Revisiting AdaGrad in Stochastic Convex Optimization: Last Iterates, High Probability, and Lower Bounds

Anonymous authors
Paper under double-blind review

Abstract

AdaGrad and AdaGrad-Norm are widely used adaptive methods, but their precise behavior in stochastic convex optimization remains less understood. We first show that AdaGrad-Norm and AdaGrad do not admit any universal **last-iterate** rate, even under sub-Gaussian noise and bounded iterates. We then prove that bounded variance alone is too weak: even with bounded iterates, it cannot yield **high-probability average-iterate** rates, and without bounded iterates it may not even guarantee convergence in expectation. Besides, we construct tight $\Omega(\log T/\sqrt{T})$ lower bounds for both AdaGrad-Norm and AdaGrad under sub-Gaussian noise, showing that $\log T$ in existing average-iterate upper bounds is unavoidable. Finally, we show that this $\log T$ loss disappears once bounded-iterate condition is imposed: under a general ABC condition, both methods achieve rates $O(1/\sqrt{T})$.

1 Introduction

AdaGrad scales its updates using the gradients observed so far, replacing a prescribed stepsize schedule by a data-dependent normalization. The original coordinate-wise method was developed in online convex optimization by McMahan & Streeter (2010) and Duchi et al. (2011). AdaGrad-Norm instead uses a single accumulator for the cumulative squared gradient norms (Ward et al., 2020). Although the methods share a common regret-analysis framework, their stochastic guarantees depend on the noise model, the output, and the behavior of the iterates.

For smooth convex objectives, average-iterate bounds are available for delayed AdaGrad-type schemes (Li & Orabona, 2019), standard AdaGrad-Norm under suitable tail assumptions (Attia & Koren, 2023; Liu et al., 2023a), and coordinate-wise AdaGrad on bounded domains (Liu et al., 2025). Log-free, noise-adaptive rates also have precedents on bounded domains or under bounded-iterate assumptions (Levy et al., 2018; Vaswani et al., 2020). These guarantees do not automatically describe the final point returned by an implementation. Our central question is therefore whether the standard updates admit an explicit, vanishing *last-iterate* rate under the same smoothness, light-tail, and stability conditions.

Our main result gives a negative answer in the general conditional-oracle model. For every proposed positive sequence $r_T \to 0$, we construct one smooth convex objective and one infinite-horizon, conditionally unbiased sub-Gaussian oracle such that the expected last-iterate error exceeds r_T by an unbounded factor along a deterministic subsequence. All iterates remain inside an arbitrarily prescribed radius. Thus the issue is not merely a missing polynomial exponent: *no prescribed vanishing rate is valid uniformly over this class.* The hard instance itself converges, separating convergence from a uniform quantitative guarantee. Its oracle is time- and history-dependent, so the result does not preclude rates under additional sampling restrictions.

A second question concerns the confidence of an *averaged* output. Averaging often gives guarantees unavailable for the last iterate, but we show that it does not repair the poor confidence dependence of unmodified AdaGrad under finite variance. Our construction combines three properties: a *fixed infinite-horizon instance*, the *uniformly averaged iterate*,

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

and *any prescribed almost-sure trajectory radius*. This output distinction matters: the lower bound in Chezhegov et al. (2025, Theorem B.4), specialized to zero momentum, is stated for the last iterate. A last-iterate lower bound alone does not establish an average-iterate lower bound. Our result excludes polynomial decay in time with only polylogarithmic dependence on the inverse failure probability.

The remaining results place these two limitations in context. Without a common stability radius, finite variance permits constant worst-case expected average error. Even under bounded, hence sub-Gaussian, noise, a scalar construction gives an $\Omega(\log T/\sqrt{T})$ average-error lower bound for the fixed-parameter unprojected updates and a corresponding inverse-square-root SGD schedule. With bounded iterates, by contrast, classical regret analysis yields explicit noise-adaptive average-error bounds under the conditional ABC model. The distinction between a fixed instance and a horizon-dependent family is kept explicit throughout: choosing a different function for each horizon does not establish nonconvergence on one fixed instance.

1.1 Our contributions

Table 1 summarizes the results in the order developed below. Every lower-bound construction is one-dimensional and applies to both AdaGrad-Norm and coordinate-wise AdaGrad.

