Adam under Generalized Smoothness with Second-Moment-Type Stochastic Gradients

Ruinan Jin
Mohamed bin Zayed University of Artificial Intelligence
jinruinan3@gmail.com

Difei Cheng
School of Computer Science and Artificial Intelligence
Aerospace Information Technology University

Ling Chen
The Ohio State University

Jun Luo
The Ohio State University

Hao Zhou
JD.com, Inc.

Youzhi Zhang*
Centre for Artificial Intelligence and Robotics,
Hong Kong Institute of Science & Innovation,
Chinese Academy of Sciences
youzhi.zhang@cair-cas.org.hk

Abstract

It has been widely observed that Adam can remain stable even when the objective function deviates significantly from global smoothness. However, under the generalized smoothness framework, existing theoretical analyses typically rely on strong tail assumptions on stochastic gradients, such as almost-sure boundedness or sub-Gaussianity. Whether one can establish the convergence of Adam on generalized smooth objectives under only second moment information on the stochastic gradients, without imposing such strong concentration assumptions, was explicitly identified as an important open direction by Li et al. [2023b]. This paper gives an affirmative answer to this question under fairly general conditions. Specifically, we prove that such strong tail assumptions are not necessary. Building on the Adam self-normalization framework of Jin et al. [2026], which was developed for classical smoothness and bounded variance, we extend the stopping-time and de-preconditioning strategy to the L_0–L_p generalized smoothness condition and a generalized second moment ABC condition. This extension shows that, even when the stochastic-gradient condition provides only second moment information and may grow along the trajectory, the stochastic trajectory of Adam remains in a locally well-behaved smoothness region, with stretched-exponential tail decay under bounded variance and global smoothness. As a consequence, we establish high-probability convergence rate guarantees over the full range $p < 2$, with a confidence dependence of order $\delta^{-1/2}$, while the stepsize prefactor depends on δ only through a single logarithmic factor. Furthermore, we construct a hard instance proving that, under only second-moment information on the stochastic gradients, this $\delta^{-1/2}$-type confidence dependence is sharp. Finally, in the more favorable regime $p < 1$, we combine the above trajectory control with polynomial-growth estimates on rare events to further obtain convergence rate guarantees in expectation.

*Corresponding author.

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

1 Introduction

Adaptive gradient methods are a standard tool in large-scale stochastic optimization, and Adam is arguably the most widely used example [Kingma and Ba, 2015, Reddi et al., 2018, Défossez et al., 2022, Li et al., 2023b]. Its practical robustness is usually attributed to two interacting mechanisms: momentum, which averages stochastic directions, and coordinatewise normalization, which rescales the update by a running average of squared gradients. These mechanisms are especially relevant when the objective is far from globally smooth. A common model for this behavior is L_0–L_p generalized smoothness [Crawshaw et al., 2022, Zhang et al., 2020, Chen et al., 2023, Li et al., 2023a, Gorbunov et al., 2025], where the local variation of the gradient is allowed to grow with the size of the gradient itself.

The main obstacle in this regime is not merely to prove a descent inequality. Since the local smoothness scale depends on the current gradient norm, one first has to show that the stochastic trajectory does not enter a region where the objective becomes too irregular. For Adam this is delicate: the algorithm normalizes stochastic gradients by $\mathbf{v}_t$, but $\mathbf{v}_t$ itself contains the raw squared stochastic gradients. Under bounded or sub-Gaussian noise, one can often control this feedback loop through direct concentration. Under a second-moment assumption alone, rare but very large stochastic-gradient spikes remain possible, so the raw gradient sequence need not have useful high-probability bounds.

This paper asks whether Adam can nevertheless be analyzed under such bounded-variance-type information. This is not merely a technical variant of existing results: after proving Adam guarantees under stronger bounded-gradient or sub-Gaussian assumptions, Li et al. [2023b] explicitly identified the bounded-variance setting as a challenging and important open direction for Adam under generalized smoothness. The answer we give is positive, but the result has a different structure from analyses based on strong tails. Our analysis builds on the self-normalized stopping-time framework of Jin et al. [2026], developed for Adam under classical smoothness and bounded variance, and extends it to generalized smoothness and generalized ABC second-moment growth. In this broader regime, self-normalization still controls the scaled quantities that actually move the iterate, not the raw stochastic gradients themselves. This distinction gives polylogarithmic trajectory localization, but controlling the adaptive denominator $\mathbf{v}_t$ in the final de-preconditioning step necessarily incurs a $\delta^{-1/2}$ factor. Moreover, unlike the classical-smooth bounded-variance result of Jin et al. [2026], the local-smoothness and noise-growth restrictions here lead to a stepsize prefactor that depends on the confidence level only through a single logarithmic factor. We show through a hard instance that the resulting $\delta^{-1/2}$ factor is not a proof artifact in the stated high-probability stationarity guarantee.

Contributions. For the horizon-dependent theoretical calibration $\beta_2 = 1 - 1/T$, we prove the following results for Adam under L_0–L_p generalized smoothness and a generalized second-moment ABC condition, using the Adam self-normalization framework of Jin et al. [2026] as a starting point.

- *Trajectory localization without tail assumptions.* We extend the self-normalized stopping-time analysis to prove that, for any confidence level δ, both the auxiliary energy and the gradient scale along the Adam trajectory are bounded by $\mathrm{polylog}(1/\delta)$ with probability at least $1 - \delta$. Under bounded variance and global smoothness, a fixed stepsize gives stretched-exponentially decaying tail probability.
- *High-probability convergence for $p < 2$.* The localization estimate yields a preconditioned energy bound. After converting it to an unweighted stationarity guarantee, we obtain

$$\frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_t)\|^2 \le \widetilde{\mathcal{O}}\Big(\frac{d}{\sqrt{\delta T}} + \frac{d^2}{T}\Big)$$

  with probability at least $1 - \delta$. The additional $\delta^{-1/2}$ factor comes from controlling the maximum adaptive denominator.
- *Sharpness of the confidence dependence.* We construct a one-dimensional hard instance showing that, under second-moment assumptions alone, the $\delta^{-1/2}$ dependence in the high-probability stationarity rate cannot generally be replaced by a polylogarithmic dependence. In the construction, a rare spike in the stochastic gradient raises the adaptive denominator and keeps the iterates in a nonstationary ramp region for a constant fraction of the horizon.

- *Expectation convergence and hard instances.* In the more structured range $p < 1$, generalized smoothness implies a polynomial distance-growth estimate for the true gradient. This allows the rare-event terms to be integrated, giving

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 \le \mathcal{O}\left(\frac{d(1+\log T)}{\sqrt{T}} + \frac{d^2(1+\log T)^2}{T}\right).$$

  For $1 \le p < 2$, we construct a fixed class of one-dimensional hard instances satisfying the same assumptions for which the calibrated Adam family has a worst-case expected squared-stationarity lower bound of $\Omega(T^{-1/3})$, uniformly over deterministic, possibly horizon-dependent, base stepsize prefactors. Thus the above expectation rate cannot extend uniformly to the full generalized ABC class for this Adam family (Proposition 4.2).

Organization. Section 2 discusses the relation to existing analyses of Adam, generalized smoothness, and high-probability stochastic optimization. Section 3 formulates the problem, states the assumptions and Algorithm 1, and fixes notation. Section 4 states the main results. Section 5 outlines the upper- and lower-bound proofs, while full proofs are deferred to the appendix.

2 Related Work

Li et al. [2023b] analyze Adam under generalized smoothness with bounded or sub-Gaussian noise. Zhang et al. [2025] and Wang et al. [2024a] obtain convergence rates under (L_0, L_1)-type smoothness with affine noise variance, for coordinatewise and scalar normalization, respectively. Theorem 4.2 is an expectation result only for $p < 1$, with second moments that may grow with the function gap, while Theorem 4.1 is a high-probability result for all $p < 2$. The closest high-probability analysis under generalized smoothness is Hong and Lin [2024], which allows exponents below 2 and obtains a polylogarithmic dependence on $1/\delta$ under an exponential-moment condition on the noise. This condition is much stronger than Assumption 3.3, which bounds only conditional second moments, so the two confidence dependences are not directly comparable. Under second moments alone, the factor $\delta^{-1/2}$ cannot in general be replaced by a polylogarithmic one (Proposition 4.1). Jin et al. [2025] establish convergence under ABC second moments. Jin et al. [2026] obtain polynomial-confidence guarantees under bounded conditional variance. We study generalized smoothness with generalized ABC growth, retaining coordinatewise normalization and distinguishing high-probability guarantees from expectation rates. Here exponential memory controls the adaptive denominator's persistence. Appendix A discusses the broader literature on adaptive methods, noise models, localization, clipping, and the comparison scopes of stochastic optimization lower bounds.

3 Preliminaries

Throughout, let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and let $\{\zeta_t\}_{t\ge1}$ denote a sequence of independent random variables driving the stochastic-gradient oracle. All random quantities are defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Expectation with respect to $\mathbb{P}$ is denoted by $\mathbb{E}[\cdot]$.

3.1 Adaptive Moment Estimation (Adam)

Adam [Kingma and Ba, 2015, Reddi et al., 2018, Wang et al., 2023a, Jin et al., 2025] is an adaptive first-order optimization method widely employed in stochastic settings. It maintains a denominator-aware first-moment variable and an exponentially weighted average of squared gradients. These moment estimates are then used to construct coordinatewise step sizes that adapt to the local geometry of the stochastic noise.

Formally, let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and let G be a stochastic-gradient oracle driven by independent random variables $\{\zeta_t\}_{t\ge1}$. At each iteration $t \ge 1$, the stochastic gradient is denoted by

$$\mathbf{g}_t = \mathsf{G}(\boldsymbol{\theta}_t; \zeta_t).$$

The algorithm maintains two auxiliary variables $\mathbf{m}_t, \mathbf{v}_t \in \mathbb{R}^d$ and computes the next iterate through a coordinatewise scaling of the momentum term. The complete procedure is summarized in Algorithm 1.

Algorithm 1 Adam

Input: Stochastic-gradient oracle G; base step size $\alpha > 0$; initial point $\boldsymbol{\theta}_1 \in \mathbb{R}^d$; $\mathbf{m}_0 = \mathbf{0}$; $\mathbf{v}_0 = v_{\text{init}}\,\mathbf{1}$ with $v_{\text{init}} > 0$; $\beta_1 \in [0,1)$; $\beta_2 \in [0,1)$; $\epsilon > 0$; number of iterations T.
for $t = 1, \dots, T-1$ **do**
  Sample ζ_t and compute $\mathbf{g}_t \leftarrow \mathsf{G}(\boldsymbol{\theta}_t; \zeta_t)$.
  Update the second-moment estimate: $\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)(\mathbf{g}_t \odot \mathbf{g}_t)$.
  Update the first-moment estimate: $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$.
  Form the coordinatewise effective step size: $\lambda_t \leftarrow \alpha \cdot (\sqrt{\mathbf{v}_t} + \epsilon)^{-1}$, where the inverse is taken element-wise.
  Update the iterate: $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \lambda_t \odot \mathbf{m}_t$.
end for
Output: iterate sequence $(\boldsymbol{\theta}_t)_{t=1}^T$.

All arithmetic operations in Algorithm 1 are executed componentwise, and the symbol $\odot$ denotes the Hadamard product. For notational convenience, we also define the initial coordinatewise preconditioning vector by $\lambda_0 := \alpha(\sqrt{v_{\text{init}}} + \epsilon)^{-1}\mathbf{1}$, where $v_{\text{init}} > 0$ is the scalar initialization in $\mathbf{v}_0 = v_{\text{init}}\mathbf{1}$. For a vector $u \in \mathbb{R}^d$, we write u_i for its i-th coordinate; for a time-indexed vector $u_t \in \mathbb{R}^d$, we write $u_{t,i}$ for its i-th coordinate.

Theoretical calibration. The results use $\alpha = \bar{\alpha}/\sqrt{T}$ and $\beta_2 = 1 - 1/T$ for an integer horizon $T \geq 10$, with fixed $\beta_1 \in [0,1)$, $v_{\text{init}} > 0$ and $\epsilon > 0$. This finite-horizon calibration, as in Jin et al. [2026] and Ghadimi and Lan [2013], fixes both parameters once T is fixed. By Lemma C.4, $\mathbf{v}_{t,i} \geq v_{\text{init}}/4$ and $\lambda_{t,i} \leq 2\lambda_{k,i}$ for $0 \leq k \leq t \leq T-1$, so the effective stepsizes are comparable along the whole run. The lower bound on $\mathbf{v}_t$ comes from the retained initialization, not from the stochastic gradients. We measure stationarity by $\frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_t)\|^2$, the average of the squared true-gradient norm over the iterates output by Algorithm 1, not a single final iterate. Time-varying β_2 schedules and restarts are outside the scope of the theorems.

Filtration. Let $\mathscr{F}_t = \sigma(\zeta_1, \dots, \zeta_t)$ for $t \geq 0$, with $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and $\mathscr{F}_\infty = \sigma\big(\cup_{t\geq 0}\mathscr{F}_t\big)$. The filtration $\{\mathscr{F}_t\}$ encodes the information available up to time t. Each stochastic gradient $\mathbf{g}_t$ is $\mathscr{F}_t$-measurable, while $\boldsymbol{\theta}_t$ is $\mathscr{F}_{t-1}$-measurable. The Euclidean norm $\|\cdot\|$ is used throughout. For nonnegative bases, a zeroth power is interpreted as one, including $0^0 = 1$.

3.2 Analytic Assumptions

We impose the following structural assumptions on the objective function and the stochastic gradient oracle.

Assumption 3.1 (Lower boundedness)**.** *The objective $f : \mathbb{R}^d \to \mathbb{R}$ is bounded from below, i.e.,*

$$f^* := \inf_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta}) > -\infty.$$

Assumption 3.2 (L_0-L_p smoothness)**.** *The function $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. There exist constants $L_0 > 0$, $L_p > 0$, and a smoothness exponent $p \in [0,2)$ such that, for all $u, u' \in \mathbb{R}^d$ satisfying $\|u - u'\| \leq L_p^{-1}$,*

$$\|\nabla f(u) - \nabla f(u')\| \leq \big(L_0 + L_p\|\nabla f(u)\|^p\big)\|u - u'\|.$$

When $p = 0$, this reduces to global smoothness up to constants. For $p > 0$, the local Lipschitz constant of the gradient is allowed to grow with the gradient norm. This captures objectives whose curvature can be much larger far from stationary regions, and it is precisely why a trajectory-localization argument is needed before applying descent estimates.

Assumption 3.3 (Generalized second-moment ABC inequality)**.** *Let $\{\mathscr{F}_t\}_{t\geq 0}$ be the natural filtration generated by the algorithm, and assume that $\boldsymbol{\theta}_t$ is $\mathscr{F}_{t-1}$-measurable. For each $t \geq 1$, the stochastic gradient $\mathbf{g}_t$ is an unbiased estimator of $\nabla f(\boldsymbol{\theta}_t)$, namely*

$$\mathbb{E}[\mathbf{g}_t \mid \mathscr{F}_{t-1}] = \nabla f(\boldsymbol{\theta}_t).$$

Moreover, there exist constants $A, B, C, \rho_1, \rho_2 \geq 0$ such that

$$\mathbb{E}\big[\|\mathbf{g}_t\|^2 \mid \mathscr{F}_{t-1}\big] \leq A\big(f(\boldsymbol{\theta}_t) - f^*\big)^{\rho_1} + B\|\nabla f(\boldsymbol{\theta}_t)\|^{\rho_2} + C.$$

This assumption is a second-moment condition, not a tail condition. It permits the conditional variance of the stochastic gradient to depend on both the function gap and the gradient norm, but it does not rule out rare large stochastic-gradient values. This is the main distinction from bounded-gradient, sub-Gaussian, or exponentially concentrated oracle models. By conditional unbiasedness,

$$\mathbb{E}[\|\mathbf{g}_t\|^2 \mid \mathscr{F}_{t-1}] = \|\nabla f(\boldsymbol{\theta}_t)\|^2 + \mathbb{E}[\|\mathbf{g}_t - \nabla f(\boldsymbol{\theta}_t)\|^2 \mid \mathscr{F}_{t-1}].$$

Hence Assumption 3.3 may equivalently be viewed as a variance-growth condition up to the deterministic term $\|\nabla f(\boldsymbol{\theta}_t)\|^2$. The classical bounded-variance assumption

$$\mathbb{E}[\|\mathbf{g}_t - \nabla f(\boldsymbol{\theta}_t)\|^2 \mid \mathscr{F}_{t-1}] \le \sigma^2$$

is recovered in variance form by taking $A = B = 0$ and $C = \sigma^2$. In the second-moment form used above, it corresponds to $A = 0$, $B = 1$, $\rho_2 = 2$, and $C = \sigma^2$.

4 Results

Throughout this section, we write

$$\Psi(x) := f(x) - f^* + 1. \tag{1}$$

We also use the standard auxiliary sequence

$$\mathbf{z}_1 := \boldsymbol{\theta}_1, \qquad \mathbf{z}_t := \frac{\boldsymbol{\theta}_t - \beta_1 \boldsymbol{\theta}_{t-1}}{1-\beta_1}, \qquad t \ge 2, \tag{2}$$

which removes the leading momentum term from the recursion. The results are organized around the two quantities that must be controlled in Adam: the trajectory itself and the adaptive denominator. Formal auxiliary lemmas, intermediate propositions, and complete proofs are deferred to the appendix.

The main statements below give the convergence orders. The order notation suppresses fixed problem and initialization parameters other than the dimension d; $\widetilde{\mathcal{O}}$ also suppresses logarithmic factors in $1/\delta$. Appendix B gives the complete inequalities; their constants are explicit but not optimized.

To display the confidence dependence of the stepsize, define

$$\eta_{\mathrm{gen}} := \mathbf{1}_{\{p>0\}} + \sqrt{A + [B-1]_+ + B\mathbf{1}_{\{\rho_2 \ne 2\}}}, \qquad [x]_+ := \max\{x, 0\}. \tag{3}$$

Theorem 4.1 (High-probability convergence, informal)**.** *Assume Assumptions 3.1–3.3. Let $T \ge 10$ and $\beta_2 = 1 - 1/T$. For every $0 < \delta < 1$, an admissible prefactor $\bar{\alpha} > 0$ can be chosen with*

$$\bar{\alpha}^{-1} = d\Big[\mathcal{O}(1) + \mathbf{1}_{\{\eta_{\mathrm{gen}}>0\}}\,\mathcal{O}(\log(1/\delta))\Big], \tag{4}$$

where η_{gen} is defined in (3)*. Adam with $\alpha = \bar{\alpha}/\sqrt{T}$ then satisfies, with probability at least $1-\delta$,*

$$\max_{1\le t\le T} \Psi(\mathbf{z}_t) = \mathcal{O}\big((1+\log(1/\delta))^{4\mathbf{1}_{\{\eta_{\mathrm{gen}}=0\}}}\big), \qquad \frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_t)\|^2 = \widetilde{\mathcal{O}}\Big(\frac{d}{\sqrt{\delta T}} + \frac{d^2}{T}\Big).$$

Here Ψ and $\mathbf{z}_t$ are defined in (1)–(2)*.*

The complete inequalities and admissible stepsizes are given in Theorem B.1 of Appendix B, where the parameter scale $\mathcal{P}$, the calibration scale $\mathcal{R}$ and the confidence height U_δ are defined. The decomposition (4) describes the upper endpoint of that stepsize interval; its two order constants are independent of T, δ, d, and for $\eta_{\mathrm{gen}} > 0$ the height U_δ is independent of δ.

Claim 4.1 (Smooth bounded-variance specialization, informal)**.** *For $p = 0, A = 0, B = 1, \rho_2 = 2$, the same result with a fixed prefactor $\bar{\alpha} > 0$ satisfying $\bar{\alpha}^{-1} = \mathcal{O}(d)$, as in* (4)*, gives*

$$\frac{1}{T}\sum_{t=1}^{T-1} \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 = \mathcal{O}(dT^{-1/2} + d^2T^{-1}).$$

Here $\eta_{\mathrm{gen}} = 0$ in (3), so the logarithmic term in (4) vanishes. The same fixed stepsize prefactor is therefore admissible at every confidence level. Integrating the quantile bound from Theorem B.1 therefore gives

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 \le \mathcal{O}(dT^{-1/2})\int_0^1 \frac{(1+\log(16/\delta))^4}{\sqrt{\delta}}\,d\delta \\ + \mathcal{O}(d^2T^{-1})\int_0^1 (1+\log(16/\delta))^8\,d\delta = \mathcal{O}(dT^{-1/2}+d^2T^{-1}).$$

The constants in the order terms are independent of T, δ, d, and both integrals are finite by the substitution $u = \log(1/\delta)$. Claim B.1 gives the explicit inequality; its full proof is in Appendix D.7.

