently small neighborhood. Only this deterministic existence and containment result is imported.

Precise provenance. The strict-saddle-manifold formulation and local invariance are explicitly recorded in Liu and Yuan (Liu & Yuan, 2024, Appendix E.2, p. 34, paragraph preceding Proposition 24); footnote 6 explains the C^3 requirement on f. The corresponding center-stable construction and local-orbit containment are also recorded in (Mertikopoulos et al., 2020, Appendix C.2). These passages invoke the classical center-manifold theorem. We do *not* identify either their Lyapunov proposition or Benaïm's Proposition 9.5 with a proof of manifold existence. Nor do we invoke the stochastic convergence theorems in those papers. This proof is complete relative to this standard, explicitly stated geometric theorem; it does not claim to reprove the general invariant-manifold existence theory.

Verification of the hypotheses. Here $B = -\nabla f$ is C^2 near $\mathcal{S}$ and $DB(s) = -\nabla^2 f(s)$ is symmetric. Spectral separation and connectedness give a constant positive unstable rank; the spectral projectors are continuous, so the associated bundles are continuous. The complementary eigenspace is E^{cs}_s, and differentiating $\nabla f = 0$ along $\mathcal{S}$ gives $\nabla^2 f(s) T_s\mathcal{S} = 0$. If Φ_t is the local flow, the variational equation at the equilibrium s gives

$$D\Phi_t(s) = \exp(tDB(s)), \qquad \|D\Phi_t(s)|_{E^{\mathrm{cs}}_s}\| \leq 1, \qquad \|D\Phi_{-t}(s)|_{E^{\mathrm{u}}_s}\| \leq e^{-c-t} \quad (t \geq 0). \tag{60}$$

The matrix-exponential identity is used *only at equilibria*, where the variational coefficient is constant. Compactness makes the neighborhoods and linearized estimates uniform. These conditions provide the uniform separation for the C^2 center-stable theorem even when there are extra center directions normal to $\mathcal{S}$; no full normal hyperbolicity or Morse–Bott condition is used. Applying the theorem gives

$$T_s\Gamma = E^{\mathrm{cs}}_s, \qquad N_s\Gamma = E^{\mathrm{u}}_s, \qquad B(y) \in T_y\Gamma \quad (y \in \Gamma \text{ near } \mathcal{S}).$$

All uses of this classical existence theorem are thus under its differentiability, compactness, and spectral hypotheses. The scalar energy and all estimates needed for our discrete algorithm are proved next, rather than imported from a stochastic-approximation theorem.

Tubular distance and directional derivatives. The normal-bundle inverse function theorem, followed by restriction near the compact set $\mathcal{S}$, provides the orthogonal projection Π on a tube $\mathcal{O}$. On a

slightly smaller tube its derivative is bounded. Put $y = \Pi(z)$ and $r_z = z - y \in N_y\Gamma$. When $r_z \neq 0$, differentiation gives

$$\nabla\mathcal{V}(z) = \frac{r_z}{\|r_z\|},$$

since $D\Pi(z)$ takes values in $T_y\Gamma$. At $z \in \Gamma$, $D\Pi(z) = P_{T_z\Gamma}$, which gives (59). These formulas imply the stated convexity, homogeneity, and Lipschitz properties in the direction variable, as well as tangent-translation invariance. The distance to a set is 1-Lipschitz; all points here lie in the tube where this local projection gives that distance.

Positive drift, using symmetry of the Hessian. At $s \in \mathcal{S}$ the quadratic form of $DB(s)$ on $N_s\Gamma = E_s^{\mathrm{u}}$ is at least c_-. By continuity of DB, continuity of the normal spaces, and compactness, shrink the tube so that

$$\langle u, DB(y + tr_z)u\rangle \geq c_-/2$$

for every unit $u \in N_y\Gamma$, $0 \leq t \leq 1$, and z under consideration. Since $B(y)$ is tangent to Γ, for $z \notin \Gamma$ and $u = r_z/\|r_z\|$,

$$\begin{aligned} D^+\mathcal{V}(z)[B(z)] &= \langle u, B(z) - B(y)\rangle \\ &= \|r_z\| \int_0^1 \langle u, DB(y + tr_z)u\rangle\, dt \geq (c_-/2)\mathcal{V}(z). \end{aligned}$$

On Γ both sides are zero. Thus (56) holds with $\lambda = c_-/2$.

Uniform lower quadratic remainder. Bounded second derivatives of local graph representations of the C^2 manifold on a compact region give constants C_Γ, C_Π such that, for sufficiently close $y, y' \in \Gamma$ in that region and unit $u \in N_y\Gamma$,

$$|\langle u, y' - y\rangle| \leq C_\Gamma \|y' - y\|^2, \qquad \|\Pi(z + t) - \Pi(z)\| \leq C_\Pi \|t\|.$$

The first inequality is Taylor's formula in a graph chart, whose first derivative is tangent at y; a finite cover gives a common constant. The second follows by integrating the bounded derivative of Π along the short segment. If $z \notin \Gamma$, take $u = (z - y)/\|z - y\|$ and $y' = \Pi(z + t)$. Then

$$\begin{aligned} \mathcal{V}(z + t) &\geq \langle u, z + t - y'\rangle \\ &\geq \mathcal{V}(z) + \langle u, t\rangle - C_\Gamma C_\Pi^2 \|t\|^2. \end{aligned}$$