1. **No universal vanishing last-iterate rate.** Theorem 3.1 shows that every prescribed $r_T \to 0$ is violated by a single fixed sub-Gaussian instance with bounded iterates. Its expected error satisfies $\mathbb{E}[\Delta_{T_n}]/r_{T_n} \to \infty$ along a deterministic subsequence. Delayed bounded pulses keep the accumulator small until an arbitrarily late time; the resulting displacement defeats the proposed rate without preventing convergence of the constructed instance.
2. **A high-probability limitation for the average iterate.** Theorem 3.2 excludes polynomial-in-T rates with only polynomial dependence on $\log(1/\delta)$ for *uniform averaging*, on one fixed infinite-horizon finite-variance instance inside any prescribed radius $D > 0$. A pulse cascade increases the accumulator through alternating small displacements. The combination of fixed-instance quantifiers, averaged output, and arbitrary stability radius distinguishes the statement from the last-iterate lower bound of Chezhegov et al. (2025); the general heavy-tail difficulty itself is already known.
3. **Average-error lower bounds without a uniform stability radius.** Theorem 3.3 gives a horizon-dependent family with constant worst-case expected average error under independent identically distributed additive finite-variance noise. Theorem 3.4 gives $\Omega(\eta\sigma \log T/\sqrt{T})$ under sub-Gaussian noise, even from a minimizer. Both concern fixed algorithm parameters. The latter construction also proves a smooth-convex average-error lower bound for inverse-square-root SGD; the final part of Appendix D shows that its SGD version can use independent bounded additive noise.
4. **An explicit noise-adaptive ABC upper bound.** Under bounded iterates, Theorem 3.5 gives $O(T^{-1} + \sqrt{C/T})$ expected average error, with initialization and the ABC coefficients explicit. The $C = 0$ case gives $O(T^{-1})$ even with state-dependent noise. This is a unified extension of classical regret and self-bounding analysis (Duchi et al., 2011; Levy et al., 2018; Vaswani et al., 2020), not a new noise-adaptation principle. It assumes, rather than proves, trajectory stability.

1.2 Related work

Last-iterate convergence versus uniform rates. Last-iterate convergence has been studied extensively for stochastic gradient methods (Bertsekas & Tsitsiklis, 2000; Zhang, 2004; Shamir & Zhang, 2013; Orabona, 2020a;b; Liu & Yuan, 2022), stochastic mirror descent (Liu & Zhou, 2023), and momentum methods based on follow-the-regularized-leader updates (Li et al., 2022). For adaptive methods, Li & Orabona (2019) analyze delayed updates, while Jin et al. (2022); Jin & Wang (2026) study qualitative stochastic convergence of current-gradient AdaGrad under their respective assumptions. The deterministic nonsmooth analysis of Preobrazhenskaia et al. (2026) concerns another setting. Theorem 3.1 does not assert a first

Theorem	Assumptions	Guarantee type	Result
3.1	Sub-Gaussian + bounded iterates	Last iterate	No universal vanishing rate[a]
3.2	Finite variance + bounded iterates	Average iterate, high probability	No polynomial rate[b] with polylog$(1/\delta)$
3.3	Finite variance	Average iterate	Lower bound[c] $\Omega(1)$
3.4	Sub-Gaussian	Average iterate	Lower bound[c] $\Omega(\log T/\sqrt{T})$
3.5	ABC condition + bounded iterates	Average iterate	Upper bound $O(T^{-1}+\sqrt{C/T})$

Table 1: Results for standard unprojected AdaGrad-Norm and AdaGrad on smooth convex objectives with conditionally unbiased oracles. Errors are in expectation unless high probability is specified. [a] For each prescribed $r_T \to 0$, one fixed infinite-horizon instance satisfies $\mathbb{E}[\Delta_{T_n}]/r_{T_n} \to \infty$. [b] One fixed infinite-horizon instance, uniformly averaged output, and any prescribed radius $D > 0$; the excluded bounds have polynomial decay in T and only polynomial dependence on $\log(1/\delta)$. [c] Horizon-dependent families with fixed algorithm parameters and no common trajectory radius. The upper bound is an explicit ABC extension of classical analysis.

qualitative convergence theorem. It instead rules out *every prescribed uniform vanishing rate* under smoothness, bounded iterates, and a general conditional sub-Gaussian oracle. The distinction is substantive: almost-sure convergence of individual trajectories does not supply a uniform expected rate for the entire oracle class.