By Cauchy–Schwarz, Claim 4.1 implies

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\| \le \left(\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2\right)^{1/2} = \mathcal{O}(\sqrt{d}\,T^{-1/4}+dT^{-1/2}).$$

For the regularized Adam calibration considered here, this recovers the expected-gradient-norm rate of Wang et al. [2023a, Theorem 2]. The implication between these moment criteria is strictly one-way: an expected norm bound alone does not control the corresponding squared moment. The algorithm in Wang et al. [2023a] uses an update without the denominator offset. The squared-gradient expectation rate in Claim 4.1 also agrees with the rate obtained by integrating the confidence bound of Jin et al. [2026, Theorem 1], whose stepsize prefactor is independent of the confidence level. Under global smoothness and bounded coordinatewise variance, assumptions much stronger than Assumptions 3.2–3.3, Li et al. [2025] obtain an expected ℓ_1 rate without logarithmic factors; Appendix A relates it to Claim 4.1, whose standard specialization carries no $\log T$ factor.

Theorem 4.2 (Expected convergence for $p < 1$, informal)**.** *Assume Assumptions 3.1–3.3 with $p < 1$, and let $T \ge 10$, $\beta_2 = 1 - 1/T$. A prefactor $\bar{\alpha}_T > 0$ with $\bar{\alpha}_T^{-1} = d[\mathcal{O}(1) + \mathbf{1}_{\{\eta_{\mathrm{gen}}>0\}}\mathcal{O}(\log T)]$ can be chosen so that Adam with $\alpha = \bar{\alpha}_T/\sqrt{T}$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 = \mathcal{O}\left(\frac{(1+\log T)^{\mathbf{1}_{\{\eta_{\mathrm{gen}}>0\}}}d}{\sqrt{T}} + \frac{(1+\log T)^{2\mathbf{1}_{\{\eta_{\mathrm{gen}}>0\}}}d^2}{T}\right).$$

Theorem B.2 in Appendix B gives the complete inequality and the polynomial-confidence calibration of the prefactor.

4.1 Rate obstructions

Proposition 4.1 (Confidence dependence, informal)**.** *Fix $\bar{\alpha} \in (0, 1]$, $v_{\mathrm{init}} > 0$, and $\epsilon > 0$. For sufficiently small $\delta > 0$ and sufficiently large T depending on δ, there are one-dimensional instances in the classical specialization $p = 0$, $A = 0$, $B = 1$, $\rho_2 = 2$ of Assumptions 3.1–3.3 for which Adam with $\beta_1 = 0$, $\beta_2 = 1 - 1/T$, and $\alpha = \bar{\alpha}/\sqrt{T}$ satisfies*

$$\frac{1}{T}\sum_{t=1}^{T-1}|\nabla f(\boldsymbol{\theta}_t)|^2 = \Omega\left(\frac{1}{\sqrt{\delta T}}\right) \quad \textit{with probability } \Omega(\delta).$$

The problem constants, initial objective gaps, and constants hidden in Ω are uniform in T, δ. This obstruction also holds for fixed prefactors admitted by Theorem B.1.

Proposition B.1 gives the precise probability, parameter ranges, and comparison with the calibrated upper bound.

Proposition 4.2 (Expected stationarity for $1 \le p < 2$, informal)**.** *Fix $p \in [1, 2)$, $\beta_1 \in [0, 1)$, $v_{\mathrm{init}} > 0$, and $\epsilon > 0$. There is a fixed class $\mathcal{C}_p$ of one-dimensional instances satisfying Assumptions 3.1–3.3, with common parameters including $A = 0$ and $\rho_2 = 3$, and uniform bounds on $|\theta_1|$, $f(\theta_1) - f^*$, and $|f'(\theta_1)|$, such that Adam with $\beta_2 = 1 - 1/T$ and $\alpha = \bar{\alpha}_T/\sqrt{T}$ satisfies, for all sufficiently large T,*

$$\inf_{\bar{\alpha}_T > 0} \sup_{(f, \mathsf{G}, \theta_1) \in \mathcal{C}_p} \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}|f'(\theta_t)|^2 \ge \Omega(T^{-1/3}).$$

The infimum is over deterministic, possibly horizon-dependent, prefactors; the class and implicit constant are independent of $T, \bar{\alpha}_T$. Consequently, the $\mathrm{polylog}(T)/\sqrt{T}$ expectation rate cannot hold uniformly for this Adam family on the full generalized ABC class.

Proposition B.2 specifies the common class parameters, the lower-bound coefficient, and the horizon range.

5 Proof Sketch

We summarize the main ideas behind the proofs. The central difficulty is to obtain high-probability trajectory control without assuming that the raw stochastic gradients concentrate. Following the self-normalized stopping-time viewpoint of Jin et al. [2026], the proof focuses on quantities that are normalized by Adam's adaptive denominator, and only later converts the resulting preconditioned estimates into the usual stationarity measure. Two growth terms enter every descent estimate. Under Assumption 3.2 the curvature coefficient $L_0 + L_p\|\nabla f(\boldsymbol{\theta}_t)\|^p$ varies along the random trajectory, and under Assumption 3.3 the conditional second moment is bounded by $C + A(f(\boldsymbol{\theta}_t) - f^*)^{\rho_1} + B\|\nabla f(\boldsymbol{\theta}_t)\|^{\rho_2}$, which depends on the same trajectory. Bounding both requires localization of the iterates, while the descent estimate that yields localization requires both bounds. We break it with the augmented energy $\mathcal{L}_t = \Psi(\mathbf{z}_t) + E_t$ from (5), stopped at the exit index σ_U from (32). Before exit, Lemmas C.1 and C.2 turn both growth terms into deterministic functions of U, the scales of (13)–(14), and the pathwise logarithmic bounds of Lemma D.1 together with the stopped martingale estimates close with conditional second moments only (Lemma D.2). The auxiliary sequence, the logarithmic energy lemma, and the de-preconditioning step (Appendix D.6) follow Jin et al. [2026]; the augmented energy with its gradient-energy telescope, the height-dependent calibration $\mathcal{R}(U)$, the conditional-to-direct comparison (Lemma C.5) with Freedman's inequality (25) before the crossing index τ from (47), the integration of the tail bounds over the confidence level for $p < 1$ (Appendix D.8), and the hard instances of Section 5.2 are specific to the present setting.

5.1 Upper Bounds

Auxiliary sequence and descent. The auxiliary sequence $\mathbf{z}_t$ from (2) removes the direct momentum lag:

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\lambda_t \odot \mathbf{g}_t + \frac{\beta_1}{1 - \beta_1}(\lambda_{t-1} - \lambda_t) \odot \mathbf{m}_{t-1}.$$

The second term records the changing denominator. Pair the auxiliary objective Ψ from (1) with the predictable gradient energy

$$E_t = \sum_i \lambda_{t-1,i}(\nabla f(\boldsymbol{\theta}_t))_i^2, \qquad \mathcal{L}_t = \Psi(\mathbf{z}_t) + E_t. \tag{5}$$

The gradient-energy difference in (5) cancels the leading true-gradient part of the same-sample denominator error. The remaining localized descent has negative drift $-E_t/8$, three centered terms, and seven nonnegative residuals; (38) gives the explicit decomposition. Its coefficients depend on the curvature and noise envelope along the trajectory.

Stopped preconditioned energy. The first ingredient turns objective localization into gradient and noise control.

Lemma 5.1 (Function-gap control, informal)**.** *Under Assumptions 3.1–3.2, with Ψ from* (1)*,*

$$\|\nabla f(x)\| = \mathcal{O}\Big(\Psi(x)^{\max\{1,(2-p)^{-1}\}}\Big).$$

If $\|x-y\| \le L_p^{-1}$*, then*

$$\Psi(x) = \mathcal{O}\Big(\Psi(y)^{\max\{1,p/(2-p)\}}\Big).$$

The constants depend only on L_0, L_p, p*.*

The complete bounds are Lemmas C.1 and C.2 in Appendix C. The first follows by taking a short step against the gradient and comparing its decrease with $f(x) - f^*$. The second integrates local smoothness between nearby points.

For $U \ge 1$, stop at

$$\sigma_U = \inf\{1 \le t \le T : \Psi(\mathbf{z}_t) > U\}, \qquad \inf \varnothing = \infty. \tag{6}$$

The calibration keeps $\boldsymbol{\theta}_t, \mathbf{z}_t$ within the comparison radius. Lemma 5.1 then bounds the true gradient before σ_U from (6) by a polynomial in U; the ABC condition gives a polynomial conditional second-moment envelope. The indicator of $t < \sigma_U$ is measurable before query t, so stopping preserves the centered terms' zero conditional means. Summing the descent reduces localization to controlling the residuals and martingales in the energy (5).

Self-normalization without raw-gradient tails. Large oracle outputs are controlled through their normalized contribution to the update.

Lemma 5.2 (Normalized energy, informal)**.** *For Algorithm 1 with* $T \ge 10$*,* $\beta_2 = 1 - 1/T$*, and* $\alpha = \bar{\alpha}/\sqrt{T}$*, for any stopping index* σ*,*

$$\sum_{t<T\wedge\sigma} \big(\|\lambda_t \odot \mathbf{g}_t\|^2 + \|\lambda_t \odot \mathbf{m}_t\|^2 + \|\lambda_{t-1} \odot \mathbf{m}_{t-1}\|^2\big) \le \mathcal{O}(\bar{\alpha}^2) \sum_i \log\left(1 + \frac{\sum_{t<T\wedge\sigma} \mathbf{g}_{t,i}^2}{T v_{\mathrm{init}}}\right).$$

The bound is pathwise.

Lemma D.1 is the formal statement. The accumulator comparison in Lemma C.4 allows logarithmic telescoping: a large sample raises both numerator and denominator, so its cumulative charge grows logarithmically. Expanding momentum into its geometric weights gives the same control for current and lagged momentum. Positive stepsize variation is bounded separately by (28).

Before σ_U from (6), the ABC envelope and Markov's inequality control the logarithm on the right. Lemma C.5 and the direct variation bound (28) control the predictable compensator of the positive stepsize variation. Freedman's inequality controls the centered descent terms. For the momentum-variation martingale, a predictable logarithmic crossing includes the first crossing update, whose size is paid for by the preceding normalized momentum energy. Thus no tail assumption on the raw gradient is needed.

From stopped energy to localization. The normalized-energy bounds close the stopped descent.

Lemma 5.3 (Localization and preconditioned energy, informal)**.** *Under Assumptions 3.1–3.3 and the calibration of Theorem 4.1, with probability at least* $1 - \delta/2$*,*

$$\max_{1\le t\le T} \Psi(\mathbf{z}_t) + \sum_{t=1}^{T-1} E_t = \mathcal{O}\big((1+\log(1/\delta))^{\mathbf{41}_{\{\eta_{\mathrm{gen}}=0\}}}\big),$$

where $\Psi, \mathbf{z}_t, E_t$ *are defined in* (1)*,* (2)*, and* (5)*.*

The precise stopped inequality and calibration appear in Lemma D.2. For the stopping height U in (6), take the order in Lemma 5.3; if $\eta_{\mathrm{gen}} > 0$, the logarithmic factor in (4) absorbs the logarithms. Absorb the centered sums and cumulative residuals into this height:

$$\mathcal{L}_{\sigma_U\wedge T} + \frac{1}{8} \sum_{t<\sigma_U\wedge T} E_t \le U/2.$$

The energy and stopping time are from (5)–(6). Since $\mathcal{L}_t \ge \Psi(\mathbf{z}_t)$, a first exit before or at T would make the left side exceed U, a contradiction. The stopping never activates, and the same inequality bounds the full preconditioned gradient energy.

From localization to stationarity. On the localization event, a second Markov bound on the stopped centered-noise energy gives, with additional failure probability at most $\delta/2$,

$$\max_{t<T,i} \mathbf{v}_{t,i} \le v_{\text{init}} + \frac{2}{T}\sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2 + \frac{\text{poly}(U)}{\delta},$$

where U is the height from (6) and the polynomial has coefficients depending only on the fixed problem parameters. Using this maximum to remove the coordinatewise weights in E_t from (5), and applying Lemma 5.3, yields the schematic bound

$$\frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_t)\|^2 \le \mathcal{O}\Big(\frac{U}{\bar{\alpha}\sqrt{T}}\Big)\left(1 + \sqrt{\frac{1}{T}\sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2} + \sqrt{\frac{\text{poly}(U)}{\delta}}\right).$$

Young's inequality absorbs the square-root gradient average into the left side. Substituting this height and the prefactor in (4) gives Theorem 4.1; Appendix D.6 supplies the complete inequalities. The square root of the noise-energy threshold introduces the $\delta^{-1/2}$ confidence factor.

Expectation bound for $p < 1$. The confidence-dependent calibration requires control of the most extreme trajectories. The crucial structural input is a distance bound on the true gradient.

Lemma 5.4 (Polynomial distance growth, informal)**.** *Under Assumption 3.2, if $0 \le p < 1$, then for every reference point x_{ref},*

$$\|\nabla f(x)\| = \mathcal{O}\Big((1 + \|x - x_{\text{ref}}\|)^{1/(1-p)}\Big), \qquad x \in \mathbb{R}^d.$$

The constant depends only on L_0, L_p, p and $\|\nabla f(x_{\text{ref}})\|$.

Lemma C.3 gives the complete inequality. Its proof applies the concavity of u^{1-p} to gradient increments along a partition of the segment from x_{ref} to x. This converts local generalized smoothness into a global polynomial distance estimate.

With the prefactor $\bar{\alpha}_T$ from Theorem 4.2, Adam's normalized displacement satisfies

$$\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\| \le 2\sqrt{d}\,\bar{\alpha}_T, \qquad \max_{t\le T} \|\boldsymbol{\theta}_t - \boldsymbol{\theta}_1\| \le 2\sqrt{d}\,\bar{\alpha}_T T,$$

as shown in (27). Combining this with Lemma 5.4, at $x_{\text{ref}} = \boldsymbol{\theta}_1$, gives

$$\frac{1}{T}\sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_t)\|^2 = \mathcal{O}\Big((1 + \sqrt{d}\,\bar{\alpha}_T T)^{2/(1-p)}\Big).$$

Fix the prefactor at the polynomial-confidence cutoff prescribed by Theorem B.2. The same run satisfies the high-probability bound at every larger confidence level. Integrating those quantiles costs only $\mathcal{O}(d(1+\log T)/\sqrt{T} + d^2(1+\log T)^2/T)$. The displayed pathwise bound controls all remaining quantiles: their interval has polynomially small length, chosen so that its product with this fixed power of T is $o(T^{-1/2})$. This gives Theorem 4.2. The exact cutoff, integral bound, and rare-event term appear in (19)–(20).

5.2 Hard Instance

Hard instance for confidence dependence. For Proposition 4.1, fix $\beta_1 = 0$ and $\bar{\alpha} \in (0, 1]$, with T, δ in the proposition's range. The one-dimensional objective has derivative $h = f'$ that vanishes at $\theta_1 = 0$ and has a positive plateau to its left. Its scales are

$$\chi_\star = (\delta T)^{-1/4}, \qquad H = \sqrt{\frac{T}{2\delta}}, \qquad L = 16\bar{\alpha}\sqrt{8}\,(\delta T)^{1/4}. \tag{7}$$

With the scales in (7), the plateau occupies $[-L, -\bar{\alpha}/8]$, with height $\chi_\star$. The oracle adds independent noise equal to $+H$ or $-H$, each with probability δ/T, and zero otherwise. Its noise variance is one.

With probability $\Omega(\delta)$, there is exactly one positive spike in the first half of the run and no other nonzero noise before T. That spike moves the iterate left into the plateau and raises the accumulator to

order δ^{-1}. Because $\beta_2 = 1 - 1/T$, its contribution persists throughout the horizon. Later steps follow the positive true gradient, but their leftward displacements are $\mathcal{O}(\bar{\alpha}\delta^{1/4}T^{-3/4})$. The width L in (7) keeps a constant fraction of the iterates on the plateau. Their squared gradients have size $(\delta T)^{-1/2}$, giving the confidence obstruction. Figure 1 shows both effects; Proposition B.1 and Appendix E give the complete construction.

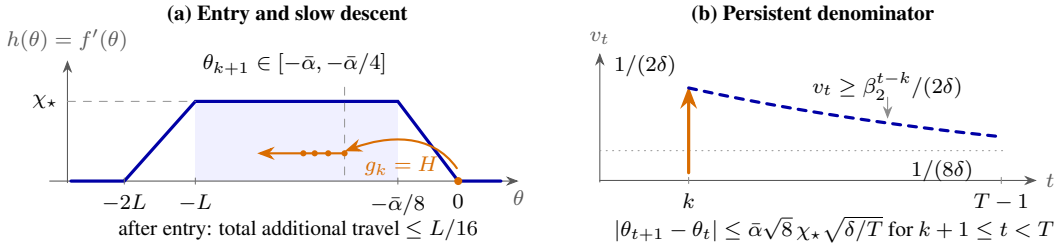


Figure 1: Confidence obstruction for $\beta_1 = 0$, with $\chi_\star, H, L$ from (7). On the event of one positive spike at time k and no other nonzero noise, Adam enters the shaded plateau and drifts slowly left. The dashed curve is a lower envelope of v_t, which bounds the subsequent travel; both panels are schematic.

Hard instance for expectation when $1 \leq p < 2$. Proposition 4.2 uses a fixed class containing a deterministic quadratic and a countable staircase family. For $f_{\mathrm{q}}(x) = x^2/2$, started at 1, each early displacement is at most α/ϵ. Thus a long initial segment stays in $[1/2, 1]$, with squared gradient at least $1/4$. This yields the $T^{-1/3}$ lower bound when $\bar{\alpha}_T$ is below a fixed multiple of $T^{-1/6}$.

For larger prefactors, use a family indexed by integers $n \geq 1$, started at $\theta_1 = 0$. Its continuous staircase derivative has levels that double across blocks of width two, with

$$h_n := f_n', \qquad h_n(x) = 2^{n+1} \quad (x \geq 2n). \tag{8}$$

At a point with $h = h_n(x) \geq 2$, the oracle returns $-h$ with probability $1 - h^{-1}$, and $2h^2 - h$ with probability h^{-1}. Its mean is h, and its second moment is $4h^3 - 3h^2$. These instances share the generalized ABC parameters in Proposition 4.2.

For h_n in (8), let $\widehat{\tau}_n$ be the first index at which the iterate is at or beyond $2n$ on the deterministic path that always selects the negative branch. The stochastic run follows this path on the event

$$E_n := \{\mathbf{g}_t = -h_n(\theta_t) \text{ for every } 1 \leq t < \widehat{\tau}_n\}. \tag{9}$$

On E_n from (9), negative momentum drives the iterate right, against the positive true gradient. Bounding the time spent in each block controls the accumulator and places an arrival before T. Using the terminal height in (8) and the event in (9), the single arrival term gives

$$\frac{1}{T}\sum_{t=1}^{T-1} \mathbb{E}|f_n'(\theta_t)|^2 \geq \frac{4^{n+1}}{T}\,\mathbb{P}(E_n).$$

Choosing n balances the growing gradient against the probability cost of this path; with the quadratic case, this yields the uniform $T^{-1/3}$ lower bound over deterministic prefactors (Figure 4 and Appendix F).

6 Conclusion

Adam admits high-probability convergence guarantees under L_0–L_p generalized smoothness and generalized ABC second-moment growth. Self-normalization and stopped augmented energy control the trajectory, while the final de-preconditioning step gives the $\delta^{-1/2}$ confidence dependence. The expectation theorem for $p < 1$ and the fixed-class lower bound for $1 \leq p < 2$ distinguish the two regimes for the calibrated Adam family.

Acknowledgments

This research is supported by the InnoHK funding.

References

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2): 165–214, 2023.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations (ICLR)*, 2019.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5396–5427, 2023.