If $z \in \Gamma$ and $P_{N_z\Gamma}t \neq 0$, take $u = P_{N_z\Gamma}t/\|P_{N_z\Gamma}t\|$ and repeat the same calculation. If the projection is zero, nonnegativity of $\mathcal{V}$ gives the inequality directly. This proves (57) uniformly, with $C_\mathcal{V} = C_\Gamma C_\Pi^2$. Choose $\mathcal{U}$ compactly inside this tube and ρ small enough that all required projected points remain in its larger domain. □

J.3 A SELF-CONTAINED DETERMINISTIC-SCALE NONCONVERGENCE LEMMA

Lemma J.4. *Let $1/2 \leq p \leq 1$, $\gamma_n = n^{-p}$, and let a nonnegative square-integrable process $\{Z_n\}$ be adapted to $\{\mathcal{H}_n\}$. Write $\Delta Z_n = Z_{n+1} - Z_n$. Suppose that there are deterministic constants $C_1, c_1 > 0$, $C_0 \geq 0$, and a deterministic N_0 such that, for all $n \geq N_0$, almost surely,*

$$|\Delta Z_n| \leq C_1\gamma_n, \tag{61}$$

$$\mathbf{1}_{\{Z_n > C_0\gamma_n\}}\mathbb{E}[\Delta Z_n \mid \mathcal{H}_n] \geq 0, \tag{62}$$

$$\mathbb{E}[Z_{n+1}^2 - Z_n^2 \mid \mathcal{H}_n] \geq c_1\gamma_n^2. \tag{63}$$

Then $\mathbb{P}(Z_n \to 0) = 0$. In particular, the scale in this lemma is deterministic; a random future sum is not part of its hypotheses.

Proof. **The endpoint $p = 1/2$.** Here $\sum_n \gamma_n^2 = \infty$. Fix a deterministic $N \geq N_0$ and condition on the event $\{Z_N \leq 1\}$. Let $\tau = \inf\{j \geq N : Z_j > 1\}$. The stopped process satisfies

$$Z_{m\wedge\tau}^2 \leq (1 + C_1\gamma_N)^2.$$

Summing (63) up to the stopped horizon yields

$$c_1 \sum_{j=N}^{m-1} \gamma_j^2\, \mathbb{P}(\tau > m \mid \mathcal{H}_N) \leq (1 + C_1\gamma_N)^2$$

on $\{Z_N \le 1\}$. Because the sum diverges, $\mathbb{P}(\tau = \infty \mid \mathcal{H}_N) = 0$ on this event. Thus the event $\{Z_j \le 1 \text{ for all } j \ge N\}$ has probability zero for every N. Its countable union contains $\{Z_j \to 0\}$, proving the endpoint. Only (61) and (63) were needed in this case.

The summable-square case $1/2 < p \le 1$**.** Define the deterministic tail

$$\mathcal{B}_n = \sum_{j=n}^{\infty} \gamma_j^2.$$

Integral comparison gives $\mathcal{B}_n = \Theta(n^{1-2p})$, and hence $\gamma_n/\sqrt{\mathcal{B}_n} \to 0$. Choose $\beta = c_1/16$ and $\kappa = \sqrt{c_1}/4$, so that

$$r_0 := \frac{2(\beta + \kappa^2)}{c_1} = \frac{1}{4} < 1.$$

Increase N_0 so that, for $n \ge N_0$,

$$C_1\gamma_n \le \kappa\sqrt{\mathcal{B}_n}, \qquad C_0\gamma_n \le \tfrac{1}{4}\sqrt{\beta\mathcal{B}_n}.$$

First stage: reach the tail-fluctuation scale. For a fixed $n \ge N_0$ let

$$T_n = \inf\{j \ge n : Z_j \ge \sqrt{\beta\mathcal{B}_j}\}.$$

If the threshold is already reached, $T_n = n$. On the complementary $\mathcal{H}_n$-measurable event, all stopped values satisfy

$$Z^2_{m\wedge T_n} \le 2(\beta + \kappa^2)\mathcal{B}_n \quad (m \ge n).$$

Indeed, before crossing use the threshold, and at crossing add at most $C_1\gamma_{T_n-1}$ to the preceding value; $\mathcal{B}_j$ is decreasing. Summing (63) for the stopped process gives

$$c_1\left(\sum_{j=n}^{m-1}\gamma_j^2\right)\mathbb{P}(T_n > m \mid \mathcal{H}_n) \le \mathbb{E}[Z^2_{m\wedge T_n} - Z_n^2 \mid \mathcal{H}_n] \le 2(\beta+\kappa^2)\mathcal{B}_n.$$

Letting $m \to \infty$ shows, in both initial cases,

$$\mathbb{P}(T_n < \infty \mid \mathcal{H}_n) \ge 1 - r_0. \tag{64}$$

Second stage: a positive chance never to return near zero. Fix $k \ge N_0$ and condition on $Z_k \ge \sqrt{\beta\mathcal{B}_k}$. Put

$$s = \tfrac{1}{2}\sqrt{\beta\mathcal{B}_k}, \qquad \tau = \inf\{j \ge k : Z_j < s\}.$$