High-probability guarantees and the averaged output. Foundational and subsequent high-probability analyses of stochastic mirror descent and gradient methods include Nemirovski et al. (2009); Gorbunov et al. (2020; 2021); Li & Orabona (2020). Clipping can control heavy-tailed gradients (Cutkosky & Mehta, 2021; Liu et al., 2024); high-probability analyses for adaptive methods include Kavis et al. (2022); Attia & Koren (2023); Hong & Lin (2024b;a). Most directly, Chezhegov et al. (2025) establish poor confidence dependence for unmodified adaptive methods and show how clipping improves it. Their Theorem B.4 concerns the event $f(x_K) - f(x^\star) \geq \varepsilon$ for the *last iterate* of their momentum-AdaGrad update; setting the momentum coefficient to zero recovers AdaGrad-Norm. Our Theorem 3.2 directly controls $f(\bar{x}_T) - f(x^\star)$ instead, on one fixed infinite-horizon instance within an arbitrary prescribed radius. The proof keeps the post-cascade iterates on the same side of the minimizer, so averaging cannot cancel the long-lived displacement. This is an averaged-output obstruction, not an inference from a last-iterate event. It remains compatible with weaker confidence bounds obtained from expectation by Markov's inequality.

Lower bounds for SGD and adaptive methods. Classical oracle-complexity and information-theoretic results establish an $\Omega(T^{-1/2})$ stochastic term for suitable convex problem classes (Nemirovskij & Yudin, 1983; Agarwal et al., 2009; Raginsky & Rakhlin, 2011). The logarithmic lower bound of Harvey et al. (2019) concerns a *nonsmooth last iterate*; it does not give a smooth-convex lower bound for uniform averaging. Lower bounds for nonconvex stationarity (Arjevani et al., 2023; Jiang et al., 2025) and structured nonconvex objectives (Saad et al., 2025) concern different settings. Recent smooth-convex SGD analyses include last-iterate upper bounds (Liu & Zhou, 2023; Garrigos et al., 2025). We are not aware of an earlier $\Omega(\log T/\sqrt{T})$ lower bound for the exact combination considered here: *uniformly averaged unprojected SGD, a smooth convex objective, bounded noise, and a fixed inverse-square-root stepsize schedule*. Theorem 3.4 and its SGD extension in Appendix D establish this statement directly. This scoped comparison is not a lower bound for all SGD schedules: properly tuned finite-horizon steps and averaging admit different guarantees (Nemirovski et al., 2009).

Adaptive methods and noise adaptation. The online analysis of McMahan & Streeter (2010); Duchi et al. (2011) provides the starting point for adaptive normalization. Related developments include adaptive steplength rules and variants of AdaGrad, RMSProp, and Adam (Yousefian et al., 2012; Mukkamala & Hein, 2017; Reddi et al., 2019), delayed AdaGrad-type schemes (Li & Orabona, 2019), and standard updates under stronger tail assumptions (Attia & Koren, 2023; Liu et al., 2023a). On a bounded convex set, Levy et al. (2018) obtain a log-free $O(LD^2/T + \sigma D/\sqrt{T})$ expected bound for scalar AdaGrad. Under bounded iterates and component-function assumptions, Vaswani et al. (2020) derive an analogous noise-adaptive transition using regret, self-bounding, and a quadratic inequality. Their residual-noise parameter differs from the conditional variance bound used here. Coordinate-wise adaptation under anisotropic smoothness and noise is studied by Liu et al. (2025). Our Theorem 3.5 keeps constants explicit under the global conditional ABC model. A broader nonconvex literature studies scalar and coordinate-wise normalization (Défossez et al., 2020; Kavis et al., 2022; Faw et al., 2022; 2023; Liu et al., 2023b; Yang et al., 2023; Wang et al., 2023; Hong & Lin, 2024b).

2 Preliminaries and setup

Problem and notation. We consider

$$\min_{x\in\mathbb{R}^d} f(x), \tag{Opt}$$

where f is differentiable, convex, and attains its minimum. Fix $x^\star \in \arg\min f$ and write

$$f^\star = f(x^\star), \qquad \Delta_t = f(x_t) - f^\star, \qquad \bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

When a hard objective depends on the horizon, we write f_T and $f_T^\star$ explicitly. The corresponding trajectory is always generated on that chosen instance. We use Euclidean norms, $u_+ = \max\{u, 0\}$, and the usual floor and ceiling functions. For nonnegative sequences, $a_T = O(b_T)$ and $a_T = \Omega(b_T)$ denote eventual upper and lower bounds up to positive constants; $a_T \asymp b_T$ means both. Unless stated otherwise, constants may depend on fixed problem and algorithm parameters, but not on T or the failure probability.