Difei Cheng, Ruinan Jin, and Bo Zhang. Convergence analysis of the last iterate in distributed stochastic gradient descent with momentum. *Neurocomputing*, 682:133413, 2026.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized SignSGD. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pages 4883–4895. Curran Associates, Inc., 2021.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 313–355. PMLR, 2022.

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

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

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053. Curran Associates, Inc., 2020.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations (ICLR)*, 2025.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209. PMLR, 2019.

Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of AdaGrad over SGD: Upper and lower bounds in stochastic non-convex optimization. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3124–3158, 2025.

Ruinan Jin, Xiaoyu Wang, and Baoxiang Wang. Stability and convergence analysis of AdaGrad for non-convex optimization via novel stopping time-based techniques. *arXiv preprint arXiv:2409.05023*, 2024.

Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing the convergence of Adam: Bridging the gap with SGD. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 27979–28030, 2025.

Ruinan Jin, Yingbin Liang, and Shaofeng Zou. Why Adam can beat SGD: Second-moment normalization yields sharper tails. *arXiv preprint arXiv:2603.03099*, 2026.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17343–17363. PMLR, 2023.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, volume 36, pages 40238–40271, 2023a.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Huan Li, Yiming Dong, and Zhouchen Lin. On the $\mathcal{O}(\sqrt{d}/K^{1/4})$ convergence rate of AdamW measured by ℓ_1 norm. In *Advances in Neural Information Processing Systems*, volume 38, pages 132360–132387, 2025.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 2019.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy L. Nguyen. High probability convergence of stochastic gradient methods. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 21884–21914, 2023.

Liam Madden, Emiliano Dall'Anese, and Stephen Becker. High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36, 2024.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. In *Advances in Neural Information Processing Systems*, volume 36, pages 24191–24222. Curran Associates, Inc., 2023.

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

Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR, 2023.

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. In *Advances in Neural Information Processing Systems*, volume 36, pages 39006–39032, 2023a.

Bohan Wang, Huishuai Zhang, Zhi-Ming Ma, and Wei Chen. Convergence of AdaGrad for non-convex objectives: Simple proofs and relaxed assumptions. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195, pages 161–190, 2023b.

Bohan Wang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, and Wei Chen. On the convergence of Adam under non-uniform smoothness: Separability from SGDM and beyond. *arXiv preprint arXiv:2403.15146*, 2024a.

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*, pages 2960–2969, 2024b.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2020.

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for RMSProp and Adam in generalized-smooth non-convex optimization with affine noise variance. *Transactions on Machine Learning Research*, 2025.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, volume 35, pages 28386–28399, 2022.

Contents

1 Introduction **2**

2 Related Work **3**

3 Preliminaries **3**
3.1 Adaptive Moment Estimation (Adam) 3
3.2 Analytic Assumptions 4

4 Results **5**
4.1 Rate obstructions 6

5 Proof Sketch **7**
5.1 Upper Bounds 7
5.2 Hard Instance 9

6 Conclusion **10**

Acknowledgments **10**

References **11**

A Extended Related Work **15**
A.1 Adam and adaptive-gradient convergence 15
A.2 Generalized smoothness and second-moment growth 15
A.3 High-probability guarantees under weak moments 16
A.4 Self-normalization and comparison scopes 16

B Formal Results **16**
B.1 Parameter scales and convergence bounds 17
B.2 Lower bounds 19

C Gradient growth and martingale tools **20**

D Proof of Theorem B.1 **22**
D.1 Deterministic normalized estimates 23
D.2 The localized augmented-energy inequality 24
D.3 Uniform bounds for the stopped gradients 27
D.4 Four stopped concentration estimates 28
D.5 Closing the stopped descent 29
D.6 De-preconditioning 31
D.7 Proof of Claim B.1 32
D.8 Proof of Theorem B.2 32

E Proof of Proposition B.1 **33**

F Proof of Proposition B.2 **35**

NeurIPS Paper Checklist **39**

A Extended Related Work

A.1 Adam and adaptive-gradient convergence

Adam combines momentum with coordinatewise second-moment normalization [Kingma and Ba, 2015]. Reddi et al. [2018] identify convergence failures and introduce AMSGrad, while Chen et al. [2019] give convergence conditions for a class of Adam-type methods. Défossez et al. [2022] analyze Adam and AdaGrad under bounded stochastic gradients. Zhang et al. [2022] study Adam without changing its update rule, including the role of its parameter choices and sampling scheme. Wang et al. [2023a] establish finite-time complexity bounds under bounded variance, and Jin et al. [2025] develop an ABC-based framework for finite-time and asymptotic Adam convergence. These analyses address how the adaptive denominator interacts with the gradient used in the same update. Last-iterate gradient convergence has also been studied for distributed stochastic gradient descent with momentum [Cheng et al., 2026].

Li et al. [2025] prove that AdamW satisfies $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_1 = \mathcal{O}(\sqrt{d}\,T^{-1/4})$ with no logarithmic factor. Their analysis assumes global smoothness and coordinatewise bounded conditional variance. These assumptions are much stronger than Assumptions 3.2–3.3: they exclude curvature growth, and they bound the variance by a constant instead of letting the second moment grow with powers of the function gap and of the gradient norm. Their problem class is therefore contained in the case $p = 0$, $A = 0$, $B = 1$, $\rho_2 = 2$ of the present setting, and their criterion is the expected ℓ_1 norm. For these reasons a comparison of the two results has limited meaning; the only common quantity is the exponent of T on that subclass. There, $\|\nabla f(\boldsymbol{\theta}_t)\|_1 \le \sqrt{d}\,\|\nabla f(\boldsymbol{\theta}_t)\|$ and the expected-norm bound obtained above from Claim 4.1 by Cauchy–Schwarz give

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|_1 = \mathcal{O}(dT^{-1/4} + d^{3/2}T^{-1/2}).$$

This recovers the exponent $T^{-1/4}$, with a dimension factor larger by $\sqrt{d}$ than in Li et al. [2025] and with constants that are not optimized. The logarithms in Theorem 4.1 depend only on $1/\delta$ and are integrated out in Claim 4.1, so this specialization carries no $\log T$ factor.

AdaGrad provides a complementary perspective on adaptive scaling. Duchi et al. [2011] develop adaptive subgradient methods for online learning and stochastic optimization. In nonconvex optimization, Li and Orabona [2019] study adaptive stepsizes, and Ward et al. [2020] give convergence guarantees for AdaGrad-Norm. Faw et al. [2022] allow unbounded gradients and affine variance; Wang et al. [2023b] use potential arguments under relaxed stochastic assumptions. Jin et al. [2024] apply stopping-time techniques to AdaGrad under global smoothness, with distinct assumptions for their finite-time and asymptotic conclusions. Jiang et al. [2025] compare coordinatewise AdaGrad and SGD through upper and lower complexity bounds with coordinatewise smoothness and noise parameters. Scalar normalization, coordinatewise accumulation, and exponential moving averages produce different denominator recursions, so their cancellation and memory estimates enter the analyses differently.

A.2 Generalized smoothness and second-moment growth

Zhang et al. [2020] motivate gradient-dependent smoothness through the behavior of gradient clipping. Crawshaw et al. [2022] study generalized SignSGD under unbounded smoothness. Chen et al. [2023] develop a generalized-smooth nonconvex framework, while Li et al. [2023a] analyze convex and nonconvex optimization under generalized smoothness. Gorbunov et al. [2025] study clipping, acceleration, and adaptivity for convex (L_0, L_1)-smooth objectives. The distinctions among local gradient inequalities, coordinatewise conditions, and Hessian growth bounds matter when transferring a convergence guarantee between these settings.

For Adam, Li et al. [2023b] establish guarantees under relaxed smoothness with bounded or sub-Gaussian noise and identify bounded variance as an open direction. Hong and Lin [2024] analyze Adam under relaxed smoothness and an exponential-moment noise envelope. Wang et al. [2024b] study random-reshuffling Adam for finite sums under non-uniform smoothness and growth conditions. Wang et al. [2024a] analyze scalar-normalized Adam under non-uniform smoothness and affine noise variance, including comparisons with momentum SGD. Zhang et al. [2025] obtain RMSProp and

Adam guarantees under coordinatewise generalized smoothness and coordinatewise affine second-moment growth, with the regularizer inside the square root. Here Assumption 3.2 is a Euclidean local gradient inequality, Assumption 3.3 allows powers of the function gap and gradient norm, and Algorithm 1 uses $\sqrt{\mathbf{v}_t} + \epsilon$.

The ABC condition belongs to a broader family of moment-growth models. Gower et al. [2019] use expected smoothness to analyze SGD under general sampling. Khaled and Richtárik [2023] develop a nonconvex second-moment model combining the function gap, squared gradient norm, and a constant noise term. The usual linear ABC form is contained in Assumption 3.3 by taking $\rho_1 = 1$ and $\rho_2 = 2$. Allowing other exponents accommodates nonlinear growth of the oracle's second moment while retaining conditional unbiasedness. This condition controls second moments without requiring almost-sure boundedness or an exponential-moment bound.

A.3 High-probability guarantees under weak moments

Ghadimi and Lan [2013] establish stochastic first-order stationarity guarantees for smooth nonconvex objectives. High-probability bounds depend on both the noise model and the algorithm: Liu et al. [2023] analyze stochastic gradient methods under sub-Gaussian noise, and Madden et al. [2024] treat norm sub-Weibull noise. Martingale inequalities such as Freedman [1975] control sums through conditional variance and bounded increments. In adaptive analyses, these tools can be applied to normalized quantities even when raw oracle values are unbounded.

Clipping gives another route under weak moment assumptions. Gorbunov et al. [2020] obtain accelerated clipped methods for convex stochastic optimization with heavy-tailed noise. Cutkosky and Mehta [2021] give high-probability nonconvex guarantees using clipped normalized momentum under moment bounds on the stochastic gradient. Nguyen et al. [2023] analyze clipped methods under centered heavy-tailed noise, and Sadiev et al. [2023] treat stochastic optimization and variational inequalities with bounded noise moments of order between one and two. Koloskova et al. [2023] characterize stochastic clipping bias and establish convergence guarantees and lower bounds for clipped SGD. Their mechanisms and assumptions differ from the unclipped, same-sample coordinatewise normalization in Algorithm 1.

A.4 Self-normalization and comparison scopes

Jin et al. [2026] connect Adam's second-moment normalization to high-probability stationarity under classical smoothness and bounded conditional variance. Their stopping-time and de-preconditioning viewpoint separates control of the normalized trajectory from control of the accumulator used in the final gradient conversion. For generalized smoothness and ABC growth, the localized curvature and noise envelope depend on the objective height. The augmented energy in the present analysis pairs the true-gradient contribution with a telescope before bounding the centered noise, yielding the explicit squared-stationarity estimate in Theorem B.1.

Lower bounds must be compared at the same oracle model, stationarity measure, and algorithmic scope. Arjevani et al. [2023] establish oracle-complexity lower bounds for smooth nonconvex stochastic optimization under bounded-variance and related oracle models. Proposition B.1 concerns the confidence factor for the calibrated Adam family with horizon-length memory. Proposition B.2 concerns expected averaged squared gradients on a fixed generalized ABC class with $\rho_2 = 3$, uniformly over deterministic base stepsize prefactors. Its benchmark is the square-root-horizon expectation rate in Theorem B.2; its quantifiers concern this Adam family rather than all stochastic first-order algorithms.

B Formal Results

This appendix gives the complete statements corresponding to Section 4. We use the auxiliary objective and sequence in (1)–(2).

B.1 Parameter scales and convergence bounds

The two gradient-growth comparison quantities used throughout are

$$K = \max\left\{(4L_p + 1)^{\max\{1,(2-p)^{-1}\}}, \sqrt{4L_0}\right\}, \tag{10}$$

$$C_{\mathrm{q}} = 1 + \frac{K}{L_p} + \frac{L_0 + L_p K^p}{2L_p^2}. \tag{11}$$

Both depend only on L_0, L_p, p. The notation $[z]_+$ means $\max\{z, 0\}$.

With Ψ from (1), the constants L_0, L_p of Assumption 3.2 and C of Assumption 3.3, the fixed parameter scale used in the high-probability analysis is

$$\begin{aligned}\mathcal{P} = 3 &+ \frac{7\beta_1}{1-\beta_1} + \frac{\beta_1^2}{(1-\beta_1)^2} \\ &+ \Psi(\boldsymbol{\theta}_1) + \|\nabla f(\boldsymbol{\theta}_1)\| + L_0 + L_p + C + \sqrt{C} \\ &+ \epsilon + v_{\mathrm{init}}^{-1} + v_{\mathrm{init}}^{-3/2}.\end{aligned} \tag{12}$$

For an auxiliary-objective height $U \geq 1$, use K and C_{q} from (10)–(11) together with A, B, ρ_1, ρ_2 of Assumption 3.3, to define the gradient radius and the growing part of the noise variance by

$$\begin{aligned}G_{\mathrm{q}}(U) &= K C_{\mathrm{q}}^{\max\{1,(2-p)^{-1}\}} U^{\max\{1,p/(2-p)^2\}}, \\ \mathcal{V}_{\mathrm{q}}(U) &= A C_{\mathrm{q}}^{\rho_1} U^{\rho_1 \max\{1,p/(2-p)\}} + \sup_{0\leq z\leq G_{\mathrm{q}}(U)} [Bz^{\rho_2} - z^2]_+.\end{aligned} \tag{13}$$

Using $\mathcal{P}$ from (12) and the local scales in (13), define the calibration scale

$$\begin{aligned}\mathcal{R}(U) = \mathcal{P} &+ L_p[G_{\mathrm{q}}(U)^p - 1]_+ + \mathcal{V}_{\mathrm{q}}(U) + \sqrt{\mathcal{V}_{\mathrm{q}}(U)} \\ &+ \sqrt{\frac{2\mathbf{1}_{\{p>0\}} G_{\mathrm{q}}(U)^2}{U+1}}.\end{aligned} \tag{14}$$

These definitions display all parameter dependence. The scale $\mathcal{P}$ is fixed, while $G_{\mathrm{q}}(U), \mathcal{V}_{\mathrm{q}}(U), \mathcal{R}(U)$ also depend on the height U. None depends on the horizon or confidence except through that height. In the proof, $C + \mathcal{V}_{\mathrm{q}}(U)$ bounds the conditional noise variance before exit (the exit index σ_U is defined in (32)).

Table 1: Quantities used in the upper-bound analysis.

Symbol	Role	Definition
Ψ	shifted objective	(1)
$\mathbf{z}_t$	auxiliary sequence	(2)
$\lambda_{t,i}$	coordinatewise stepsize	Algorithm 1
$K,\ C_{\mathrm{q}}$	gradient-growth comparison	(10)–(11)
$\mathcal{P}$	fixed parameter scale	(12)
$G_{\mathrm{q}}(U),\ \mathcal{V}_{\mathrm{q}}(U)$	gradient radius, growing noise variance	(13)
$\mathcal{R}(U)$	calibration scale	(14)
$\mathcal{X}$	domination scale	(34)
U_δ	confidence height	(15)
η_{gen}	stepsize excess factor	(3)
$\Delta_{t,i}$	stepsize variation	(26)
σ_U	exit time above height U	(32)
$E_t,\ \mathcal{L}_t$	gradient energy, Lyapunov process	(33)
Z_n	stopped logarithmic energy	(30)
τ	logarithmic-energy crossing time	(47)

Theorem B.1 (High-probability convergence)**.** *Assume Assumptions 3.1–3.3, and use the scales in* (12)–(14)*. Let $T \geq 10$, $\beta_2 = 1 - 1/T$, and $\alpha = \bar{\alpha}/\sqrt{T}$ in Algorithm 1. For $0 < \delta < 1$, with η_{gen} from* (3)*, set*

$$U_\delta = 2^{20}\mathcal{P}^8(1 + \log(16/\delta))^{4\mathbf{1}_{\{\eta_{\mathrm{gen}}=0\}}}. \tag{15}$$

With $\mathcal{R}$ defined in (14)*, for every*

$$0 < \bar{\alpha} \le 2^{-100} d^{-1} (1 + \log(16/\delta))^{-\mathbf{1}_{\{\eta_{\mathrm{gen}} > 0\}}} \mathcal{R}(U_\delta)^{-48}, \tag{16}$$

with probability at least $1 - \delta$, where Ψ and $\mathbf{z}_t$ are defined in (1)–(2)*,*

$$\begin{aligned} \max_{1 \le t \le T} \Psi(\mathbf{z}_t) &\le U_\delta, \\ \frac{1}{T} \sum_{t=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_t)\|^2 &\le \frac{8 U_\delta (\epsilon + \sqrt{v_{\mathrm{init}}})}{\bar{\alpha}\sqrt{T}} + \frac{32 U_\delta^2}{\bar{\alpha}^2 T} \\ &\quad + \frac{16 U_\delta \sqrt{C + \mathcal{V}_{\mathrm{q}}(U_\delta)}}{\bar{\alpha}\sqrt{\delta T}}. \end{aligned} \tag{17}$$

In particular, the upper endpoint of the displayed stepsize interval gives $\widetilde{\mathcal{O}}(d(\delta T)^{-1/2} + d^2 T^{-1})$ stationarity, where the suppressed factors are powers of $1 + \log(1/\delta)$. If $\eta_{\mathrm{gen}} > 0$, then, by (15)*, $U_\delta = 2^{20}\mathcal{P}^8$ does not depend on δ, so the localization bound is $\mathcal{O}(1)$, and δ enters* (16) *only through the single factor $1 + \log(16/\delta)$.*

At the upper endpoint of (16) with $\eta_{\mathrm{gen}} = 0$ (see (3)), the three coefficients in (17), with $\mathcal{P}$ from (12), are $2^{123} d\,\mathcal{P}^{56}(1 + \log(16/\delta))^4(\epsilon + \sqrt{v_{\mathrm{init}}})$, $2^{245} d^2 \mathcal{P}^{112}(1 + \log(16/\delta))^8$ and $2^{124} d\,\mathcal{P}^{56}(1 + \log(16/\delta))^4\sqrt{C}$; the second term is dominated by the first once $T \ge 2^{244} d^2 \mathcal{P}^{112}(1+\log(16/\delta))^8(\epsilon + \sqrt{v_{\mathrm{init}}})^{-2}$. When $\eta_{\mathrm{gen}} > 0$, the factor $\mathcal{P}^{48}(1 + \log(16/\delta))^4$ in the first and third coefficients, and its square in the second coefficient and in this threshold, are replaced by $\mathcal{R}(2^{20}\mathcal{P}^8)^{48}(1 + \log(16/\delta))$ and its square, with $\mathcal{R}$ from (14) evaluated at the height $2^{20}\mathcal{P}^8$ of (15), and $\sqrt{C}$ by $\sqrt{C + \mathcal{V}_{\mathrm{q}}(2^{20}\mathcal{P}^8)}$. These constants are not optimized. The quantities U_δ and $\mathcal{V}_{\mathrm{q}}$ in (17) are defined in (15) and (13), respectively.

Claim B.1 (Smooth bounded-variance specialization)**.** *Under the assumptions of Theorem B.1, let $p = 0, A = 0, B = 1, \rho_2 = 2$ in Assumptions 3.2 and 3.3, and let $\mathcal{P}$ be as in* (12)*. For every fixed $0 < \bar{\alpha} \le 2^{-100} d^{-1} \mathcal{P}^{-48}$,*

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 &\le \frac{2^{23}\mathcal{P}^8}{\bar{\alpha}\sqrt{T}} \sum_{j=0}^{4} \binom{4}{j} (1 + \log 16)^{4-j} j! \big(\epsilon + \sqrt{v_{\mathrm{init}}} + 2^{j+2}\sqrt{C}\big) \\ &\quad + \frac{2^{45}\mathcal{P}^{16}}{\bar{\alpha}^2 T} \sum_{j=0}^{8} \binom{8}{j} (1 + \log 16)^{8-j} j!. \end{aligned} \tag{18}$$

In particular, at $\bar{\alpha} = 2^{-100} d^{-1} \mathcal{P}^{-48}$ the left-hand side is $\mathcal{O}(dT^{-1/2} + d^2 T^{-1})$.

Here $\eta_{\mathrm{gen}} = 0$ in (3), $\mathcal{V}_{\mathrm{q}}(U) = 0$ and $\mathcal{R}(U) = \mathcal{P}$ for all $U \ge 1$, by (13)–(14). Thus the same run satisfies (17), with U_δ from (15), at every confidence level, and integrating that bound gives the claim.