Before τ, $Z_j \ge s > C_0\gamma_j$, so (62) makes the stopped process a submartingale. Its Doob decomposition can be written

$$Z_{j\wedge\tau} = Z_k + M_j + A_j, \qquad M_k = A_k = 0,$$

where, writing $b_i = \mathbb{E}[\Delta Z_i \mid \mathcal{H}_i]$,

$$A_j = \sum_{i=k}^{j-1}\mathbf{1}_{\{i<\tau\}}b_i, \qquad M_j = \sum_{i=k}^{j-1}\mathbf{1}_{\{i<\tau\}}(\Delta Z_i - b_i).$$

The indicators are predictable, A_j is nondecreasing, and M_j is a martingale. These assertions are made on the $\mathcal{H}_k$-measurable starting event under discussion; equivalently, restrict the conditional measure to that event. The conditional variance of each martingale increment is at most $C_1^2\gamma_i^2$, not a bound obtained by squaring its crude absolute bound. Thus by (61) and conditional orthogonality,

$$\mathbb{E}[M_j^2 \mid \mathcal{H}_k] \le C_1^2\mathcal{B}_k =: v.$$

For any $t > 0$, apply the maximal inequality to the nonnegative submartingale $(M_j - t)^2$, conditionally on $\mathcal{H}_k$. For every finite horizon m this gives

$$\mathbb{P}\left(\min_{k\le j\le m} M_j < -s \mid \mathcal{H}_k\right) \le \frac{v + t^2}{(s+t)^2}.$$

Taking $t = v/s$ and then $m \to \infty$ bounds this probability by $v/(v + s^2)$. A crossing of Z_j below s would force $M_j < -s$, since $Z_k \geq 2s$ and $A_j \geq 0$. Therefore

$$\mathbb{P}(\tau = \infty \mid \mathcal{H}_k) \geq \frac{s^2}{v + s^2} = \frac{\beta}{4C_1^2 + \beta} =: \pi_0 > 0 \tag{65}$$

on the indicated starting event. On $\{\tau = \infty\}$ the process does not converge to zero.

Conclude by conditional probabilities. Let $G = \{Z_j \text{ does not converge to } 0\}$. Apply (65) at each possible value k of T_n, use the tower property, and sum over k. Equation (64) yields

$$\mathbb{E}[\mathbf{1}_G \mid \mathcal{H}_n] \geq \pi_0(1 - r_0) > 0 \quad (n \geq N_0).$$

The bounded martingale on the left converges to $\mathbf{1}_G$, because $G \in \sigma(\cup_n \mathcal{H}_n)$. Hence $\mathbf{1}_G \geq \pi_0(1 - r_0) > 0$ almost surely, so $\mathbb{P}(G) = 1$. This proves the lemma without conditioning on a future adaptive denominator. □

J.4 Verification of the local conditional inequalities

Let $\mathcal{U}, \mathcal{O}, \mathcal{V}$ be from Lemma J.3 and set

$$a_- := c_\eta/q, \qquad a_+ := C_\eta/q, \qquad C_* := C_{\mathrm{r}} + C_{\mathcal{V}} H^2.$$

Consider an index at which the following predictable conditions hold:

$$z_n \in \mathcal{U}, \qquad c_\eta \gamma_n \leq \eta_n \leq C_\eta \gamma_n, \qquad \|v_n\| \leq C_v \gamma_n. \tag{66}$$

Then $a_- \gamma_n \leq \alpha_n \leq a_+ \gamma_n$, and the algebraic bounds in Lemma J.2 give $\|\mathfrak{r}_n\| \leq C_{\mathrm{r}} \alpha_n$. After a sufficiently large *deterministic* index, $H\alpha_n \leq H a_+ \gamma_n$ is smaller than both the margin from $\mathcal{U}$ to $\mathcal{O}^c$ and the radius ρ in (57). All energy evaluations below are then in their domain.

Let $\Delta\mathcal{V}_n = \mathcal{V}(z_{n+1}) - \mathcal{V}(z_n)$. The distance Taylor bound and (53) imply

$$\Delta\mathcal{V}_n \geq \lambda \alpha_n \mathcal{V}(z_n) + \alpha_n D^+\mathcal{V}(z_n)[-\zeta_n] - C_* \alpha_n^2. \tag{67}$$

Here is the justification that avoids the false splitting of a nonlinear derivative. Off Γ the derivative is linear, so split the gradient drift and the noise, and use its norm-one bound for $\mathfrak{r}_n$. On Γ, $-\nabla f(z_n)$ is tangent and $\mathcal{V}(z_n) = 0$; the tangent-translation identity in (59) removes that drift exactly. The directional Lipschitz bound then removes $\mathfrak{r}_n$ at cost at most $\|\mathfrak{r}_n\|$. Both cases give (67).

Because the right derivative is convex and $\mathbb{E}_n[\zeta_n] = 0$, conditional Jensen gives

$$\mathbb{E}_n\left[D^+\mathcal{V}(z_n)[-\zeta_n]\right] \geq 0.$$

Moreover,

$$\mathbb{E}_n\left[\left(D^+\mathcal{V}(z_n)[-\zeta_n]\right)^+\right] \geq \chi_m. \tag{68}$$

Off Γ, use the unit vector $-\nabla\mathcal{V}(z_n)$ in Assumption 4. On Γ, choose any unit normal u_n: the derivative is $\left\|P_{N_{z_n}\Gamma}\zeta_n\right\|$, which dominates $\langle \zeta_n, u_n\rangle^+$. A predictable unit normal can be chosen without a global frame: take the first coordinate vector e_j whose normal projection is nonzero and normalize $P_{N_{z_n}\Gamma}e_j$. The projection matrix is continuous and the first-nonzero rule is Borel measurable. Continuity in the direction and a countable dense set also extend (4) to predictable unit directions.