Theorem B.2 (Expected convergence for $p < 1$)**.** *Assume Assumptions 3.1–3.3 with $p < 1$, and let $T \ge 10$, $\beta_2 = 1 - 1/T$ in Algorithm 1. Choose*

$$\begin{aligned} A_{\mathrm{tail}} > \frac{1}{2} + \frac{2}{1-p}, \qquad \delta_T = T^{-A_{\mathrm{tail}}}, \qquad \alpha = \bar{\alpha}_T/\sqrt{T}, \\ \bar{\alpha}_T = 2^{-100} d^{-1} (1 + \log(16/\delta_T))^{-\mathbf{1}_{\{\eta_{\mathrm{gen}} > 0\}}} \mathcal{R}(U_{\delta_T})^{-48}, \end{aligned} \tag{19}$$

where η_{gen}, $\mathcal{R}$ and U_δ are defined in (3)*,* (14) *and* (15)*. Then*

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T-1} \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 &\le \frac{8(\epsilon + \sqrt{v_{\mathrm{init}}})}{\bar{\alpha}_T\sqrt{T}} \int_{\delta_T}^{1} U_\delta \, d\delta + \frac{32}{\bar{\alpha}_T^2 T} \int_{\delta_T}^{1} U_\delta^2 \, d\delta \\ &\quad + \frac{16}{\bar{\alpha}_T\sqrt{T}} \int_{\delta_T}^{1} \frac{U_\delta \sqrt{C + \mathcal{V}_{\mathrm{q}}(U_\delta)}}{\sqrt{\delta}} \, d\delta \\ &\quad + \delta_T (1 + 2T)^{2/(1-p)} \\ &\qquad \times \big[(1 + \|\nabla f(\boldsymbol{\theta}_1)\|)^{1-p} + (1-p)(L_0 + L_p)\big]^{2/(1-p)}. \end{aligned} \tag{20}$$

The integrands use only U_δ and $\mathcal{V}_{\mathrm{q}}$ from (15) *and* (13)*; the cutoff and prefactor are specified by* (19)*. In particular,*

$$\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 = \begin{cases} \mathcal{O}\left(\dfrac{d(1+\log T)}{\sqrt{T}} + \dfrac{d^2(1+\log T)^2}{T}\right), & \eta_{\mathrm{gen}} > 0, \\ \mathcal{O}\left(\dfrac{d}{\sqrt{T}} + \dfrac{d^2}{T}\right), & \eta_{\mathrm{gen}} = 0. \end{cases}$$

The horizon enters these orders only through $\bar{\alpha}_T^{-1}$. The integrals in (20) are at most their values over $(0,1]$, which are finite and independent of T, and the last term is $o(T^{-1/2})$ because $A_{\mathrm{tail}} > 1/2 + 2/(1-p)$. By (19), $\bar{\alpha}_T^{-1}$ is linear in $1+\log(16/\delta_T) = 1+\log 16 + A_{\mathrm{tail}}\log T$ when $\eta_{\mathrm{gen}} > 0$ (see (3)), whereas $\bar{\alpha}_T^{-1} = 2^{100} d\,\mathcal{P}^{48}$, with $\mathcal{P}$ from (12), does not depend on T when $\eta_{\mathrm{gen}} = 0$.

B.2 Lower bounds

Proposition B.1 (Sharpness of the $\delta^{-1/2}$ dependence)**.** *Fix $\bar{\alpha} \in (0,1]$, $v_{\mathrm{init}} > 0$, and $\epsilon > 0$. For every integer $T \geq 4$ and every*

$$0 < \delta \leq \min\left\{\frac{1}{8}, \frac{1}{8(v_{\mathrm{init}} + \epsilon^2)}\right\}, \qquad \delta T \geq 1,$$

there is a one-dimensional lower-bounded C^1 objective and an unbiased oracle satisfying Assumptions 3.1–3.3 with $p = 0$, $A = 0$, $B = 1$, $\rho_1 = 0$, $\rho_2 = 2$, $C = 1$, and smoothness constants independent of T, δ, such that

$$\theta_1 = 0, \qquad f'(\theta_1) = 0, \qquad f(\theta_1) - f^* \leq 32\sqrt{8}\,\bar{\alpha},$$

and Adam (Algorithm 1) with $\beta_1 = 0$, $\beta_2 = 1 - 1/T$, and $\alpha = \bar{\alpha}/\sqrt{T}$ satisfies

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T-1}|\nabla f(\boldsymbol{\theta}_t)|^2 \geq \frac{1}{4\sqrt{\delta T}}\right) \geq \frac{\delta}{4}.$$

Whenever $\delta T \geq (8/\bar{\alpha})^4$, the same construction (the objective in (53) *and the oracle noise in* (54)*, with scales from* (52)*) satisfies the common smoothness constants $L_0 = L_p = 1$.*

The polynomial confidence obstruction holds within this common classical parameter class, including prefactors admitted by Theorem B.1. Indeed, each fixed

$$0 < \bar{\alpha} \leq 2^{-100}\left(100 + \epsilon + v_{\mathrm{init}}^{-1} + v_{\mathrm{init}}^{-3/2}\right)^{-48}$$

satisfies its classical stepsize condition (16) for every constructed member with $\delta T \geq (8/\bar{\alpha})^4$. Reparameterizing the construction of Proposition B.1 by $\delta = 8\eta$ gives, for sufficiently small $\eta > 0$ and sufficiently large horizons,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T-1}|f'(\theta_t)|^2 \geq \frac{1}{8\sqrt{2}\sqrt{\eta T}}\right) \geq 2\eta.$$

Thus a uniform $\mathrm{polylog}(1/\eta)/\sqrt{T}$ guarantee at confidence $1-\eta$ is impossible for these fixed calibrated prefactors. The comparison concerns the same averaged squared-gradient statistic and the $\beta_1 = 0$ subfamily of Algorithm 1.

Proposition B.2 (A lower bound on expected stationarity for $1 \leq p < 2$)**.** *Fix $p \in [1,2)$, $\beta_1 \in [0,1)$, $v_{\mathrm{init}} > 0$, and $\epsilon > 0$. There exists a fixed class $\mathcal{C}_p$ of one-dimensional objectives, stochastic-gradient oracles, and initial points satisfying Assumptions 3.1–3.3 with*

$$L_0 = L_p = 1, \qquad A = 0, \quad B = 4, \quad C = 1, \quad \rho_1 = 0, \quad \rho_2 = 3,$$

and

$$f^* = 0, \qquad |\theta_1| \leq 1, \qquad |f'(\theta_1)| \leq 2, \qquad f(\theta_1) - f^* \leq 2,$$

such that Adam (Algorithm 1) with $\beta_2 = 1 - 1/T$ and $\alpha = \bar{\alpha}_T/\sqrt{T}$ satisfies, for every sufficiently large integer T,

$$\inf_{\bar{\alpha}_T>0} \sup_{(f,\mathsf{G},\theta_1)\in\mathcal{C}_p} \frac{1}{T}\sum_{t=1}^{T-1} \mathbb{E}|f'(\theta_t)|^2 \geq \frac{\epsilon(1-\beta_1)^2}{2^{19}(1+\sqrt{v_{\mathrm{init}}}+\epsilon)^2}\, T^{-1/3}. \tag{21}$$

The class, whose members are given in (56)*,* (58)*, and* (59)*, is independent of T and $\bar{\alpha}_T$; the horizon threshold depends only on $\beta_1, v_{\mathrm{init}}, \epsilon$.*

The infimum in Proposition B.2 is over deterministic scaled base stepsizes for the stated Adam recursion. Its stationarity measure is the same expected average of squared true-gradient norms as in Theorem B.2. For every fixed $k \geq 0$,

$$\frac{T^{-1/3}}{\log^k(eT)/\sqrt{T}} = \frac{T^{1/6}}{\log^k(eT)} \longrightarrow \infty.$$

Consequently, no uniform bound of order $\log^k(eT)/\sqrt{T}$ holds for this Adam family on the full class in the proposition. The comparison is with the square-root-horizon expectation rate available when $p < 1$ (see (20)); the lower bound concerns the Adam family and the generalized ABC class containing $\rho_2 = 3$.

C Gradient growth and martingale tools

Throughout this section, Ψ is the auxiliary objective in (1).

Lemma C.1 (Gradient growth from the function gap)**.** *Under Assumptions 3.1–3.2, with K defined in* (10)*,*

$$\|\nabla f(x)\| \leq K\Psi(x)^{\max\{1,(2-p)^{-1}\}} \qquad (x \in \mathbb{R}^d). \tag{22}$$

Proof. The conclusion is immediate when $\nabla f(x) = 0$. Otherwise take

$$y = x - \frac{\nabla f(x)}{L_0 + L_p \max\{\|\nabla f(x)\|^p, \|\nabla f(x)\|\}}.$$

This displacement is at most $1/L_p$. Integrating the local gradient inequality of Assumption 3.2 along the segment from x to y gives

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_0 + L_p\|\nabla f(x)\|^p}{2}\|y - x\|^2.$$

The chosen reciprocal stepsize is at least the local smoothness coefficient, so lower boundedness (Assumption 3.1) implies

$$f(x) - f^* \geq \frac{\|\nabla f(x)\|^2}{2[L_0 + L_p \max\{\|\nabla f(x)\|^p, \|\nabla f(x)\|\}]}.$$

Either $\|\nabla f(x)\|^2 \leq 4L_0(f(x) - f^*)$, or the last display yields

$$\|\nabla f(x)\|^2 \leq 4L_p(f(x) - f^*)\max\{\|\nabla f(x)\|^p, \|\nabla f(x)\|\}.$$

Distinguishing the two entries of this maximum proves

$$\|\nabla f(x)\| \leq \max\left\{\sqrt{4L_0(f(x) - f^*)},\, [4L_p(f(x) - f^*)]^{1/(2-p)},\, 4L_p(f(x) - f^*)\right\}.$$

Each entry is at most $K\Psi(x)^{\max\{1,(2-p)^{-1}\}}$, by the definition of K in (10) and $\Psi(x) \geq 1$ (see (1)). □

Lemma C.2 (Comparison of nearby function gaps)**.** *Under Assumptions 3.1–3.2, if $\|x - y\| \leq 1/L_p$, then, with Ψ from* (1) *and C_{q} from* (11)*,*

$$\Psi(x) \leq C_{\mathrm{q}}\Psi(y)^{\max\{1,p/(2-p)\}}. \tag{23}$$

Proof. The integrated local gradient inequality of Assumption 3.2 and the gradient bound (22) of Lemma C.1, with K from (10), give

$$\Psi(x) \le \Psi(y) + \frac{K}{L_p}\Psi(y)^{\max\{1,(2-p)^{-1}\}} + \frac{L_0 + L_p K^p}{2L_p^2}\Psi(y)^{p\max\{1,(2-p)^{-1}\}}.$$

Since $\Psi(y) \ge 1$, all three powers are bounded by the power $\max\{1, p/(2-p)\}$. Indeed this exponent is one for $p \le 1$ and is $p/(2-p)$ for $p \ge 1$. Substituting the definition of C_{q} in (11) gives (23). □

Lemma C.3 (Polynomial distance growth for $p < 1$)**.** *Under Assumption 3.2, for $0 \le p < 1$ and any $x, x_{\mathrm{ref}} \in \mathbb{R}^d$,*

$$(1 + \|\nabla f(x)\|)^{1-p} \le (1 + \|\nabla f(x_{\mathrm{ref}})\|)^{1-p} + (1-p)(L_0 + L_p)\|x - x_{\mathrm{ref}}\|. \tag{24}$$

Consequently,

$$\|\nabla f(x)\| \le \left[(1 + \|\nabla f(x_{\mathrm{ref}})\|)^{1-p} + (1-p)(L_0 + L_p)\right]^{1/(1-p)} \times (1 + \|x - x_{\mathrm{ref}}\|)^{1/(1-p)}.$$

Proof. Partition the segment from x_{ref} to x into finitely many subsegments of length at most $1/L_p$, with consecutive endpoints z_{j-1}, z_j. The reverse triangle inequality, combined with Assumption 3.2 on each subsegment, gives

$$\|\nabla f(z_j)\| - \|\nabla f(z_{j-1})\| \le (L_0 + L_p)(1 + \|\nabla f(z_{j-1})\|)^p \|z_j - z_{j-1}\|.$$

The increasing concave function $u \mapsto u^{1-p}$, on $u > 0$, therefore satisfies

$$\begin{aligned}
&(1 + \|\nabla f(z_j)\|)^{1-p} - (1 + \|\nabla f(z_{j-1})\|)^{1-p} \\
&\quad \le (1-p)(1 + \|\nabla f(z_{j-1})\|)^{-p}\big(\|\nabla f(z_j)\| - \|\nabla f(z_{j-1})\|\big) \\
&\quad \le (1-p)(L_0 + L_p)\|z_j - z_{j-1}\|.
\end{aligned}$$

Summing over the partition proves (24). The second bound follows by factoring $1 + \|x - x_{\mathrm{ref}}\|$ and taking the positive power $1/(1-p)$. □

Lemma C.4 (Finite-horizon accumulator comparison)**.** *For Algorithm 1 with $T \ge 10$ and $\beta_2 = 1 - 1/T$, every $0 \le t \le T-1$ satisfies*

$$\mathbf{v}_{t,i} \ge \frac{1}{4}\left(v_{\mathrm{init}} + \frac{1}{T}\sum_{k=1}^{t}\mathbf{g}_{k,i}^2\right).$$

Moreover, the coordinatewise stepsizes $\lambda_{t,i}$ of Algorithm 1 satisfy $\lambda_{t,i} \le 2\lambda_{k,i}$ for $0 \le k \le t \le T-1$.

Proof. Unrolling the second-moment recursion of Algorithm 1 with $\beta_2 = 1 - 1/T$ gives

$$\mathbf{v}_{t,i} = (1 - 1/T)^t v_{\mathrm{init}} + \frac{1}{T}\sum_{k=1}^{t}(1 - 1/T)^{t-k}\mathbf{g}_{k,i}^2.$$

The elementary bound $(1 - 1/T)^T \ge 1/4$ holds for $T \ge 2$: the function $u \mapsto u\log(1 - 1/u)$ is increasing for $u > 1$, since $\log(1 - z) > -z/(1 - z)$ for $0 < z < 1$, and its value at $u = 2$ is $\log(1/4)$. Thus every weight in the display is at least $1/4$. The same recursion started at k gives $\mathbf{v}_{t,i} \ge \mathbf{v}_{k,i}/4$; taking square roots and retaining $\epsilon > 0$ proves the comparison of the stepsizes $\lambda_{t,i}$ defined in Algorithm 1. □

Lemma C.5 (Conditional-to-direct moment comparison)**.** *For nonnegative random variables Z_t adapted to a filtration $\{\mathscr{F}_t\}$ (in the applications, the one of Assumption 3.3) and a finite horizon n, write*

$$Y_n = \sum_{t=1}^{n}\mathbb{E}[Z_t \mid \mathscr{F}_{t-1}].$$

For every real $s \geq 1$,

$$\|Y_n\|_{L^s} \leq s \left\| \sum_{t=1}^{n} Z_t \right\|_{L^s}, \qquad \|X\|_{L^s} := (\mathbb{E}|X|^s)^{1/s}.$$

The same result applies to a predictable truncation by multiplying Z_t by its $\mathscr{F}_{t-1}$-measurable inclusion indicator.

Proof. The case $s = 1$ follows from the tower property. For $s > 1$, fix a nonnegative dual random variable W with $\|W\|_{L^{s/(s-1)}} = 1$, and let $M_t = \mathbb{E}[W \mid \mathscr{F}_t]$. Then

$$\mathbb{E}[WY_n] = \sum_{t=1}^{n} \mathbb{E}[M_{t-1} Z_t] \leq \mathbb{E}\left[\left(\max_{0 \leq t \leq n} M_t \right) \sum_{u=1}^{n} Z_u \right].$$

Hölder's inequality, Doob's maximal inequality, and contraction of conditional expectation give

$$\mathbb{E}[WY_n] \leq \left\| \max_{0 \leq t \leq n} M_t \right\|_{L^{s/(s-1)}} \left\| \sum_{u=1}^{n} Z_u \right\|_{L^s} \leq s \left\| \sum_{u=1}^{n} Z_u \right\|_{L^s}.$$

Taking the supremum over the dual variables proves the claim. If needed, apply the argument first to truncated random variables and pass to the limit by monotone convergence; when the right side is infinite the inequality holds in the extended sense. □

We use the scalar Freedman inequality [Freedman, 1975] in the following form. For bounded real martingale differences D_t (in the applications, with respect to the filtration $\{\mathscr{F}_t\}$ of Assumption 3.3) and $0 < \delta < 1$,

$$\mathbb{P}\left(\sum_{t=1}^{n} D_t > \sqrt{2 \left\| \sum_{t=1}^{n} \mathbb{E}[D_t^2 \mid \mathscr{F}_{t-1}] \right\|_{L^\infty} \log(1/\delta)} + \frac{2}{3} \max_{1 \leq t \leq n} \|D_t\|_{L^\infty} \log(1/\delta) \right) \leq \delta. \tag{25}$$

When both bounds vanish, the martingale is identically zero and the strict upper-tail event is empty.

D Proof of Theorem B.1

Throughout this proof, K, C_{q} are defined in (10)–(11), $\mathcal{P}$ in (12), and $G_{\mathrm{q}}, \mathcal{V}_{\mathrm{q}}, \mathcal{R}$ in (13)–(14). The auxiliary objective Ψ is defined in (1), the auxiliary sequence $\mathbf{z}_t$ in (2), and the effective stepsizes $\lambda_{t,i}$ in Algorithm 1.

Proof framework. Figure 2 records the dependencies among the auxiliary estimates and convergence results. The stopped concentration estimates close the localized energy descent; ordinary-gradient conversion then proves Theorem B.1. Its two expectation consequences use the calibrations and integration arguments in Appendices D.7 and D.8, respectively. For $p < 1$, the distance bound in Lemma C.3 supplies the pathwise tail contribution.

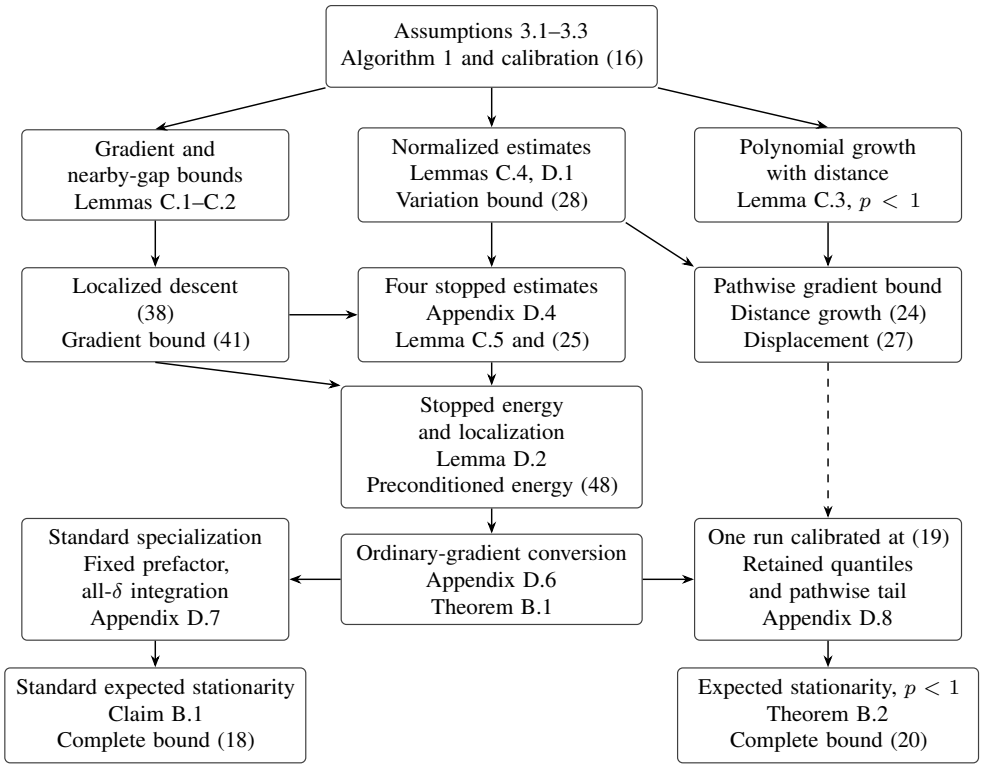


Figure 2: Formal dependencies for Theorems B.1 and B.2 and Claim B.1. Solid arrows connect the localized estimates, stopped concentration, energy closure, and gradient conversion. The standard branch takes $p = 0, A = 0, B = 1, \rho_2 = 2$. The two integration branches use one fixed run under their respective calibrations; the dashed arrow supplies the deterministic contribution of extreme quantiles for $p < 1$.