Since $\lambda\alpha_n\mathcal{V}(z_n) \geq 0$, equations (67)–(68) give

$$\mathbb{E}_n[\Delta\mathcal{V}_n] \geq \lambda\alpha_n\mathcal{V}(z_n) - C_*\alpha_n^2, \tag{69}$$

$$\mathbb{E}_n[(\Delta\mathcal{V}_n)^+] \geq \chi_m\alpha_n - C_*\alpha_n^2 \geq \tfrac{1}{2}\chi_m\alpha_n, \tag{70}$$

$$|\Delta\mathcal{V}_n| \leq H\alpha_n. \tag{71}$$

The last inequality follows from the Lipschitz property of distance. For sufficiently large deterministic indices, the middle inequality holds uniformly under (66). Jensen then gives $\mathbb{E}_n[(\Delta\mathcal{V}_n)^2] \geq \chi_m^2\alpha_n^2/4$.

If $\mathcal{V}(z_n) > (C_*/\lambda)\alpha_n$, the drift in (69) is nonnegative. Otherwise,

$$2\mathcal{V}(z_n)\mathbb{E}_n[\Delta\mathcal{V}_n] \geq -\frac{2C_*^2}{\lambda}\alpha_n^3.$$

Consequently, increasing the deterministic starting index once more if necessary,

$$\mathbb{E}_n[\mathcal{V}(z_{n+1})^2 - \mathcal{V}(z_n)^2] \geq \frac{\chi_m^2}{4}\alpha_n^2 - \frac{2C_*^2}{\lambda}\alpha_n^3 \geq \frac{\chi_m^2}{8}\alpha_n^2. \tag{72}$$

When $C_* = 0$, the same inequalities hold without the corresponding smallness restrictions. Every index bound used here is deterministic because $\alpha_n \leq a_+\gamma_n$ on the predictable good event.

J.5 STOPPING, EVENTUAL RESIDENCE, AND THE CONCLUSION

Now, we present the formal proof Theorem 3.8 below.

Proof of Theorem 3.8. Fix any deterministic N beyond all the index bounds just specified, including the smoothing threshold in Assumption 4. Let

$$\tau_N = \inf\{n \geq N : \text{one of the conditions in (66) fails}\}.$$

Use the filtration $\mathcal{H}_n = \mathcal{F}_{n-1}$; then τ_N is a stopping time. Define a nonnegative scalar process, for $n \geq N$, by

$$Z_N = \begin{cases} \mathcal{V}(z_N), & \tau_N > N, \\ 1, & \tau_N = N, \end{cases}$$

and

$$Z_{n+1} - Z_n = \mathbf{1}_{\{n<\tau_N\}}[\mathcal{V}(z_{n+1}) - \mathcal{V}(z_n)] + \mathbf{1}_{\{n\geq\tau_N\}}\gamma_n. \tag{73}$$

The expression involving $\mathcal{V}$ is evaluated *only* on $\{n < \tau_N\}$; equivalently, interpret the increment definition piecewise rather than as multiplication by an undefined value outside its domain. On that event $z_n \in \mathcal{U}$ and the buffer guarantees $z_{n+1} \in \mathcal{O}$, including the exit step. Up to the exit the energy telescopes, and after exit positive increments preserve nonnegativity. The initial value is bounded and every increment has a deterministic bound, so Z_n is square-integrable at each finite index.

Equations (69)–(72) verify the hypotheses of Lemma J.4 before exit, with

$$C_1 = \max\{Ha_+, 1\}, \qquad C_0 = \frac{C_* a_+}{\lambda}, \qquad c_1 = \min\{\chi_m^2 a_-^2/8, 1\}.$$

For example, $Z_n > C_0\gamma_n$ implies $Z_n > (C_*/\lambda)\alpha_n$, giving the required nonnegative drift. After exit, $\Delta Z_n = \gamma_n$ and $Z_{n+1}^2 - Z_n^2 = 2Z_n\gamma_n + \gamma_n^2 \geq \gamma_n^2$, so the same scalar hypotheses hold there. Lemma J.4 yields

$$\mathbb{P}(Z_n \to 0) = 0 \quad \text{for every such deterministic } N. \tag{74}$$

On the probability-one event in Lemmas J.1–J.2, all the clock and momentum conditions hold from some index onward, and $v_n \to 0$. If $\operatorname{dist}(w_n, \mathcal{S}) \to 0$, then also $\operatorname{dist}(z_n, \mathcal{S}) \to 0$, since $z_n - w_n = \theta v_n/q$. Compactness of $\mathcal{S}$ implies eventual residence in the neighborhood $\mathcal{U}$. There is therefore an integer N for which $\tau_N = \infty$ and

$$Z_n = \mathcal{V}(z_n) \leq \operatorname{dist}(z_n, \mathcal{S}) \longrightarrow 0.$$

Each such event has probability zero by (74). Taking the countable union over N proves $\mathbb{P}\big(\operatorname{dist}(w_n, \mathcal{S}) \longrightarrow 0\big) = 0$. This completes the proof of Theorem 3.8. □

Table 3: Selected tensor-decomposition hyperparameters.

p	ZO-SGD η	ZO-AdaMM η	R-AdaZO η	ZO-AdaNesterov (a, b, θ)
10	5×10^{-2}	3×10^{-3}	1×10^{-3}	$(0.06, 0.36, 0.75)$
20	5×10^{-2}	2×10^{-3}	5×10^{-4}	$(0.14, 1, 0.70)$
50	5×10^{-2}	2×10^{-3}	5×10^{-4}	$(0.20, 5, 0.80)$

K Additional Experiments and Parameter Analysis

Note that we implemented (ZO-AdaNesterov) on all experiments with $\epsilon = 0$. We state other detailed description for experiments below.

K.1 High-Dimensional Tensor Decomposition

We minimize

$$\min_{u_1,\dots,u_K\in\mathbb{R}^p} \|\mathcal{T} - u_1 \otimes \cdots \otimes u_K\|_F^2 ,$$

where $\mathcal{T}$ is a fixed normalized rank-one target tensor. The distinction between optimization and ambient dimensions is important: concatenating the factors gives Kp trainable scalars, while the reconstructed tensor has p^K entries. We use $K = 3$ and $p \in \{10, 20, 50\}$, corresponding to optimization dimensions 30, 60, and 150 and ambient tensor sizes 10^3, 20^3, and 50^3.

All methods use the same target, initialization, forward spherical estimator, $m = 10$ directions per step, and smoothing schedule $\mu_t = 10^{-4}/t$. Each run lasts 2,000 steps in float64. Table 3 lists the selected parameters; ZO-AdaMM and R-AdaZO use $(\beta_1, \beta_2) = (0.9, 0.999)$.

The five-seed mean final losses of ZO-AdaNesterov are 1.77×10^{-15}, 1.26×10^{-14}, and 5.89×10^{-14} for $p = 10, 20, 50$, respectively, and are the lowest among the compared methods at every width. Its mean log-loss area under the curve is also the best in each case, confirming that the improvement is not limited to the final iteration.

Sensitivity analysis. Figure 2 varies a, b, θ, μ_1, and m at $p = 20$. The coupled initial scale $a/\sqrt{b}$ controls the early plateau; moderate momentum ($\theta \in [0.6, 0.8]$) is robust, while $\theta = 0$ slows convergence and $\theta = 0.9$ increases variability. A smaller smoothing radius reduces finite-difference bias on this deterministic float64 objective. Increasing m reduces direction noise at a fixed step budget but costs additional queries, so $m = 10$ is used as a variance–cost compromise.

K.2 Full-Parameter LLM Fine-Tuning on GLUE

We formulate each GLUE task as a prompt-based objective and optimize the model parameters directly. No task-specific classification head, adapter, or backpropagated gradient is used: the optimizer observes only the scalar loss returned by a forward evaluation. Let $P(x)$ denote the prompted version of an input x, let v_j be the verbalizer for class j, and let $z_j(x; w)$ be its language-model score under parameters w. For each classification task, the optimized mini-batch objective is the verbalizer cross-entropy

$$\mathcal{L}_{\text{cls}}(w) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp z_{y_i}(x_i; w)}{\sum_{j=1}^{C} \exp z_j(x_i; w)}.$$

Thus, the loss compares only the task verbalizers and does not use the remaining vocabulary logits as negative classes. For STS-B, the five verbalizers are assigned the ordered scores $s = (0, 1.25, 2.5, 3.75, 5)$. We form the differentiable prediction

$$\widehat{y}_i(w) = \sum_{j=1}^{5} s_j \ \text{softmax}(z(x_i; w))_j$$

and optimize $\mathcal{L}_{\text{STS-B}}(w) = B^{-1}\sum_i (\widehat{y}_i(w) - y_i)^2$. The primary metrics follow GLUE: the mean of accuracy and F1 for MRPC, accuracy for QNLI, RTE, and SST-2, and the mean of Pearson and Spearman correlations for STS-B.