D.1 Deterministic normalized estimates

With the effective stepsizes $\lambda_{t,i}$ and accumulators $\mathbf{v}_{t,i}$ of Algorithm 1, let

$$\widetilde{\lambda}_{t-1,i} = \frac{\bar{\alpha}}{\sqrt{T-1}(\sqrt{\mathbf{v}_{t-1,i}} + \epsilon)}, \qquad \Delta_{t,i} = \widetilde{\lambda}_{t-1,i} - \lambda_{t,i}. \tag{26}$$

The accumulator recursion gives, for all $t < T$,

$$\mathbf{v}_{t,i} \geq v_{\mathrm{init}}/4, \qquad \lambda_{t,i} \leq \frac{2\bar{\alpha}}{\sqrt{T v_{\mathrm{init}}}}, \qquad |\lambda_{t,i}\mathbf{g}_{t,i}| \leq \bar{\alpha}, \qquad 0 \leq \Delta_{t,i} \leq 2\lambda_{t-1,i}.$$

Moreover, $\lambda_{t,i} \leq 2\lambda_{k,i}$ for $k \leq t < T$. Expanding the momentum recursion therefore yields

$$\|\lambda_t \odot \mathbf{m}_t\| \leq 2\sqrt{d}\,\bar{\alpha}. \tag{27}$$

For $\Delta_{t,i}$ from (26), the exact decomposition

$$\Delta_{t,i} = \lambda_{t-1,i} - \lambda_{t,i} + \bar{\alpha}\left(\frac{1}{\sqrt{T-1}} - \frac{1}{\sqrt{T}}\right)\frac{1}{\sqrt{\mathbf{v}_{t-1,i}} + \epsilon}$$

implies, for every $n \leq T-1$,

$$\begin{aligned}\sum_{t=1}^{n}\sum_{i}\Delta_{t,i} &\leq \frac{d\bar{\alpha}}{\sqrt{T}(\sqrt{v_{\mathrm{init}}} + \epsilon)} + \frac{2d\bar{\alpha}(T-1)}{\sqrt{v_{\mathrm{init}}}}\left(\frac{1}{\sqrt{T-1}} - \frac{1}{\sqrt{T}}\right) \\ &\leq \frac{2d\bar{\alpha}}{\sqrt{T v_{\mathrm{init}}}}.\end{aligned} \tag{28}$$

For the last step use $(T-1)(1/\sqrt{T-1} - 1/\sqrt{T}) < 1/(2\sqrt{T})$.

Lemma D.1 (Stopped normalized energies)**.** *For Algorithm 1 with $T \geq 10$, $\beta_2 = 1 - 1/T$, and $\alpha = \bar{\alpha}/\sqrt{T}$, let σ be a stopping index taking values in $\{1, 2, \ldots\} \cup \{\infty\}$. For every $1 \leq n \leq T$,*

$$\sum_{t<n\wedge\sigma} \left(\|\lambda_t \odot \mathbf{g}_t\|^2 + \|\lambda_t \odot \mathbf{m}_t\|^2 + \|\lambda_{t-1} \odot \mathbf{m}_{t-1}\|^2\right) \leq 132\bar{\alpha}^2 \sum_i \log\left(1 + \frac{\sum_{t<n\wedge\sigma} \mathbf{g}_{t,i}^2}{T v_{\text{init}}}\right). \tag{29}$$

Proof. For the stopping index σ, put

$$Z_n = \sum_i \log\left(1 + \frac{\sum_{t<n\wedge\sigma} \mathbf{g}_{t,i}^2}{T v_{\text{init}}}\right). \tag{30}$$

The accumulator lower bound $\mathbf{v}_{t,i} \geq (v_{\text{init}} + T^{-1}\sum_{k\leq t} \mathbf{g}_{k,i}^2)/4$, followed by the logarithmic sum inequality, gives

$$\sum_{t<n\wedge\sigma} \|\lambda_t \odot \mathbf{g}_t\|^2 \leq 4\bar{\alpha}^2 Z_n,$$
$$\sum_{t<n\wedge\sigma} \|\lambda_t \odot \mathbf{m}_t\|^2, \quad \sum_{t<n\wedge\sigma} \|\lambda_{t-1} \odot \mathbf{m}_{t-1}\|^2 \leq 64\bar{\alpha}^2 Z_n. \tag{31}$$

For the momentum bounds, fix i. Since $\mathbf{m}_0 = \mathbf{0}$, $\mathbf{m}_{t,i} = (1-\beta_1)\sum_{k=1}^t \beta_1^{t-k}\mathbf{g}_{k,i}$, and the weights $(1-\beta_1)\beta_1^{t-k}$ sum to at most one. Jensen's inequality and $\lambda_{t,i} \leq 2\lambda_{k,i}$ for $k \leq t$ give

$$(\lambda_{t,i}\mathbf{m}_{t,i})^2 \leq (1-\beta_1)\sum_{k=1}^t \beta_1^{t-k}(\lambda_{t,i}\mathbf{g}_{k,i})^2 \leq 4(1-\beta_1)\sum_{k=1}^t \beta_1^{t-k}(\lambda_{k,i}\mathbf{g}_{k,i})^2.$$

Sum over $t < n \wedge \sigma$ and i, exchange the sums over t and k, and use $(1-\beta_1)\sum_{t\geq k}\beta_1^{t-k} \leq 1$:

$$\sum_{t<n\wedge\sigma} \|\lambda_t \odot \mathbf{m}_t\|^2 \leq 4 \sum_{k<n\wedge\sigma} \|\lambda_k \odot \mathbf{g}_k\|^2 \leq 16\bar{\alpha}^2 Z_n \leq 64\bar{\alpha}^2 Z_n.$$

Because $\mathbf{m}_0 = \mathbf{0}$, the lagged sum $\sum_{t<n\wedge\sigma} \|\lambda_{t-1} \odot \mathbf{m}_{t-1}\|^2$ consists of the terms of the current sum with $t \leq (n \wedge \sigma) - 2$, so it obeys the same bound. All three bounds hold pathwise, hence also at the random truncation index $n \wedge \sigma$. Adding the three inequalities in (31) gives (29). □

D.2 The localized augmented-energy inequality

Fix $0 < \delta < 1$, put $U = U_\delta$ as defined in (15), and let $\bar{\alpha}$ satisfy (16). With Ψ from (1) and $\mathbf{z}_t$ from (2), define

$$\sigma_U = \inf\{1 \leq t \leq T : \Psi(\mathbf{z}_t) > U\}, \qquad \inf \varnothing = \infty, \tag{32}$$

and write, with $\lambda_{t-1,i}$ from Algorithm 1,

$$h_t = \nabla f(\boldsymbol{\theta}_t), \quad \xi_t = \mathbf{g}_t - h_t, \quad E_t = \sum_i \lambda_{t-1,i} h_{t,i}^2, \quad \mathcal{L}_t = \Psi(\mathbf{z}_t) + E_t. \tag{33}$$

For $\mathcal{R}$ defined in (14), the calibration (16) gives

$$2 \leq \mathcal{R}(U) \leq \mathcal{X}, \qquad \mathcal{X} := (2^{100} d\bar{\alpha}\,(1 + \log(16/\delta))^{\mathbf{1}_{\{\eta_{\text{gen}}>0\}}})^{-1/48}. \tag{34}$$

The domination scale $\mathcal{X}$ of (34) depends only on d, $\bar{\alpha}$ and δ, with the stepsize excess factor η_{gen} defined in (49); the lower bound $\mathcal{X} \geq 2$ follows from $\mathcal{P} \geq 3$ (see (12)) and $\mathcal{R}(U) \geq \mathcal{P}$. Thus every nonnegative summand in (14) is bounded by the same upper bound. For every $0 \leq k \leq 48$,

$$d\bar{\alpha}\,(1 + \log(16/\delta))^{\mathbf{1}_{\{\eta_{\text{gen}}>0\}}}\mathcal{X}^k = 2^{-100}\mathcal{X}^{k-48} \leq 2^{-100}. \tag{35}$$

Since $d \geq 1$ and $1 + \log(16/\delta) \geq 1$, the same bound holds with $d\bar{\alpha}\,(1+\log(16/\delta))^{\mathbf{1}_{\{\eta_{\text{gen}}>0\}}}$ replaced by $d\bar{\alpha}$, by $\sqrt{d}\,\bar{\alpha}$ or by $\bar{\alpha}$.

In both cases $\eta_{\mathrm{gen}} = 0$ and $\eta_{\mathrm{gen}} > 0$, the confidence height (15), $\mathcal{P} \le \mathcal{R}(U)$ and (34) give

$$\begin{aligned}(1+\log(16/\delta))^{\mathbf{1}_{\{\eta_{\mathrm{gen}}=0\}}} &\le U^{1/4},\\ U^{1/4} &\le 2^5\mathcal{P}^2(1+\log(16/\delta))\\ &\le 2^5\mathcal{X}^2(1+\log(16/\delta)).\end{aligned} \tag{36}$$

Consequently, for every $0 \le k \le 48$ and every $0 \le b \le 3/4$,

$$d\bar{\alpha}\,(1+\log(16/\delta))\mathcal{X}^k U^b \le 2^{-100}U. \tag{37}$$

Indeed, if $\eta_{\mathrm{gen}} > 0$, then (37) is (35) multiplied by $U^b \le U$. If $\eta_{\mathrm{gen}} = 0$, then (35) holds with $d\bar{\alpha}$ alone, and the first line of (36) gives $(1+\log(16/\delta))U^b \le U^{b+1/4} \le U$. In the closing steps below, the factor $1+\log(16/\delta)$ appears with power at most one and is absorbed through (37).

The stepsize (16) and the domination (34) ensure all local-distance requirements, because

$$(1+6\frac{\beta_1}{1-\beta_1})\sqrt{d}L_p\bar{\alpha} \le \mathcal{X}^2 d\bar{\alpha} < 1.$$

For the displacement of $\boldsymbol{\theta}_t$, the momentum expansion more precisely gives $\|\lambda_t \odot \mathbf{m}_t\| \le (1+\beta_1)\sqrt{d}\,\bar{\alpha} \le (1+6\frac{\beta_1}{1-\beta_1})\sqrt{d}\,\bar{\alpha}$; this also covers small values of β_1. The displacements of $\mathbf{z}_t$ (defined in (2)) and the distance between $\mathbf{z}_t$ and $\boldsymbol{\theta}_t$ satisfy the same local-radius requirement. In particular, before σ_U (see (32)), Lemmas C.2 and C.1, with the definitions (10)–(11) and (13), give

$$\Psi(\boldsymbol{\theta}_t) \le C_{\mathrm{q}}U^{\max\{1,p/(2-p)\}}, \qquad \|h_t\|, \|\nabla f(\mathbf{z}_t)\| \le G_{\mathrm{q}}(U).$$

With G_{q} and $\mathcal{R}$ defined in (13)–(14), the corresponding smoothness coefficient is at most

$$L_0 + L_p \max\{1, G_{\mathrm{q}}(U)^p\} \le \mathcal{R}(U) \le \mathcal{X}.$$

Conditional unbiasedness and the variance envelope in (13) give, for the noise ξ_t from (33),

$$\begin{aligned}\mathbb{E}_{t-1}\|\xi_t\|^2 &= \mathbb{E}_{t-1}\|\mathbf{g}_t\|^2 - \|h_t\|^2\\ &\le C + \mathcal{V}_{\mathrm{q}}(U) \le \mathcal{R}(U) \le \mathcal{X}.\end{aligned}$$

Here and below $\mathbb{E}_{t-1}$ denotes conditional expectation given $\mathscr{F}_{t-1}$.

For $t < \sigma_U$, with $\mathcal{L}_t$ and E_t from (33), the following inequality holds:

$$\mathcal{L}_{t+1} - \mathcal{L}_t \le -\tfrac{1}{8}E_t + D_{t,1} + D_{t,2} + D_{t,3} + \sum_{j=1}^{7} P_{t,j}, \tag{38}$$

where, with $\Delta_{t,i}$ from (26),

$$\begin{aligned}D_{t,1} &= \mathbb{E}_{t-1}\sum_i \lambda_{t,i}h_{t,i}\mathbf{g}_{t,i} - \sum_i \lambda_{t,i}h_{t,i}\mathbf{g}_{t,i},\\ D_{t,2} &= \sum_i h_{t,i}^2(\mathbb{E}_{t-1}\Delta_{t,i} - \Delta_{t,i}),\\ D_{t,3} &= \frac{\beta_1}{1-\beta_1}\sum_i[(\lambda_{t-1,i}-\lambda_{t,i}) - \mathbb{E}_{t-1}(\lambda_{t-1,i}-\lambda_{t,i})]\nabla_i f(\mathbf{z}_t)\mathbf{m}_{t-1,i},\end{aligned} \tag{39}$$

and, with the domination scale $\mathcal{X}$ from (34),

$$
\begin{aligned}
P_{t,1} &= \Bigg(\frac{\beta_1^2\mathcal{X}}{2(1-\beta_1)^2} \\
&\qquad + \frac{\beta_1^2\mathcal{X}^2\bar{\alpha}}{4(1-\beta_1)^2\sqrt{v_{\text{init}}}}\Bigg)\|\lambda_{t-1}\odot\mathbf{m}_{t-1}\|^2, \\
P_{t,2} &= \frac{3\mathcal{X}}{2}\|\lambda_t\odot\mathbf{g}_t\|^2, \\
P_{t,3} &= \frac{\beta_1^2}{(1-\beta_1)^2}\mathcal{X}\sum_i(\lambda_{t-1,i}-\lambda_{t,i})^2\mathbf{m}_{t-1,i}^2, \\
P_{t,4} &= \frac{140d\mathcal{X}^2\bar{\alpha}}{\sqrt{v_{\text{init}}}}\|\lambda_t\odot\mathbf{m}_t\|^2, \\
P_{t,5} &= \mathcal{X}\sum_i\mathbb{E}_{t-1}\Delta_{t,i}, \\
P_{t,6} &= \frac{8\beta_1(1+\beta_1)}{(1-\beta_1)^2}\frac{T}{T-1}\sum_i\Bigg[\frac{\lambda_{t-1,i}^2\mathbf{m}_{t-1,i}^2}{\bar{\alpha}} \\
&\qquad \times\Big(|h_{t,i}|+\mathcal{X}^{1/2}+\frac{\epsilon}{\sqrt{T}+\sqrt{T-1}}\Big)\Bigg], \\
P_{t,7} &= \frac{\beta_1}{T(1-\beta_1)}\sum_i\lambda_{t-1,i}|\nabla_i f(\mathbf{z}_t)\mathbf{m}_{t-1,i}|.
\end{aligned}
\tag{40}
$$

To prove (38), start from the exact identity, with $\mathbf{z}_t$ from (2),

$$
\mathbf{z}_{t+1}-\mathbf{z}_t = -\lambda_t\odot\mathbf{g}_t+\frac{\beta_1}{1-\beta_1}(\lambda_{t-1}-\lambda_t)\odot\mathbf{m}_{t-1}.
$$

Generalized smoothness at $\mathbf{z}_t$, and, for h_t from (33), $\|\nabla f(\mathbf{z}_t)-h_t\|\le\frac{\beta_1}{1-\beta_1}\mathcal{X}\|\lambda_{t-1}\odot\mathbf{m}_{t-1}\|$, give

$$
\begin{aligned}
\Psi(\mathbf{z}_{t+1})-\Psi(\mathbf{z}_t) \le &-\sum_i\lambda_{t,i}h_{t,i}\mathbf{g}_{t,i}+\frac{\beta_1}{1-\beta_1}\sum_i(\lambda_{t-1,i}-\lambda_{t,i})\nabla_i f(\mathbf{z}_t)\mathbf{m}_{t-1,i} \\
&+\frac{\beta_1^2\mathcal{X}}{2(1-\beta_1)^2}\|\lambda_{t-1}\odot\mathbf{m}_{t-1}\|^2 \\
&+\frac{3\mathcal{X}}{2}\|\lambda_t\odot\mathbf{g}_t\|^2 \\
&+\frac{\beta_1^2\mathcal{X}}{(1-\beta_1)^2}\|(\lambda_{t-1}-\lambda_t)\odot\mathbf{m}_{t-1}\|^2.
\end{aligned}
$$

Center the first complete adaptive product. By the definitions in (26), $\lambda_t=\widetilde{\lambda}_{t-1}-\Delta_t$. Conditional Cauchy–Schwarz and $\Delta_{t,i}\le 2\lambda_{t-1,i}$ give

$$
|h_{t,i}|\mathbb{E}_{t-1}[\Delta_{t,i}|\mathbf{g}_{t,i}|]\le\tfrac{1}{2}\lambda_{t-1,i}h_{t,i}^2+\mathbb{E}_{t-1}\Delta_{t,i}\,\mathbb{E}_{t-1}\mathbf{g}_{t,i}^2.
$$

The predictable first product is therefore at most $-E_t/2+\sum_i(h_{t,i}^2+\mathcal{X})\mathbb{E}_{t-1}\Delta_{t,i}$, with E_t from (33). The realized part of the latter sum has the exact decomposition

$$
\begin{aligned}
\sum_i h_{t,i}^2\Delta_{t,i} =& E_t-E_{t+1}+\sum_i(\widetilde{\lambda}_{t-1,i}-\lambda_{t-1,i})h_{t,i}^2 \\
&+\sum_i\lambda_{t,i}(h_{t+1,i}^2-h_{t,i}^2).
\end{aligned}
$$

For $T\ge 10$, the middle term is at most $E_t/16$. Put $q=h_{t+1}-h_t$, so $\|q\|\le\mathcal{X}\|\lambda_t\odot\mathbf{m}_t\|$ with $\mathcal{X}$ from (34). Expanding the last term and using $\lambda_t\le 2\lambda_{t-1}$ gives

$$
\sum_i\lambda_{t,i}(2h_{t,i}q_i+q_i^2)\le\tfrac{1}{8}E_t+\frac{34\mathcal{X}^2\bar{\alpha}}{\sqrt{Tv_{\text{init}}}}\|\lambda_t\odot\mathbf{m}_t\|^2.
$$

Consequently the realized sum is bounded by $E_t - E_{t+1} + E_t/4 + P_{t,4}$, with E_t and $P_{t,4}$ defined in (33) and (40).

For the lagged-momentum product, centering gives $D_{t,3}$ from (39). The deterministic difference $\lambda_{t-1} - \widetilde{\lambda}_{t-1}$ has magnitude at most λ_{t-1}/T, yielding $P_{t,7}$ in (40). For $\Delta_{t,i}$ from (26), direct rationalization gives

$$\mathbb{E}_{t-1}\Delta_{t,i} \le \frac{T}{T-1}\frac{|h_{t,i}| + \mathcal{X}^{1/2} + \epsilon/(\sqrt{T}+\sqrt{T-1})}{\bar{\alpha}}\lambda_{t-1,i}^2.$$

Since $\mathbf{z}_t$, $\mathbf{m}_{t-1}$ and λ_{t-1} are $\mathscr{F}_{t-1}$-measurable, the remaining part is

$$\frac{\beta_1}{1-\beta_1}\sum_i \mathbb{E}_{t-1}\Delta_{t,i}\,\nabla_i f(\mathbf{z}_t)\mathbf{m}_{t-1,i}.$$

Fix i. Apply the arithmetic–geometric mean inequality to $\frac{1}{4}\lambda_{t-1,i}^{1/2}|\nabla_i f(\mathbf{z}_t)|$ and $\frac{2\beta_1}{1-\beta_1}\lambda_{t-1,i}^{-1/2}\mathbb{E}_{t-1}\Delta_{t,i}|\mathbf{m}_{t-1,i}|$, whose product is half of the left side below. Then use $(\mathbb{E}_{t-1}\Delta_{t,i})^2 \le 2\lambda_{t-1,i}\mathbb{E}_{t-1}\Delta_{t,i}$, which follows from $0 \le \Delta_{t,i} \le 2\lambda_{t-1,i}$. This gives

$$\begin{aligned}\frac{\beta_1}{1-\beta_1}\mathbb{E}_{t-1}\Delta_{t,i}|\nabla_i f(\mathbf{z}_t)||\mathbf{m}_{t-1,i}| &\le \tfrac{1}{16}\lambda_{t-1,i}(\nabla_i f(\mathbf{z}_t))^2 + \frac{4\beta_1^2}{(1-\beta_1)^2}\frac{(\mathbb{E}_{t-1}\Delta_{t,i})^2}{\lambda_{t-1,i}}\mathbf{m}_{t-1,i}^2\\ &\le \tfrac{1}{16}\lambda_{t-1,i}(\nabla_i f(\mathbf{z}_t))^2 + \frac{8\beta_1^2}{(1-\beta_1)^2}\mathbb{E}_{t-1}\Delta_{t,i}\,\mathbf{m}_{t-1,i}^2.\end{aligned}$$

The rationalization bound above and $\beta_1^2 \le \beta_1(1+\beta_1)$ give

$$\begin{aligned}\frac{8\beta_1^2}{(1-\beta_1)^2}\mathbb{E}_{t-1}\Delta_{t,i}\,\mathbf{m}_{t-1,i}^2 \le \frac{8\beta_1(1+\beta_1)}{(1-\beta_1)^2}\frac{T}{T-1}\frac{\lambda_{t-1,i}^2\mathbf{m}_{t-1,i}^2}{\bar{\alpha}}\\ \times\Big(|h_{t,i}| + \mathcal{X}^{1/2} + \frac{\epsilon}{\sqrt{T}+\sqrt{T-1}}\Big).\end{aligned}$$

Summing over i, the remaining part is at most $\frac{1}{16}\sum_i \lambda_{t-1,i}(\nabla_i f(\mathbf{z}_t))^2 + P_{t,6}$, where $P_{t,6}$ is defined in (40). Finally, with E_t from (33),

$$\tfrac{1}{16}\sum_i \lambda_{t-1,i}(\nabla_i f(\mathbf{z}_t))^2 \le \tfrac{1}{8}E_t + \frac{\beta_1^2\mathcal{X}^2\bar{\alpha}}{4(1-\beta_1)^2\sqrt{v_{\text{init}}}}\|\lambda_{t-1}\odot\mathbf{m}_{t-1}\|^2.$$

Combining the three descent coefficients gives $-1/2 + 1/4 + 1/8 = -1/8$, proving the claim. Multiplication by $\mathbf{1}_{\{t<\sigma_U\}}$ preserves the martingale-difference property because this event is $\mathscr{F}_{t-1}$-measurable.

D.3 Uniform bounds for the stopped gradients

For the stopping index and gradients defined in (32)–(33), before σ_U, with $\mathcal{X}$ from (34) and $U = U_\delta$ from (15),

$$\|h_t\|,\ \|\nabla f(\mathbf{z}_t)\| \le 6\mathcal{X}^3\sqrt{U}. \tag{41}$$

For $p > 0$, the last term of (14) and (34) imply

$$\sqrt{\frac{2G_{\mathrm{q}}(U)^2}{U+1}} \le \mathcal{X}.$$

Since $U + 1 \le 2U$, the bound $\|h_t\|, \|\nabla f(\mathbf{z}_t)\| \le G_{\mathrm{q}}(U)$ before σ_U gives

$$\begin{aligned}\|h_t\|,\ \|\nabla f(\mathbf{z}_t)\| &\le \mathcal{X}\sqrt{U}\\ &\le 6\mathcal{X}^3\sqrt{U},\end{aligned}$$

which is (41) for $p > 0$. The definitions (10)–(11) give $K, C_{\mathrm{q}} \ge 1$; with G_{q} from (13), this gives $G_{\mathrm{q}}(U) \ge U$, hence

$$\mathcal{X} \ge \sqrt{\frac{2G_{\mathrm{q}}(U)^2}{U+1}} \ge \sqrt{U} \qquad (p > 0). \tag{42}$$

For $p = 0$, the local Lipschitz condition extends by partitioning any segment to global $(L_0 + L_p)$-smoothness. Before σ_U, $\Psi(\mathbf{z}_t) \le U$ by (32), and (34) gives $L_0 + L_p \le \mathcal{P} \le \mathcal{X}$, with $\mathcal{P}$ from (12). Hence

$$\|\nabla f(\mathbf{z}_t)\| \le \sqrt{2(L_0 + L_p)U} \le \sqrt{2\mathcal{X}U}.$$

Since $(L_0 + L_p)\beta_1/(1 - \beta_1) \le \mathcal{P}^2 \le \mathcal{X}^2$, (35) gives

$$\|h_t - \nabla f(\mathbf{z}_t)\| \le \frac{2(L_0 + L_p)\beta_1}{1 - \beta_1}\sqrt{d}\,\bar{\alpha} \le 2\sqrt{d}\,\bar{\alpha}\mathcal{X}^2 \le 1.$$

With $U \ge 1$ (see (15)) and $\mathcal{X} \ge 1$,

$$\begin{aligned}\|h_t\| &\le \sqrt{2\mathcal{X}U} + 1\\ &\le 6\mathcal{X}^3\sqrt{U}.\end{aligned}$$

The same bound holds for $\|\nabla f(\mathbf{z}_t)\|$. This proves (41) in this case. The all-path displacement bounds

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_1\| \le 2\sqrt{d}\,\bar{\alpha}T, \qquad \|\mathbf{z}_t - \mathbf{z}_1\| \le \left(1 + \frac{6\beta_1}{1 - \beta_1}\right)\sqrt{d}\,\bar{\alpha}T$$

also give, before σ_U,

$$\|h_t\|^2,\ \|\nabla f(\mathbf{z}_t)\|^2 \le 36\mathcal{X}^6 \min\{U, 1 + d\bar{\alpha}^2T^2\} \qquad (p = 0). \tag{43}$$

The elementary inequality

$$\frac{\min\{U, d\bar{\alpha}^2T^2\}}{\sqrt{T}} \le d^{1/4}\sqrt{\bar{\alpha}}\,U^{3/4} \tag{44}$$

follows by considering $T \le \sqrt{U}/(\sqrt{d}\,\bar{\alpha})$ and its complement.

D.4 Four stopped concentration estimates

Throughout this subsection $U = U_\delta$ is the height (15), and every closing step absorbs the factor $1 + \log(16/\delta)$ into U only through (37). Use $\sigma = \sigma_U$, the stopping index defined in (32), in (30) and (31).

First, (41) and the noise envelope (13), bounded through (34), show that the stopped raw second-moment sum has expectation at most $T[36\mathcal{X}^6U + \mathcal{X}]$. Markov's inequality and Jensen's inequality across coordinates give, for Z_T from (30), except on an event of probability $e^{-(1+\log(16/\delta))}$,

$$\begin{aligned}Z_T &\le d\{(1 + \log(16/\delta)) + \log[1 + 37\mathcal{X}^7U]\}\\ &\le 2^7d\mathcal{X}^2(1 + \log(16/\delta)).\end{aligned} \tag{45}$$

For the last inequality, bound $1 + 37\mathcal{X}^7U$ by $38\mathcal{X}^7U$ and split its logarithm into three parts. Since $\mathcal{X} \ge 2$ by (34), $\log(38\mathcal{X}) \le 4\mathcal{X}$ and $6\log(\mathcal{X}) \le 6\mathcal{X}$, while $\log U \le 2U^{1/4}$ and (36) give $\log U \le 2^6\mathcal{X}^2(1 + \log(16/\delta))$. Using $\mathcal{X} \ge 2$ and $1 + \log(16/\delta) \ge 1$ once more, $1 \le \mathcal{X}^2/4$ and $10\mathcal{X} \le 5\mathcal{X}^2$, so the braces are at most $(1/4 + 5 + 2^6)\mathcal{X}^2(1 + \log(16/\delta)) \le 2^7\mathcal{X}^2(1 + \log(16/\delta))$.

Second, Lemma C.5, at exponent $1 + \log(16/\delta)$, and (28) for $\Delta_{t,i}$ from (26) imply, with σ_U from (32), except on an event of probability $e^{-(1+\log(16/\delta))}$,

$$\sum_{t<T} \mathbf{1}_{\{t<\sigma_U\}} \sum_i \mathbb{E}_{t-1}\Delta_{t,i} \le \frac{2e(1 + \log(16/\delta))d\bar{\alpha}}{\sqrt{T v_{\text{init}}}}. \tag{46}$$

Indeed the $L^{1+\log(16/\delta)}$ norm of the left side is at most $(1 + \log(16/\delta))$ times the deterministic upper bound for the corresponding realized stopped sum; Markov's inequality at e times that norm gives the stated probability.

Third, the first two martingales in (39) combine exactly as

$$D_{t,12} := D_{t,1} + D_{t,2} = \mathbb{E}_{t-1}\sum_i \lambda_{t,i}h_{t,i}\xi_{t,i} - \sum_i \lambda_{t,i}h_{t,i}\xi_{t,i}.$$

Here $\lambda_{t,i}$ is from Algorithm 1 and h_t, ξ_t are defined in (33). Using the gradient estimate (41), the noise envelope (13), and the coefficient bound (34), the stopped predictable quadratic variation is at most

$$144\bar{\alpha}^2\mathcal{X}^8 U.$$

To see this, bound each conditional variance by the uncentered square, then use Cauchy–Schwarz, $\lambda_{t,i} \le 2\bar{\alpha}/\sqrt{T v_{\mathrm{init}}}$, and $\mathbb{E}_{t-1}\|\xi_t\|^2 \le \mathcal{X}$. The stopped increment magnitude is at most

$$12\sqrt{d}\,\bar{\alpha}\mathcal{X}^4\sqrt{U} + 144\bar{\alpha}\mathcal{X}^7\frac{U}{\sqrt{T}}.$$

For $p = 0$, (43) permits replacing U in the second summand by $\min\{U, 1 + d\bar{\alpha}^2T^2\}$. The one-sided Freedman inequality (25) applies at failure level $e^{-(1+\log(16/\delta))}$. Its square-root contribution is at most $17\bar{\alpha}\mathcal{X}^4(1+\log(16/\delta))^{1/2}U^{1/2}$, since $\sqrt{2\cdot 144} \le 17$; the first jump contribution is at most $8\sqrt{d}\,\bar{\alpha}\mathcal{X}^4(1+\log(16/\delta))U^{1/2}$. For $p > 0$, (42) gives $U \le \mathcal{X}\sqrt{U}$ and therefore bounds the remaining jump contribution by $96\bar{\alpha}\mathcal{X}^8(1+\log(16/\delta))U^{1/2}$. For $p = 0$, use the replacement above, $\min\{U, 1+d\bar{\alpha}^2T^2\} \le 1 + \min\{U, d\bar{\alpha}^2T^2\}$, (44), $d^{1/4}\sqrt{\bar{\alpha}} \le \sqrt{d\bar{\alpha}} \le 1$ and $U \ge 1$; they bound it by $192\bar{\alpha}\mathcal{X}^7(1+\log(16/\delta))U^{3/4}$. Since $d, U, 1+\log(16/\delta) \ge 1$ and $\mathcal{X} \ge 2$ by (34), in both cases the three contributions sum to at most

$$217\sqrt{d}\,\bar{\alpha}\mathcal{X}^8(1+\log(16/\delta))U^{3/4} < U/32,$$

the last step by $\sqrt{d} \le d$ and (37), for the height $U = U_\delta$ of (15). Therefore, except on an event of probability $e^{-(1+\log(16/\delta))}$,

$$\sum_{t<T}\mathbf{1}_{\{t<\sigma_U\}}D_{t,12} \le U/32.$$

Fourth, for Z_n from (30), introduce the predictable crossing

$$\tau = \inf\{1 \le n \le T : Z_n > 2^7 d\mathcal{X}^2(1+\log(16/\delta))\}, \qquad \inf\varnothing = \infty. \tag{47}$$

Stop $D_{t,3}$ from (39) at $t < \sigma_U$ and $t < \tau$. Conditional on the past, $\nabla f(\mathbf{z}_t)$ and $\mathbf{m}_{t-1}$ are fixed ($\mathbf{z}_t$ is the auxiliary sequence (2) and $\mathbf{m}_{t-1}$ the momentum of Algorithm 1), while $|\lambda_{t-1,i} - \lambda_{t,i}| \le 3\lambda_{t-1,i}$. Its quadratic variation is bounded, using (41) and (31), by

$$\frac{9\beta_1^2}{(1-\beta_1)^2}[36\mathcal{X}^6U]\sum_{t<T}\mathbf{1}_{\{t<\sigma_U, t<\tau\}}\|\lambda_{t-1}\odot\mathbf{m}_{t-1}\|^2 \le 2^{22}d\bar{\alpha}^2\mathcal{X}^{10}(1+\log(16/\delta))U.$$

The summand at time t uses $\mathbf{m}_{t-1}$. Let t be an included time, that is, $t < T$, $t < \sigma_U$ and $t < \tau$. Since $\mathbf{m}_0 = \mathbf{0}$, $\sum_{s\le t}\|\lambda_{s-1}\odot\mathbf{m}_{s-1}\|^2 = \sum_{s<t}\|\lambda_s\odot\mathbf{m}_s\|^2$. The momentum bound in (31), with $\sigma = \sigma_U$ and $n = t$, for which $n \wedge \sigma_U = t$, bounds this sum by $64\bar{\alpha}^2 Z_t$. Because $t < \tau$, (47) gives $Z_t \le 2^7 d\mathcal{X}^2(1+\log(16/\delta))$. Taking t to be the largest included time bounds the stopped sum above. The increment magnitude is at most $72\sqrt{d}\,\bar{\alpha}\mathcal{X}^5\sqrt{U}$. Freedman's inequality (25) therefore gives an upper deviation at most

$$2^{12}\sqrt{d}\,\bar{\alpha}\mathcal{X}^5(1+\log(16/\delta))U^{1/2} < U/32$$

with failure probability $e^{-(1+\log(16/\delta))}$; here $\sqrt{2\cdot 2^{22}} + \frac{2}{3}\cdot 72 \le 2^{12}$, and the last step uses $\sqrt{d} \le d$ and (37). On (45), $\tau > T$ by (47), so this is the desired estimate for the stopped $D_{t,3}$ sum.

A union bound gives all four events, namely (45), (46) and the two Freedman bounds for the stopped $D_{t,12}$ and $D_{t,3}$ sums, simultaneously with probability at least $1 - 4e^{-(1+\log(16/\delta))} > 1 - \delta/2$.

D.5 Closing the stopped descent

Lemma D.2 (Stopped energy and localization)**.** *Assume Assumptions 3.1–3.3, $T \ge 10$, $\beta_2 = 1 - 1/T$, and $\alpha = \bar{\alpha}/\sqrt{T}$. For $0 < \delta < 1$, take $U = U_\delta$ from* (15) *and $\bar{\alpha}$ in the interval* (16)*. With the stopping index and energies defined in* (32)–(33)*, with probability at least $1 - \delta/2$,*

$$\mathcal{L}_{\sigma_U\wedge T} + \frac{1}{8}\sum_{t<\sigma_U\wedge T}E_t \le U/2, \qquad \sigma_U > T, \qquad \sum_{t=1}^{T-1}E_t \le 4U.$$

Proof. On (45) and (46), (31), (41) and the residual definitions (40) give the following cumulative bounds, with $\mathcal{X}$ from (34) and all sums being over $t < \sigma_U \wedge T$ for σ_U from (32):

$$\sum_t P_{t,1} \le 2^{13} d\bar{\alpha}^2 \mathcal{X}^7 (1 + \log(16/\delta)),$$
$$\sum_t P_{t,2} \le 2^{10} d\bar{\alpha}^2 \mathcal{X}^3 (1 + \log(16/\delta)),$$
$$\sum_t P_{t,3} \le 2^{17} d\bar{\alpha}^2 \mathcal{X}^5 (1 + \log(16/\delta)),$$
$$\sum_t P_{t,4} \le 2^{21} d^2 \bar{\alpha}^3 \mathcal{X}^6 (1 + \log(16/\delta)),$$
$$\sum_t P_{t,5} \le 6 d\bar{\alpha} \mathcal{X}^3 (1 + \log(16/\delta)),$$
$$\sum_t P_{t,6} \le 2^{22} d\bar{\alpha} \mathcal{X}^7 (1 + \log(16/\delta)) U^{1/2},$$
$$\sum_t P_{t,7} \le 12\sqrt{d}\, \bar{\alpha} \mathcal{X}^5 U^{1/2}.$$

For example, the coefficient in $P_{t,6}$ from (40) is at most $512\mathcal{X}^5\sqrt{U}/\bar{\alpha}$, and its lagged-momentum sum is at most $64\bar{\alpha}^2 Z_T$, with Z_T defined in (30). This gives exactly its displayed bound. For $P_{t,5}$, use (46); for $P_{t,7}$, use (27) directly. The other four use (31) together with $64\bar{\alpha}^2 Z_T \le 2^{13} d\bar{\alpha}^2 \mathcal{X}^2 (1 + \log(16/\delta))$ from (45). By (34), $\mathcal{P} \le \mathcal{R}(U) \le \mathcal{X}$ (with $\mathcal{P}$ and $\mathcal{R}$ from (12) and (14)) and $\mathcal{X} \ge 2$, so the power $\mathcal{X}^k$ is nondecreasing in $k \ge 0$; this is used below to enlarge exponents. For $P_{t,1}$, $\beta_1^2/(1-\beta_1)^2 \le \mathcal{P} \le \mathcal{X}$, $v_{\text{init}}^{-1/2} \le 1 + v_{\text{init}}^{-1} \le \mathcal{P}$ and $\mathcal{X}^4 \bar{\alpha} \le 2^{-100}$ from (35) bound its coefficient by

$$\frac{\mathcal{X}^2}{2} + \frac{\mathcal{X}^4 \bar{\alpha}}{4} \le \mathcal{X}^2,$$

so $\sum_t P_{t,1} \le 2^{13} d\bar{\alpha}^2 \mathcal{X}^4 (1 + \log(16/\delta))$; enlarging the exponent gives the displayed bound. For $P_{t,2}$, with Z_T from (30),

$$\sum_t P_{t,2} \le \frac{3\mathcal{X}}{2} \cdot 4\bar{\alpha}^2 Z_T \le 768 d\bar{\alpha}^2 \mathcal{X}^3 (1 + \log(16/\delta)),$$

and $768 < 2^{10}$ gives the displayed bound. For $P_{t,3}$, with λ_t from Algorithm 1, $|\lambda_{t-1} - \lambda_t| \le 3\lambda_{t-1}$ and $\beta_1^2/(1-\beta_1)^2 \le \mathcal{X}$ give

$$\sum_t P_{t,3} \le 9\mathcal{X}^2 \cdot 64\bar{\alpha}^2 Z_T \le 9 \cdot 2^{13} d\bar{\alpha}^2 \mathcal{X}^4 (1 + \log(16/\delta)),$$

and $9 \cdot 2^{13} < 2^{17}$ with an enlarged exponent gives the displayed bound. For $P_{t,4}$, $v_{\text{init}}^{-1/2} \le \mathcal{P} \le \mathcal{X}$ gives

$$\sum_t P_{t,4} \le 140 d\bar{\alpha} \mathcal{X}^3 \cdot 64\bar{\alpha}^2 Z_T \le 140 \cdot 2^{13} d^2 \bar{\alpha}^3 \mathcal{X}^5 (1 + \log(16/\delta)),$$

and $140 \cdot 2^{13} < 2^{21}$ with an enlarged exponent gives the displayed bound. By (35), $d\bar{\alpha} \le 1$. Together with $d, U, 1 + \log(16/\delta) \ge 1$ and the monotonicity in k noted above, each of the seven bounds is at most its constant times $d\bar{\alpha}\mathcal{X}^7 (1 + \log(16/\delta)) U^{3/4}$. The seven constants sum to less than 2^{23}, so the seven bounds sum to at most

$$2^{23} d\bar{\alpha} \mathcal{X}^7 (1 + \log(16/\delta)) U^{3/4} < U/32,$$

the last step by (37). The energy definition (33), the parameter scale (12), and the height (15) give

$$\mathcal{L}_1 \le 2\mathcal{P} < U/32.$$

Summing (38) to $(\sigma_U \wedge T) - 1$ on all four good events of Appendix D.4 yields

$$\mathcal{L}_{\sigma_U \wedge T} + \frac{1}{8} \sum_{t < \sigma_U \wedge T} E_t \le U/2.$$

The energy definition (33) gives $\mathcal{L}_t \ge \Psi(\mathbf{z}_t)$, with Ψ from (1), so a crossing $\sigma_U \le T$, for the stopping index σ_U of (32), contradicts this inequality. Thus

$$\sigma_U > T, \qquad \sum_{t=1}^{T-1} E_t \le 4U \tag{48}$$

on an event of probability at least $1 - \delta/2$. □

D.6 De-preconditioning

For the stopping index (32) and variance envelope (13), the same predictable stopping gives

$$\mathbb{E} \sum_{t<T} \mathbf{1}_{\{t < \sigma_U\}} \|\xi_t\|^2 \le T[C + \mathcal{V}_{\mathrm{q}}(U)].$$

Here C is the constant of Assumption 3.3 and ξ_t is defined in (33). With an additional failure probability at most $\delta/2$, this stopped sum divided by T is at most

$$\frac{2}{\delta}[C + \mathcal{V}_{\mathrm{q}}(U)].$$

On its intersection with (48), the accumulator recursion of Algorithm 1, the variables in (33), and $\|\mathbf{g}_t\|^2 \le 2\|h_t\|^2 + 2\|\xi_t\|^2$ imply

$$\max_{t<T,i} \mathbf{v}_{t,i} \le v_{\mathrm{init}} + \frac{2}{T} \sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2 + \frac{4}{\delta}[C + \mathcal{V}_{\mathrm{q}}(U)].$$

With $U = U_\delta$ from (15) and $\mathcal{V}_{\mathrm{q}}$ from (13), (48) consequently gives

$$\frac{1}{T} \sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2 \le \frac{4U(\epsilon + \sqrt{v_{\mathrm{init}}})}{\bar{\alpha}\sqrt{T}} + \frac{4U}{\bar{\alpha}\sqrt{T}} \sqrt{\frac{2}{T} \sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2} + \frac{8U\sqrt{C + \mathcal{V}_{\mathrm{q}}(U)}}{\bar{\alpha}\sqrt{\delta T}}.$$

The elementary inequality

$$\frac{4U}{\bar{\alpha}\sqrt{T}} \sqrt{\frac{2}{T} \sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2} \le \frac{1}{2T} \sum_{s=1}^{T-1} \|\nabla f(\boldsymbol{\theta}_s)\|^2 + \frac{16U^2}{\bar{\alpha}^2 T}$$

gives (17). The union bound gives total failure at most δ, proving Theorem B.1.