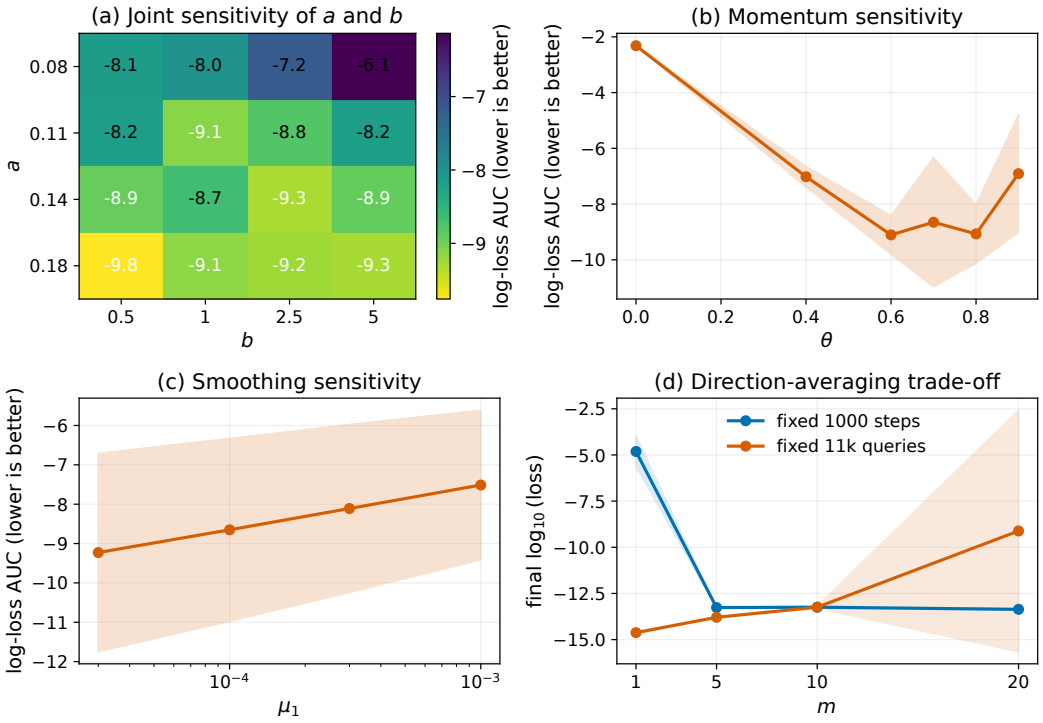

Figure 2: Sensitivity of ZO-AdaNesterov on tensor decomposition with $K = 3$ and $p = 20$. Unless varied, $(a, b, \theta, \mu_1, m) = (0.14, 1, 0.7, 10^{-4}, 10)$.

Table 4: Prompt templates and label verbalizers used for RoBERTa-large.

Task	Prompt template	Label verbalizers
SST-2	⟨sentence⟩ It was `[MASK]`.	negative → terrible; positive → great
MRPC	The sentences are `[MASK]`. Sentence 1: ⟨sentence1⟩ Sentence 2: ⟨sentence2⟩	not equivalent → different; equivalent → equivalent
QNLI	The answer is `[MASK]`. Question: ⟨question⟩ Context: ⟨sentence⟩	entailment → yes; not entailment → no
RTE	The hypothesis is `[MASK]`. Premise: ⟨sentence1⟩ Hypothesis: ⟨sentence2⟩	entailment → true; not entailment → false
STS-B	The sentences are `[MASK]`. Sentence 1: ⟨sentence1⟩ Sentence 2: ⟨sentence2⟩	unrelated, different, similar, equivalent, identical

RoBERTa-large prompts. RoBERTa-large evaluates z_j as the masked-language-model logit of verbalizer v_j at the unique `[MASK]` position. The exact templates and label mappings are listed below; angle-bracketed text denotes a field copied from the corresponding GLUE example.

RoBERTa-large protocol. All 355,412,057 parameters are trainable. At each step, we use an estimator with $m = 10$ random unit-sphere directions and a Euclidean smoothing radius $\mu_1 = 0.01$. The batch size is 64, the maximum sequence length is 64 for SST-2 and 128 for the remaining tasks, and the formal seeds are $\{7, 17, 27, 37, 47\}$. We use fixed training/validation slices of $1000/372$ (SST-2), $1000/208$ (MRPC), $5000/500$ (QNLI), $1000/128$ (RTE), and $1000/500$ (STS-B). MRPC, QNLI, RTE, and STS-B are optimized for 300 steps, while SST-2 is optimized for 500 steps. Hyperparameters are selected on a disjoint tuning slice using the same candidate budget for every method; the reported numbers are the five-seed validation means and standard deviations.

Qwen3.5-2B prompts. We use the model's chat template with thinking disabled, no system instruction, and no in-context demonstrations. The task prompt is placed in the user turn, and z_j is read from the next-token logits at the first assistant position. For each verbalizer, we retain all valid

Table 5: User-turn prompt templates and next-token verbalizers used for Qwen3.5-2B.

Task	Prompt template	Verbalizers
SST-2	Classify the sentiment of the following sentence as negative or positive. Sentence: ⟨sentence⟩ Sentiment:	negative, positive
MRPC	Do the following two sentences have the same meaning? Answer yes or no. Sentence 1: ⟨sentence1⟩ Sentence 2: ⟨sentence2⟩ Answer:	no, yes
QNLI	Does the context contain the answer to the question? Answer yes or no. Question: ⟨question⟩ Context: ⟨sentence⟩ Answer:	yes, no
STS-B	Describe the semantic similarity of the following sentences as unrelated, different, similar, equivalent, or identical. Sentence 1: ⟨sentence1⟩ Sentence 2: ⟨sentence2⟩ Similarity:	unrelated, different, similar, equivalent, identical

one-token variants among the lowercase, capitalized, leading-space lowercase, and leading-space capitalized forms. If V_j is this set, its class score is the log-mean-exp

$$z_j = \log\left(\frac{1}{|V_j|}\sum_{v\in V_j}\exp \ell_v\right),$$

where ℓ_v is the next-token logit. This avoids making the result depend on a single capitalization or tokenizer boundary convention. The exact user-turn prompts are:

Qwen3.5-2B protocol. We optimize all 1,881,825,088 parameters of the text language model and tied output head. The unused vision encoder is frozen because the GLUE inputs are text-only. We evaluate in bfloat16 with maximum sequence length 128. Each step uses $m = 50$ directions and $\mu_1 = 10^{-4}$, together with batch size 128. We run 30 steps per task and report the resulting five-seed validation means and standard deviations in Table 2.

Interpretation. ZO-AdaNesterov achieves the highest mean on every reported task for both model families. For RoBERTa-large, its gains over the strongest baseline are largest on QNLI and RTE, at 4.12 and 2.66 percentage points, respectively. For Qwen3.5-2B, it reaches 77.50, 88.08, 83.18, 91.78, and 90.12 on MRPC, QNLI, RTE, SST-2, and STS-B, respectively. The baselines do not have a universal ordering—for example, ZO-AdaMM is stronger than R-AdaZO on RoBERTa-large STS-B, whereas R-AdaZO is stronger on MRPC—so the consistent advantage of ZO-AdaNesterov is not explained by a single uniformly dominant baseline.