Confidence dependence of the admissible stepsize. For the decomposition (4), the coefficient in (3) is

$$\eta_{\mathrm{gen}} = \mathbf{1}_{\{p>0\}} + \sqrt{A + [B-1]_+ + B\mathbf{1}_{\{\rho_2 \ne 2\}}}. \tag{49}$$

Here A, B, ρ_2 are the constants of Assumption 3.3. It vanishes exactly when $p = 0$, $A = 0$, $B \le 1$, and $B = 0$ or $\rho_2 = 2$. In that case $Bz^{\rho_2} - z^2 \le 0$ for all $z \ge 0$, so (13)–(14) give $\mathcal{V}_{\mathrm{q}}(U) = 0$ and, since $p = 0$, $\mathcal{R}(U) = \mathcal{P}$ for every $U \ge 1$. If $\eta_{\mathrm{gen}} > 0$, then (15) gives $U_\delta = 2^{20}\mathcal{P}^8$ for every δ. Choose the upper endpoint of (16). Its reciprocal is therefore

$$\bar{\alpha}^{-1} = \begin{cases} 2^{100} d\, \mathcal{P}^{48}, & \eta_{\mathrm{gen}} = 0, \\ 2^{100} d\, \mathcal{R}(2^{20}\mathcal{P}^8)^{48}(1 + \log(16/\delta)), & \eta_{\mathrm{gen}} > 0, \end{cases} \tag{50}$$

where $\mathcal{P}$ and $\mathcal{R}$ are defined in (12) and (14). Neither $\mathcal{P}$ nor $\mathcal{R}(2^{20}\mathcal{P}^8)$ depends on T, δ, d. Since $1 + \log(16/\delta) = (1 + \log 16) + \log(1/\delta)$, the second case of (50) is the sum of a term independent of δ and $2^{100} d\, \mathcal{R}(2^{20}\mathcal{P}^8)^{48} \log(1/\delta)$. This proves (4); its logarithmic term is present only when $\eta_{\mathrm{gen}} > 0$.

D.7 Proof of Claim B.1

Proof. When $p = 0, A = 0, B = 1, \rho_2 = 2$, (49) gives $\eta_{\mathrm{gen}} = 0$, so (15) reads $U_\delta = 2^{20}\mathcal{P}^8(1 + \log(16/\delta))^4$ and (16) carries no logarithmic factor. The definitions (13)–(14) give $\mathcal{V}_{\mathrm{q}}(U) = 0$ and $\mathcal{R}(U) = \mathcal{P}$ for every $U \geq 1$, with $\mathcal{P}$ given by (12). The same fixed prefactor is therefore admissible for all confidence levels. The bound (17) is a decreasing function of δ. For $1 \leq r < 2$, quantile integration bounds

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{s=1}^{T-1}\|\nabla f(\boldsymbol{\theta}_s)\|^2\right)^r\right]$$

by the integral of the r-th power of the right-hand side of (17) over $0 < \delta < 1$. The only singular endpoint integral is of the form

$$\int_0^1 \delta^{-r/2}(1 + \log(16/\delta))^{4r}\, d\delta < \infty.$$

All remaining integrals are finite moments of $\log(1/\delta)$, and thus

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{s=1}^{T-1}\|\nabla f(\boldsymbol{\theta}_s)\|^2\right)^r\right] = \mathcal{O}(T^{-r/2}).$$

For example,

$$\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T-1}\|\nabla f(\boldsymbol{\theta}_s)\|^2\right] \leq \frac{8}{\bar{\alpha}\sqrt{T}}\int_0^1 U_\delta(\epsilon + \sqrt{v_{\mathrm{init}}} + 2\sqrt{C}\,\delta^{-1/2})\, d\delta + \frac{32}{\bar{\alpha}^2 T}\int_0^1 U_\delta^2\, d\delta.$$

All parameter dependence in these finite integrals is specified by (15) and (12). For each integer $j \geq 0$ and real $s > 0$, substitution $u = \log(1/\delta)$ followed by integration by parts gives

$$\int_0^1 \delta^{s-1}\log^j(1/\delta)\, d\delta = \int_0^\infty e^{-su}u^j\, du = \frac{j!}{s^{j+1}}. \tag{51}$$

Expand the fourth and eighth powers in (15) by the binomial theorem. Applying (51) with $s = 1$ to the deterministic terms and with $s = 1/2$ to the noise term yields (18), proving Claim B.1. □

D.8 Proof of Theorem B.2

Use $\delta_T, \bar{\alpha}_T, A_{\mathrm{tail}}$ from (19) and the confidence height (15). The value $\mathcal{R}(U_\delta)$, with $\mathcal{R}$ from (14), does not depend on δ: if $\eta_{\mathrm{gen}} > 0$, then $U_\delta = 2^{20}\mathcal{P}^8$ by (15), and if $\eta_{\mathrm{gen}} = 0$, then $\mathcal{R}(U) = \mathcal{P}$ for every $U \geq 1$ (with $\mathcal{P}$ from (12)), as shown after (49). Since $(1 + \log(16/\delta))^{-\mathbf{1}\{\eta_{\mathrm{gen}} > 0\}}$ is nondecreasing in δ, the fixed prefactor $\bar{\alpha}_T$ satisfies (16) for every $\delta \in [\delta_T, 1)$. Theorem B.1 consequently applies to the same run with prefactor $\bar{\alpha}_T$ at every $\delta \in [\delta_T, 1)$. These individual tail bounds can therefore be integrated over the confidence level.

The bound in (17), with the denominator fixed at $\bar{\alpha}_T$ from (19), is decreasing in δ. By (15) and (13), U_δ and $C + \mathcal{V}_{\mathrm{q}}(U_\delta)$ grow at most polynomially in $1 + \log(16/\delta)$. Thus integrating the bound over $[\delta_T, 1]$ gives

$$\mathcal{O}\left(\frac{1}{\bar{\alpha}_T\sqrt{T}} + \frac{1}{\bar{\alpha}_T^2 T}\right).$$

For the remaining quantiles $\delta < \delta_T$, with δ_T from (19), Lemma C.3 at $\boldsymbol{\theta}_1 = \mathbf{z}_1$ and $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_1\| \leq 2\sqrt{d}\,\bar{\alpha}_T T \leq 2T$ give the deterministic bound

$$\frac{1}{T}\sum_{s=1}^{T-1}\|\nabla f(\boldsymbol{\theta}_s)\|^2 \leq \left[(1 + \|\nabla f(\boldsymbol{\theta}_1)\|)^{1-p} + (1-p)(L_0 + L_p)\right]^{2/(1-p)} \times (1 + 2T)^{2/(1-p)}.$$

Their integral is at most this bound times δ_T. Since $1 + 2T \le 3T$, $\delta_T(1 + 2T)^{2/(1-p)} \le 3^{2/(1-p)}T^{2/(1-p)-A_{\text{tail}}}$, which is $o(T^{-1/2})$ by the choice $A_{\text{tail}} > 1/2 + 2/(1-p)$ in (19). Adding this contribution to the integral of (17) over $[\delta_T, 1]$, with U_δ, $\mathcal{V}_{\text{q}}$ and $\bar{\alpha}_T$ defined in (15), (13) and (19), respectively, gives the complete inequality (20). Finally, (19) and (50) at $\delta = \delta_T$ give $\bar{\alpha}_T^{-1} = 2^{100}d\,\mathcal{P}^{48}$ when $\eta_{\text{gen}} = 0$ and $\bar{\alpha}_T^{-1} = 2^{100}d\,\mathcal{R}(2^{20}\mathcal{P}^8)^{48}(1 + \log 16 + A_{\text{tail}} \log T)$ when $\eta_{\text{gen}} > 0$. The integrals in (20) are at most their values over $(0, 1]$, which are finite and independent of T and d. Hence the three integral terms of (20) are $\mathcal{O}(d/\sqrt{T} + d^2/T)$ when $\eta_{\text{gen}} = 0$ and $\mathcal{O}(d(1 + \log T)/\sqrt{T} + d^2(1 + \log T)^2/T)$ when $\eta_{\text{gen}} > 0$, while the rare-event term is $o(T^{-1/2})$. This proves Theorem B.2.

E Proof of Proposition B.1

Proof. We give the construction for the one-dimensional case. Let

$$\chi_\star := (\delta T)^{-1/4}, \qquad H := \sqrt{\frac{T}{2\delta}}, \qquad L := 16\bar{\alpha}\sqrt{8}\,(\delta T)^{1/4}. \tag{52}$$

Using the height $\chi_\star$ and width L from (52), define the continuous piecewise-affine function

$$h(u) = \begin{cases} 0, & u \le -2L, \\ \chi_\star(u + 2L)/L, & -2L \le u \le -L, \\ \chi_\star, & -L \le u \le -\bar{\alpha}/8, \\ -8\chi_\star u/\bar{\alpha}, & -\bar{\alpha}/8 \le u \le 0, \\ 0, & u \ge 0, \end{cases} \qquad f(u) = \int_0^u h(r)\,dr. \tag{53}$$

The pieces in (53) agree at their endpoints. Since $\delta T \ge 1$, the scales in (52) satisfy $\chi_\star \le 1$ and $L \ge 16\bar{\alpha}\sqrt{8}$. Thus $f \in C^1(\mathbb{R})$, its derivative is globally Lipschitz with constant at most $8\chi_\star/\bar{\alpha} \le 8/\bar{\alpha}$, and it satisfies the smoothness assumption (Assumption 3.2) with $p = 0$ and $L_0 = L_p = 4/\bar{\alpha}$. Moreover, h in (53) is nonnegative and vanishes outside $[-2L, 0]$. Using $L\chi_\star$ from (52) gives

$$f^* = -\int_{-2L}^{0} h(u)\,du > -\infty, \qquad f(0) - f^* \le 2L\chi_\star = 32\sqrt{8}\,\bar{\alpha}.$$

Set $\boldsymbol{\theta}_1 = 0$, $\mathbf{m}_0 = 0$, and $\mathbf{v}_0 = v_{\text{init}} > 0$ in Algorithm 1. Let the oracle be

$$\mathbf{g}_t = h(\boldsymbol{\theta}_t) + \omega_t,$$

where the independent noise variables, with spike height H from (52), satisfy

$$\omega_t = \begin{cases} H, & \text{with probability } \delta/T, \\ -H, & \text{with probability } \delta/T, \\ 0, & \text{with probability } 1 - 2\delta/T. \end{cases} \tag{54}$$

The noise law (54) and the spike height in (52) give $\mathbb{E}[\omega_t] = 0$ and $\mathbb{E}[\omega_t^2] = 1$, so

$$\mathbb{E}[\mathbf{g}_t \mid \mathscr{F}_{t-1}] = h(\boldsymbol{\theta}_t) = \nabla f(\boldsymbol{\theta}_t), \qquad \mathbb{E}[\mathbf{g}_t^2 \mid \mathscr{F}_{t-1}] = h(\boldsymbol{\theta}_t)^2 + 1.$$

Thus Assumption 3.3 holds with $A = 0$, $B = 1$, $\rho_1 = 0$, $\rho_2 = 2$, and $C = 1$, independently of T, δ.

For the noise in (54), let

$$\mathcal{S} := \bigcup_{k=1}^{\lfloor T/2 \rfloor} \{\omega_k = H, \quad \omega_t = 0 \text{ for } 1 \le t < T,\ t \ne k\}. \tag{55}$$

Thus exactly one positive spike occurs in the first $\lfloor T/2 \rfloor$ iterations, and all other noise values before T vanish. Since $\delta \le 1/8$,

$$\mathbb{P}(\mathcal{S}) = \left\lfloor \frac{T}{2} \right\rfloor \frac{\delta}{T} \left(1 - \frac{2\delta}{T}\right)^{T-2} \ge \frac{\delta}{4}.$$

Indeed, $\lfloor T/2 \rfloor / T \ge 1/3$, and Bernoulli's inequality bounds the survival factor below by $1 - 2\delta \ge 3/4$. Figure 3 illustrates the entry into the constant-gradient region and the ensuing slow motion.

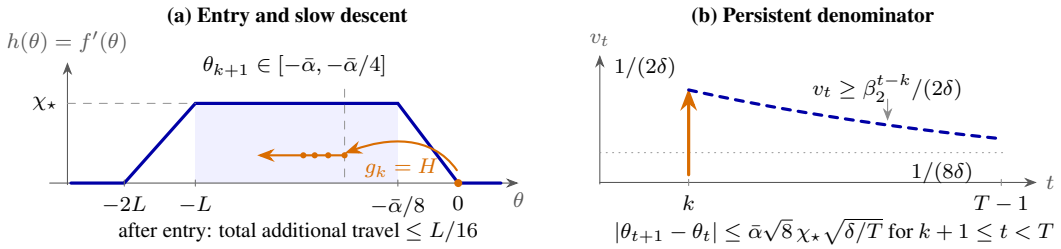


Figure 3: Confidence lower-bound dynamics on the single-spike event $\mathcal{S}$ from (55), with $\beta_1 = 0$. The scales $\chi_\star, H, L$ are defined in (52), and (a) shows the derivative in (53). At the spike time k, the positive oracle output sends the iterate left into the shaded plateau. Subsequent steps remain leftward but cannot leave it before the horizon. Panel (b) shows a lower envelope of the accumulator, not an exact accumulator trajectory: the spike contribution decays only by β_2^{t-k} and stays above $1/(8\delta)$. Both panels are schematic and not to scale.

On the event $\mathcal{S}$ in (55), let $k \leq T/2$ denote the spike time. Before time k, the iterate remains at 0, because $h(0) = 0$ in (53) and the noise is zero. At time k, $\mathbf{g}_k = H$, with H from (52), and

$$\mathbf{v}_k \leq v_{\text{init}} + \frac{H^2}{T} = v_{\text{init}} + \frac{1}{2\delta}, \qquad \mathbf{v}_k \geq \frac{H^2}{T} = \frac{1}{2\delta}.$$

The confidence range on δ in Proposition B.1 ensures

$$\sqrt{1 + 2\delta v_{\text{init}}} + \epsilon\sqrt{2\delta} \leq \frac{\sqrt{5}+1}{2} < 4.$$

Substituting H from (52), the resulting Adam step satisfies

$$\boldsymbol{\theta}_{k+1} = -\frac{\bar{\alpha}}{\sqrt{T}}\frac{H}{\sqrt{\mathbf{v}_k} + \epsilon} = -\frac{\bar{\alpha}}{\sqrt{2\delta\mathbf{v}_k} + \epsilon\sqrt{2\delta}}.$$

Because $\mathbf{v}_k \geq 1/(2\delta)$, we have $|\boldsymbol{\theta}_{k+1}| \leq \bar{\alpha}$. The upper accumulator bound gives

$$\sqrt{2\delta\mathbf{v}_k} + \epsilon\sqrt{2\delta} \leq \sqrt{1 + 2\delta v_{\text{init}}} + \epsilon\sqrt{2\delta} \leq 4,$$

and hence $|\boldsymbol{\theta}_{k+1}| \geq \bar{\alpha}/4$. Therefore

$$\boldsymbol{\theta}_{k+1} \in [-\bar{\alpha}, -\bar{\alpha}/4] \subset [-L, -\bar{\alpha}/8],$$

so the iterate enters the constant-gradient part of (53).

We claim that $\boldsymbol{\theta}_t \in [-L, -\bar{\alpha}/8]$ for all $t = k+1, \ldots, T-1$, where L is defined in (52). We prove this by induction. On $\mathcal{S}$ in (55), there are no further noise spikes. As long as the iterates remain in the constant part of (53), $\mathbf{g}_t = h(\boldsymbol{\theta}_t) = \chi_\star$, the updates move monotonically to the left, and the accumulator remains large:

$$\mathbf{v}_t \geq \beta_2^{t-k}\mathbf{v}_k \geq \frac{1}{8\delta}, \qquad k \leq t \leq T-1.$$

Here we used $(1 - 1/T)^T \geq 1/4$ for $T \geq 2$. Therefore each later displacement in the ramp (53), whose height $\chi_\star$ is given in (52), is bounded by

$$|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t| = \frac{\bar{\alpha}}{\sqrt{T}}\frac{\chi_\star}{\sqrt{\mathbf{v}_t} + \epsilon} \leq \bar{\alpha}\sqrt{8}\,\frac{\chi_\star\sqrt{\delta}}{\sqrt{T}}.$$

Summing over at most T steps and using the definitions of $\chi_\star, L$ in (52) gives

$$\sum_{t=k+1}^{T-1} |\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t| \leq \bar{\alpha}\sqrt{8}\,\chi_\star\sqrt{\delta T} = \bar{\alpha}\sqrt{8}\,(\delta T)^{1/4} \leq \frac{L}{16}.$$

With L from (52), for every later time covered by the induction,

$$\boldsymbol{\theta}_t \geq \boldsymbol{\theta}_{k+1} - \frac{L}{16} \geq -\bar{\alpha} - \frac{L}{16} \geq -L, \qquad \boldsymbol{\theta}_t \leq \boldsymbol{\theta}_{k+1} \leq -\frac{\bar{\alpha}}{4} \leq -\frac{\bar{\alpha}}{8}.$$

This closes the induction and proves $\boldsymbol{\theta}_t \in [-L, -\bar{\alpha}/8]$. The definition (53) then gives

$$|\nabla f(\boldsymbol{\theta}_t)| = \chi_\star, \qquad t = k+1, \dots, T-1.$$

Because $k \le T/2$, substituting $\chi_\star$ from (52) yields on the event (55)

$$\frac{1}{T}\sum_{t=1}^{T-1} |\nabla f(\boldsymbol{\theta}_t)|^2 \ge \frac{T/2-1}{T}\chi_\star^2 \ge \frac{1}{4\sqrt{\delta T}},$$

where $T \ge 4$ was used in the last step. Together with $\mathbb{P}(\mathcal{S}) \ge \delta/4$, this proves the probability and stationarity assertions.