K.3 Frozen-Backbone Image Classification

CIFAR-10 with ResNet-18. We evaluate the optimizers under a linear-probe protocol using an ImageNet-pretrained ResNet-18. We replace the original output layer with a randomly initialized ten-class linear classifier and optimize only this layer on CIFAR-10, while keeping the feature extractor and batch-normalization statistics fixed. This protocol gives a 5,130-dimensional optimization problem and provides the same pretrained representation and initialization to every method. All methods use an identical mini-batch sequence, $m = 10$ forward spherical directions, and smoothing radius $\mu_1 = 5 \times 10^{-3}$. We report the mean and standard deviation over seeds $\{1, 2, 3, 4, 5\}$.

Table 6: CIFAR-10 linear-probe validation results (mean ± standard deviation over five seeds).

Method	Loss ↓	Accuracy (%) ↑
ZO-SGD	1.5776 ± 0.0202	43.982 ± 0.899
ZO-AdaMM	1.5781 ± 0.0247	43.834 ± 1.105
R-AdaZO	1.5852 ± 0.0357	44.968 ± 1.054
ZO-AdaNesterov	$\mathbf{1.5596 \pm 0.0348}$	$\mathbf{45.150 \pm 1.327}$

ZO-AdaNesterov obtains the best mean endpoint loss and accuracy. It also has the lowest normalized validation-loss AUC (1.7508 versus 1.7746 for the next-best R-AdaZO) and reaches validation loss 1.6 approximately 50 steps earlier, showing that its advantage is reflected in both convergence speed and endpoint performance.

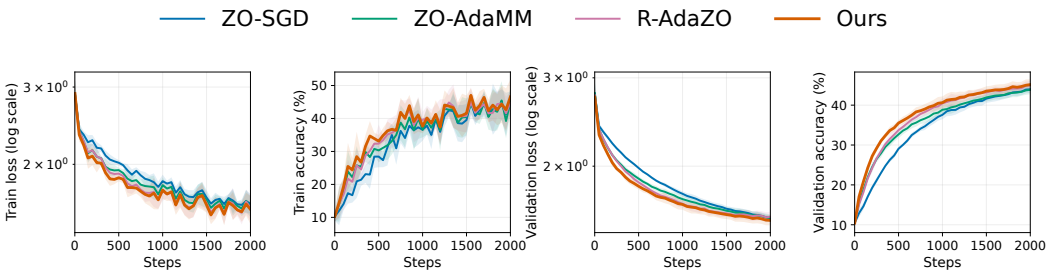

Figure 3: CIFAR-10 linear-probe optimization trajectories. Lines show five-seed means and shaded regions show one standard deviation.

Table 7: Tiny-ImageNet ViT-L/16 linear-probe endpoints after 1,500 steps (five-seed mean ± standard deviation). Accuracy is reported in percent.

Method	Train loss ↓	Train acc. ↑	Val. loss ↓	Val. acc. ↑
ZO-SGD	1.6828 ± 0.0564	$\mathbf{70.938 \pm 2.417}$	1.7388 ± 0.0134	70.590 ± 0.339
ZO-AdaMM	2.5955 ± 0.0438	69.766 ± 2.729	2.6159 ± 0.0221	69.254 ± 0.490
R-AdaZO	2.0582 ± 0.0708	65.391 ± 3.321	2.1200 ± 0.0223	64.678 ± 0.244
ZO-AdaNesterov	$\mathbf{1.5947 \pm 0.0596}$	70.547 ± 2.658	$\mathbf{1.6461 \pm 0.0125}$	$\mathbf{70.892 \pm 0.288}$

Tiny-ImageNet with ViT-L/16. We further evaluate the methods on Tiny-ImageNet-200 (Le & Yang, 2015) using a ViT-L/16 (Dosovitskiy, 2020) pretrained on ImageNet-1K (Deng et al., 2009). We freeze the transformer backbone, replace its ImageNet head with a randomly initialized 200-class linear classifier, and optimize only this shared linear head. The resulting problem has 205,000 trainable parameters (a 1024×200 weight matrix and 200 biases), while the complete network contains approximately 303.5 million parameters. This setting isolates optimizer behavior in a substantially higher-dimensional problem than the CIFAR-10 probe while retaining identical pretrained features for all methods.

We use all 100,000 training images and all 10,000 validation images. Inputs are resized to 224×224 and normalized with the ImageNet channel statistics. Training uses random resized crops and horizontal flips; validation uses deterministic resizing. The mini-batch size is 256, and the validation set is evaluated every 100 steps. At a fixed seed, every method receives the same head initialization, augmented mini-batch sequence, and spherical-direction sequence. The frozen-backbone validation features are cached once and reused; this changes only evaluation time and not the optimization trajectory.

All methods run for 1,500 steps with $m = 50$ directions spherical estimator. We use the seeds $\{181, 191, 201, 211, 221\}$ and report the mean and sample standard deviation. The optimized objective is mini-batch cross-entropy. Hyperparameters are selected from method-specific grids and then frozen across the five reported seeds.