It remains to compare the construction on a common parameter class with the calibrated upper bound. If $\delta T \ge (8/\bar{\alpha})^4$, then the height in (52) satisfies $\chi_\star \le \bar{\alpha}/8$, so the Lipschitz constant $8\chi_\star/\bar{\alpha}$ of h from (53) is at most one. Thus all these instances satisfy the common constants

$$p = 0, \quad L_0 = L_p = 1, \quad A = 0, \quad B = 1, \quad \rho_1 = 0, \quad \rho_2 = 2, \quad C = 1.$$

Their initial gradient is zero because $h(0) = 0$ in (53), and their initial gap is at most $32\sqrt{8}\,\bar{\alpha} \le 91$. In this specialization and with $d = 1, \beta_1 = 0$, the baseline parameter sum in (16) is therefore bounded by

$$100 + \epsilon + v_{\text{init}}^{-1} + v_{\text{init}}^{-3/2}.$$

Every fixed positive prefactor at most $2^{-100}(100 + \epsilon + v_{\text{init}}^{-1} + v_{\text{init}}^{-3/2})^{-48}$ is consequently admitted by Theorem B.1 for the entire constructed class, independently of T and δ.

Now use $\delta = 8\eta$. For each sufficiently small $\eta > 0$, choose any integer $T \ge 10$ large enough that $8\eta T \ge (8/\bar{\alpha})^4$. The already proved estimates on the event $\mathcal{S}$ from (55) give

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T-1} |f'(\theta_t)|^2 \ge \frac{1}{8\sqrt{2}\sqrt{\eta T}}\right) \ge 2\eta.$$

For every fixed $k \ge 0$, $\eta^{-1/2}/\log^k(e/\eta) \to \infty$ as $\eta \downarrow 0$. Hence the lower threshold eventually exceeds any uniform bound of order $\log^k(e/\eta)/\sqrt{T}$, while the probability of exceeding that bound is strictly larger than η. Any additional term of order $\mathrm{polylog}(1/\eta)/T$ can also be made smaller by increasing T. This establishes the asserted confidence obstruction for the same stationarity statistic and an admissible fixed prefactor. □

F Proof of Proposition B.2

In the one-dimensional constructions below, θ_t, m_t, v_t denote the scalar iterate, momentum and accumulator of Algorithm 1.

Proof. Let $\mathcal{C}_p$ consist of the deterministic quadratic instance

$$f_{\mathrm{q}}(x) = x^2/2, \qquad \mathsf{G}_{\mathrm{q}}(x) = x, \qquad \theta_1 = 1, \tag{56}$$

and the following countable family, indexed by integers $n \ge 1$:

$$h_n(x) = \begin{cases} 0, & x \le -2, \\ x+2, & -2 \le x \le 0, \\ 2^{j+1}, & 2j \le x \le 2j+1, \quad 0 \le j < n, \\ 2^{j+1}(x-2j), & 2j+1 \le x \le 2j+2, \quad 0 \le j < n, \\ 2^{n+1}, & x \ge 2n, \end{cases} \tag{57}$$

$$f_n(x) = \int_{-2}^{x} h_n(y)\,dy, \qquad \theta_1 = 0. \tag{58}$$

The pieces in (57) agree at their common endpoints. In particular, h_n is continuous and nondecreasing, $f_n \in C^1(\mathbb{R})$, $f_n^* = 0$, $f_n(0) = 2$, and $f_n'(0) = 2$.

We first verify the exact smoothness condition for (57)–(58). Suppose $u < v$ and $v - u \le 1$. The interval intersects the interior of at most one ramp $(2j+1, 2j+2)$. If it intersects that ramp, then $u > 2j$, so $h_n(u) \ge 2^{j+1}$. The slope on that ramp is 2^{j+1}, which is at most $h_n(u)^p$ because $p \ge 1$. If the interval instead intersects $(-2, 0)$, the slope there is one; all remaining slopes on the interval are zero. Thus

$$0 \le h_n(v) - h_n(u) \le \big(1 + h_n(u)^p\big)(v - u).$$

Monotonicity gives the same bound with $h_n(v)^p$ on the right, proving Assumption 3.2 in both orders. The quadratic satisfies the same assumption since $|f_{\mathrm{q}}'(u) - f_{\mathrm{q}}'(v)| = |u - v|$ by (56).

For the staircase gradient h_n in (57), write $h = h_n(x)$ and define the oracle by

$$\mathsf{G}_n(x) = \begin{cases} h, & 0 \le h < 2, \\ -h \quad \text{with probability } 1 - h^{-1}, & h \ge 2, \\ 2h^2 - h \quad \text{with probability } h^{-1}, & h \ge 2. \end{cases} \tag{59}$$

At different calls use independent uniform random variables to select the branch. For $h = h_n(x) \ge 2$, the law in (59) gives

$$\mathbb{E}\mathsf{G}_n(x) = h, \qquad \mathbb{E}\mathsf{G}_n(x)^2 = 4h^3 - 3h^2 \le 4h^3.$$

For $0 \le h < 2$, the deterministic branch of (59) has second moment $h^2 \le 4h^3 + 1$: when $h \le 1$, use $h^2 \le 1$, and when $h \ge 1$, use $h^2 \le h^3$. The same two cases, with $|x|$, show $x^2 \le 4|x|^3 + 1$ for the quadratic in (56). Hence all members of $\mathcal{C}_p$, given by (56), (58), and (59), satisfy the stated unbiasedness and ABC conditions.

For the quadratic in (56), as long as the preceding iterates lie in $[1/2, 1]$, the momentum recursion of Algorithm 1 gives $0 \le m_t \le 1$. Therefore

$$0 \le \theta_t - \theta_{t+1} = \frac{\alpha m_t}{\sqrt{v_t} + \epsilon} \le \frac{\alpha}{\epsilon}.$$

By induction, the first $\min\{T-1, \lfloor \epsilon/(2\alpha) \rfloor + 1\}$ iterates lie in $[1/2, 1]$. For $T \ge 2$, with $\alpha = \bar{\alpha}_T/\sqrt{T}$ as in Proposition B.2, this proves

$$\frac{1}{T}\sum_{t=1}^{T-1} |f_{\mathrm{q}}'(\theta_t)|^2 \ge \min\left\{\frac{1}{8}, \frac{\epsilon}{8\bar{\alpha}_T\sqrt{T}}\right\}. \tag{60}$$

In particular, when

$$\bar{\alpha}_T \le \frac{2^{16}(1 + \sqrt{v_{\mathrm{init}}} + \epsilon)^2}{(1-\beta_1)^2}\, T^{-1/6}, \tag{61}$$

the bound in (21) follows from (60). Indeed, $\epsilon(1-\beta_1)^2/[2^{19}(1 + \sqrt{v_{\mathrm{init}}} + \epsilon)^2] < 1/8$.

For larger stepsizes, namely those violating (61), consider the deterministic path obtained by always selecting the negative branch $-h_n(\theta_t)$ of (59) until the path reaches $2n$. Call the intervals

$$[2j, 2j+2), \qquad 0 \le j < n, \tag{62}$$

the staircase blocks. Extend the same deterministic recursion for as many steps as needed to define its hitting times

$$\widehat{\tau}_j := \inf\{t \ge 1 : \theta_t \ge 2j\}, \qquad 0 \le j \le n. \tag{63}$$

The infimum of the empty set is interpreted as infinity. We next prove that these times are finite and that

$$\widehat{\tau}_{j+1} - \widehat{\tau}_j \le \frac{2^{16}(1 + \sqrt{v_{\mathrm{init}}} + \epsilon)^2}{(1-\beta_1)^2}\left(\frac{1}{\min\{1, \bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{2^{j+1}\min\{1, \bar{\alpha}_T\}}\right), \quad 0 \le j < n. \tag{64}$$

Along the negative-branch path of (59), $m_t < 0$, $|m_t| \ge (1-\beta_1)h_n(\theta_t)$, and $\theta_{t+1} > \theta_t$. The bound for the hitting times in (63) is proved inductively. A block in (62) that is skipped has zero residence time, so suppose $\widehat{\tau}_j < \widehat{\tau}_{j+1}$. In each preceding occupied block k, the definition (57) gives $h_n(\theta_s) < 2^{k+2}$. The induction hypothesis, namely (64) for the earlier blocks, and geometric sums give

$$\frac{1}{4^{j+1}}\sum_{s < \widehat{\tau}_j} h_n(\theta_s)^2 \le \frac{2^{18}(1 + \sqrt{v_{\mathrm{init}}} + \epsilon)^2}{(1-\beta_1)^2}\left(\frac{1}{3\min\{1, \bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{2^{j+1}\min\{1, \bar{\alpha}_T\}}\right). \tag{65}$$

For $j = 0$, the sum on the left is empty, which starts the induction.

Let

$$N_j := \left\lfloor \frac{2^{16}(1+\sqrt{v_{\text{init}}}+\epsilon)^2}{(1-\beta_1)^2} \left(\frac{1}{\min\{1,\bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{2^{j+1}\min\{1,\bar{\alpha}_T\}} \right) \right\rfloor, \tag{66}$$

the integer part of the right side of (64). Suppose the path has not left block j after N_j further updates. For all those updates, $2^{j+1} \le h_n(\theta_t) < 2^{j+2}$, and the accumulator recursion of Algorithm 1 and (65) imply

$$v_t \le v_{\text{init}} + \frac{2^{19}(1+\sqrt{v_{\text{init}}}+\epsilon)^2\, 4^{j+1}}{(1-\beta_1)^2 T} \left(\frac{1}{\min\{1,\bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{2^{j+1}\min\{1,\bar{\alpha}_T\}} \right).$$

Here we used $v_t \le v_{\text{init}} + T^{-1}\sum_{s\le t} h_n(\theta_s)^2$.

To verify that the N_j updates in (66) traverse a block of (62), consider their total displacement. Since the quantity whose integer part defines N_j is at least 2^{16}, its integer part is at least half that quantity. Also $\sqrt{v_{\text{init}}}+\epsilon \le 1+\sqrt{v_{\text{init}}}+\epsilon$. Thus the total displacement is at least

$$\frac{\dfrac{2^{15}(1+\sqrt{v_{\text{init}}}+\epsilon)^2}{1-\beta_1}\left(\dfrac{1}{\min\{1,\bar{\alpha}_T\}} + \dfrac{\sqrt{T}}{2^{j+1}}\right)}{\dfrac{(1+\sqrt{v_{\text{init}}}+\epsilon)\sqrt{T}}{2^{j+1}} + \dfrac{2^8\sqrt{8}(1+\sqrt{v_{\text{init}}}+\epsilon)}{1-\beta_1}\sqrt{\dfrac{1}{\min\{1,\bar{\alpha}_T\}^2} + \dfrac{\sqrt{T}}{2^{j+1}\min\{1,\bar{\alpha}_T\}}}}. \tag{67}$$

The square root in the denominator satisfies

$$\sqrt{\frac{1}{\min\{1,\bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{2^{j+1}\min\{1,\bar{\alpha}_T\}}} \le \frac{1}{\min\{1,\bar{\alpha}_T\}} + \frac{\sqrt{T}}{2^{j+1}}.$$

Indeed, the square of the right side exceeds the radicand by a nonnegative quantity. The other term in the denominator is bounded using $\sqrt{T}/2^{j+1} \le 1/\min\{1,\bar{\alpha}_T\} + \sqrt{T}/2^{j+1}$. Consequently (67) is at least

$$\frac{2^{15}(1+\sqrt{v_{\text{init}}}+\epsilon)}{(1-\beta_1)+2^8\sqrt{8}} \ge \frac{2^{15}}{1+2^{10}} > 2.$$

This contradicts residence in a block of width two (see (62)) and proves (64), including finiteness of the next hitting time. This induction also allows arbitrary overshoots and skipped blocks.

Summing (64) over the blocks in (62) and using $\sum_{j\ge 0} 2^{-j-1} = 1$ gives, for the hitting time defined in (63),

$$\widehat{\tau}_n - 1 \le \frac{2^{16}(1+\sqrt{v_{\text{init}}}+\epsilon)^2}{(1-\beta_1)^2}\left(\frac{n}{\min\{1,\bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{\min\{1,\bar{\alpha}_T\}}\right). \tag{68}$$

The adverse path and its hitting times in (63) are deterministic. Let

$$E_n := \{\mathbf{g}_t = -h_n(\theta_t) \text{ for every } 1 \le t < \widehat{\tau}_n\}, \tag{69}$$

the event that the oracle in (59) selects its negative branch up to that hitting time. Since $\log(1-x) \ge -2x$ for $0 \le x \le 1/2$, the residence bound (64) gives

$$\begin{aligned}\log \mathbb{P}(E_n) &\ge -2\sum_{j=0}^{n-1} \frac{\widehat{\tau}_{j+1}-\widehat{\tau}_j}{2^{j+1}} \\ &\ge -\frac{2^{18}(1+\sqrt{v_{\text{init}}}+\epsilon)^2}{(1-\beta_1)^2}\left(\frac{1}{\min\{1,\bar{\alpha}_T\}^2} + \frac{\sqrt{T}}{\min\{1,\bar{\alpha}_T\}}\right).\end{aligned} \tag{70}$$

Choose the objective in (58) with

$$n = \left\lfloor \frac{(1-\beta_1)^2 T \min\{1,\bar{\alpha}_T\}^2}{2^{18}(1+\sqrt{v_{\text{init}}}+\epsilon)^2} \right\rfloor. \tag{71}$$

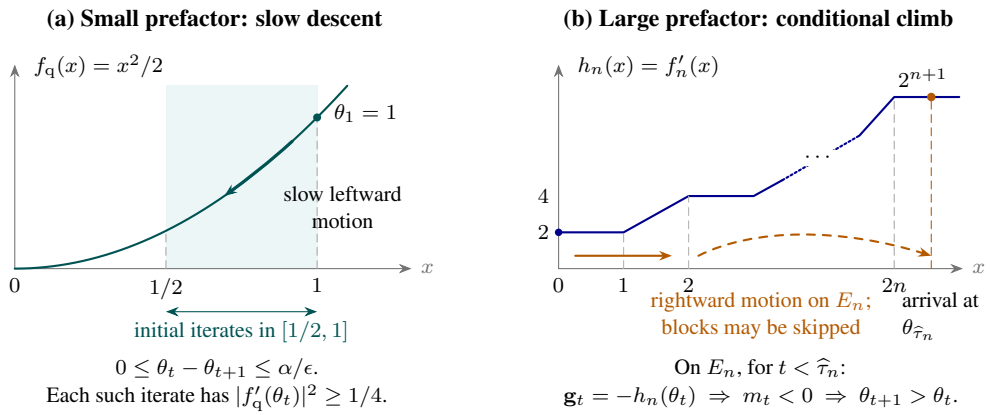
When (61) fails and T is sufficiently large, the integer in (71) is positive,

$$\min\{1, \bar{\alpha}_T\} \geq \frac{2^{16}(1+\sqrt{v_{\text{init}}}+\epsilon)^2}{(1-\beta_1)^2} T^{-1/6}, \tag{72}$$

and (68) yields $\widehat{\tau}_n \leq T/2 + 1 \leq T - 1$. On E_n from (69), (57) gives true gradient 2^{n+1} at the deterministic index $\widehat{\tau}_n$. Therefore

$$\frac{1}{T}\sum_{t=1}^{T-1} \mathbb{E}|f_n'(\theta_t)|^2 \geq \frac{4^{n+1}}{T}\mathbb{P}(E_n).$$

Figure 4 depicts the initial slow descent on the quadratic and the conditional climb to this arrival gradient.



For $h = h_n(x) \geq 2$, the oracle selects $-h$ with probability $1 - h^{-1}$, and $2h^2 - h$ with probability h^{-1}, so its mean is h.

Figure 4: Dynamics of the two instances in Proposition B.2. The quadratic (56) retains an initial segment of iterates in $[1/2, 1]$, as quantified by (60). The staircase is the derivative h_n in (57); its plateaus are joined continuously by linear ramps. The rightward arrows are conditional on the negative-oracle event E_n in (69), with the two branches given by (59). They end at the first arrival $\widehat{\tau}_n$ from (63), where $h_n(\theta_{\widehat{\tau}_n}) = 2^{n+1}$. Choosing n as in (71) places this index before T; its single squared-gradient contribution gives the lower bound $4^{n+1}\mathbb{P}(E_n)/T$. Arrows indicate directions schematically; intermediate staircase blocks are omitted.

Using the choice (71), which gives $n+1 \geq (1-\beta_1)^2 T \min\{1, \bar{\alpha}_T\}^2/[2^{18}(1+\sqrt{v_{\text{init}}}+\epsilon)^2]$, together with the event-probability bound (70), the logarithm of the right side is bounded below by

$$\frac{(1-\beta_1)^2 \log 4}{2^{18}(1+\sqrt{v_{\text{init}}}+\epsilon)^2} T \min\{1, \bar{\alpha}_T\}^2 - \frac{2^{18}(1+\sqrt{v_{\text{init}}}+\epsilon)^2}{(1-\beta_1)^2}\left(\frac{\sqrt{T}}{\min\{1, \bar{\alpha}_T\}} + \frac{1}{\min\{1, \bar{\alpha}_T\}^2}\right) - \log T.$$

This expression is increasing in $\min\{1, \bar{\alpha}_T\} > 0$. Substituting (72) gives

$$\left(\frac{2^{14}(1+\sqrt{v_{\text{init}}}+\epsilon)^2 \log 4}{(1-\beta_1)^2} - 4\right) T^{2/3} - \frac{(1-\beta_1)^2}{2^{14}(1+\sqrt{v_{\text{init}}}+\epsilon)^2} T^{1/3} - \log T.$$

The coefficient of $T^{2/3}$ is positive, so this is nonnegative for all sufficiently large T, uniformly over stepsizes in this case. Hence the expected stationarity average is at least one for this member of $\mathcal{C}_p$, the objective f_n from (58) with n from (71). Together with the quadratic case (60) and the fact that $\epsilon(1-\beta_1)^2/[2^{19}(1+\sqrt{v_{\text{init}}}+\epsilon)^2] < 1$, this proves (21). □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and Section 1 describe trajectory localization, high-probability convergence, the confidence obstruction, and the expectation result for $p < 1$. Section 4 states their probability, parameter, and comparison scopes under Assumptions 3.1–3.3.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The theoretical calibration in Section 3 specifies the horizon-dependent choice $\beta_2 = 1 - 1/T$. The results distinguish the ranges $p < 1$ and $1 \leq p < 2$. The statements and discussion of Propositions B.1 and B.2 specify the algorithm family and quantifiers of the lower bounds. Section 3 states the scope of the finite-horizon calibration, and the remark after Theorem B.1 states that the constants are not optimized.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: Assumptions 3.1–3.3 precede the main results. Section 5 explains the argument, and the appendix gives the auxiliary estimates and proofs of all stated results.

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: The submission is a theoretical contribution; it does not report computational experiments whose claims depend on empirical reproduction.

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

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

   Answer: [N/A]

   Justification: There are no empirical experiments. The stated algorithm, assumptions, and complete mathematical proofs specify the theoretical results.

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: The paper contains no training or benchmarking experiments; Algorithm 1 and the theoretical parameter choices ($\alpha = \bar{\alpha}/\sqrt{T}$, $\beta_2 = 1 - 1/T$, etc.) are specified for the analysis rather than for an empirical protocol.

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: No stochastic experiments are reported; statistical statements are probabilistic guarantees in the theorems (confidence level δ, expectations), not empirical error bars.

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: The paper does not include computational experiments requiring hardware disclosure.

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is mathematical analysis of a standard optimization algorithm under stated assumptions, with no human subjects, crowdsourcing, or high-risk model or dataset release described in the manuscript.

10. **Broader impacts**

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

    Answer: [No]

    Justification: The manuscript does not include a dedicated discussion of societal impacts. The contribution is foundational theory on stochastic optimization; any societal effects would be indirect (e.g., through improved understanding of widely used training dynamics) rather than tied to a specific deployment studied here.

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release models, datasets, or other high-misuse-risk assets.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The work does not build on third-party codebases, datasets, or pretrained model checkpoints for experiments; prior algorithms and results are cited bibliographically in the usual way.

13. **New assets**

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

    Answer: [N/A]

    Justification: No new datasets, code repositories, or model checkpoints are introduced or released with the submission.

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

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

    Answer: [N/A]

    Justification: The research does not involve human subjects or crowdsourcing.

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

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

    Answer: [N/A]

    Justification: The paper does not involve human subjects research.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: Language models assisted proof development and manuscript preparation. The mathematical arguments and their assumptions are provided explicitly in the paper.