Table 7 reports the fixed-budget endpoints. ZO-AdaNesterov achieves the lowest training loss (1.5947 ± 0.0596) and validation loss (1.6461 ± 0.0125), as well as the highest validation accuracy ($70.892 \pm 0.288\%$). Compared with ZO-SGD, it reduces the mean validation loss from 1.7388 to 1.6461 and improves the mean validation accuracy by 0.302 percentage points. ZO-SGD attains the highest training accuracy, while ZO-AdaNesterov remains close, with mean accuracies of 70.938% and 70.547%, respectively. Overall, ZO-AdaNesterov provides the best validation performance under the same optimization and query budgets.

The training columns are measured on the current mini-batch and are consequently noisier than the complete-validation metrics. Figure 4 presents the complete trajectories. The uncertainty bands are intentionally light to keep overlapping five-seed curves legible.

K.4 Black-Box Adversarial Attack

We evaluate an untargeted per-image black-box attack on a pretrained CIFAR-10 ResNet-18 victim with 95.87% clean test accuracy. The victim is fixed throughout optimization: the attacker observes

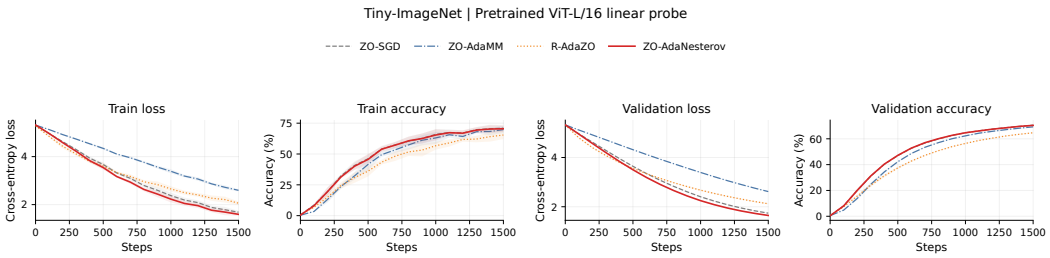

Figure 4: Tiny-ImageNet linear-probe optimization with an ImageNet-pretrained ViT-L/16 and a shared 200-class head. The four panels show training loss, training accuracy, validation loss, and validation accuracy over 1,500 steps. Lines are five-seed means, and shaded regions show one sample standard deviation.

only its output logits and never accesses its parameters or gradients. Images are represented in $(0,1)^{3\times32\times32}$ and normalized by the CIFAR-10 channel statistics before being passed to the victim. We attack the same first 100 correctly classified test images in every run.

Attack objective. For a clean image x with true label y, we optimize an unconstrained variable $\delta \in \mathbb{R}^{3\times32\times32}$ initialized at zero. The change of variables

$$x'(\delta) = \frac{1}{2}\left[\tanh(\operatorname{arctanh}(2x-1)+\delta)+1\right]$$

keeps every adversarial pixel in $(0,1)$. If $Z_j(x')$ is the victim logit for class j, the optimized C&W-style untargeted objective is (Carlini & Wagner, 2017; Chen et al., 2019)

$$F(\delta) = \|x'(\delta)-x\|_2^2 + c\max\left\{Z_y(x'(\delta)) - \max_{j\neq y} Z_j(x'(\delta)) + \kappa,\ 0\right\}.$$

We set $c=1$ and confidence margin $\kappa=0$. An image is counted as successfully attacked once $\arg\max_j Z_j(x') \neq y$; its run terminates at the first such step. ASR is the fraction of the 100 images satisfying this condition, and the reported ℓ_2 value is $\|x'-x\|_2$ in the unnormalized pixel space.

The selected optimizer settings are: learning rate 0.2 for ZO-SGD; learning rate 9×10^{-4} and numerical constant 10^{-5} for ZO-AdaMM; learning rate 2.2×10^{-4}, $(\beta_1,\beta_2)=(0.9,0.999)$, and numerical constant 10^{-5} for R-AdaZO; and $(a,b,\theta)=(0.88,20,0.6)$ for ZO-AdaNesterov. Unsuccessful images are assigned the maximum iteration and query budgets when computing speed, so a method is not rewarded for a lower success rate.

Table 8: Per-image black-box attack results (mean $\pm$ standard deviation over five seeds). Iterations and queries are penalized for unsuccessful images.

Method	Iters. ($\times10^3$) $\downarrow$	Queries ($\times10^3$) $\downarrow$	Speedup	ASR (%) $\uparrow$	ℓ_2
ZO-SGD	0.234 ± 0.000	2.805 ± 0.005	$1.00\times$	80.0 ± 1.6	1.199 ± 0.013
ZO-AdaMM	0.190 ± 0.003	2.283 ± 0.030	$0.99\times$	86.4 ± 1.9	1.001 ± 0.007
R-AdaZO	0.182 ± 0.002	2.181 ± 0.019	$1.03\times$	88.0 ± 2.5	1.093 ± 0.006
ZO-AdaNesterov	$\mathbf{0.179\pm0.004}$	$\mathbf{2.152\pm0.052}$	$\mathbf{1.05\times}$	$\mathbf{90.0\pm1.0}$	0.976 ± 0.006

ZO-AdaNesterov achieves the best overall attack performance, requiring the fewest penalized iterations and queries while attaining the highest ASR. Compared with the tuned ZO-SGD baseline, it reduces the average iteration and query counts by approximately 23% and improves the ASR by 10.0 percentage points, with a measured speedup of $1.05\times$. It also outperforms ZO-AdaMM and R-AdaZO in both attack efficiency and success rate.