Oracle and filtration. At x_t, the oracle returns $g_t = \nabla f(x_t) + \xi_t$. The initialization x_1 is deterministic, $\mathcal{F}_0$ is trivial, and $\mathcal{F}_t = \sigma(g_1, \dots, g_t)$, so x_t is $\mathcal{F}_{t-1}$-measurable. Every conditional assumption below refers to the history before the current query and holds for all $t \geq 1$. The oracle may depend on time and history; independent identically distributed noise is imposed only where stated. We distinguish one fixed infinite-horizon instance from a family whose objective or oracle changes with T.

To conduct the theoretical analysis, we introduce some assumptions below.

Assumption 1. For some $L > 0$, for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Assumption 2. The oracle satisfies $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla f(x_t)$.

These two assumptions are used throughout. The results differ in which of the following moment conditions is imposed.

Assumption 3 (Sub-Gaussian-square noise)**.** For some $\sigma > 0$,

$$\mathbb{E}\left[\exp\left(\frac{\|\xi_t\|^2}{\sigma^2}\right)\middle|\mathcal{F}_{t-1}\right] \leq e.$$

Assumption 4 (Bounded variance)**.** For some $\sigma \geq 0$, $\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_{t-1}] \leq \sigma^2$.

Assumption 5 (ABC condition)**.** There are $A, B, C \geq 0$ such that

$$\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_{t-1}] \leq A\Delta_t + B\|\nabla f(x_t)\|^2 + C. \tag{1}$$

Remark 2.1. Conditional Jensen's inequality shows that Assumption 3 implies Assumption 4. Together with unbiasedness, bounded variance implies the ABC condition with $(A, B, C) = (0, 1, \sigma^2)$. The ABC framework therefore includes both preceding noise models (Khaled & Richtárik, 2020). The case $C = 0$ permits state-dependent noise; it does not require every gradient to be deterministic.

Algorithms. For $b_0, \eta > 0$, AdaGrad-Norm uses

$$b_t^2 = b_0^2 + \sum_{s=1}^{t} \|g_s\|^2, \qquad x_{t+1} = x_t - \eta \frac{g_t}{b_t}. \tag{AdaGrad-Norm}$$

For $b_{0,i}, \eta > 0$, coordinate-wise AdaGrad uses

$$b_{t,i}^2 = b_{0,i}^2 + \sum_{s=1}^{t} g_{s,i}^2, \qquad x_{t+1,i} = x_{t,i} - \eta \frac{g_{t,i}}{b_{t,i}}, \quad 1 \le i \le d. \tag{AdaGrad}$$

Both updates are unprojected, and each denominator includes the current gradient; it is generally not $\mathcal{F}_{t-1}$-measurable. In dimension one, the algorithms coincide when their initial accumulators agree. All lower bounds exploit this common scalar update, whereas the upper bound treats the two normalizations separately and allows arbitrary positive $b_{0,i}$.

3 Main results

We first rule out a universal last-iterate rate, then show that averaging does not remove the high-probability obstruction under finite variance. The next two results address expected average error without a common stability radius. The final theorem gives the complementary noise-adaptive upper bound when stability is available.

3.1 No universal vanishing last-iterate rate

Our first result concerns the expected error at the actual final iterate. Neither smoothness, sub-Gaussian noise, nor an arbitrarily small trajectory radius supplies a universal vanishing rate in the conditional-oracle model.

Theorem 3.1 (Arbitrarily slow last-iterate convergence)**.** *Fix $L, \sigma, b_0, \eta, D > 0$. For every prescribed sequence $r_T > 0$ with $r_T \to 0$, there exist a single convex L-smooth quadratic $f : \mathbb{R} \to \mathbb{R}$, a single infinite-horizon oracle satisfying Assumptions 2 and 3, and deterministic $T_n \to \infty$, such that both algorithms, initialized at $x_1 = x^\star = 0$, satisfy $|x_t - x^\star| \le D$ almost surely for all $t \ge 1$ and*

$$\frac{\mathbb{E}[f(x_{T_n}) - f^\star]}{r_{T_n}} \longrightarrow \infty. \tag{2}$$

Proof sketch. At the minimizer of a small quadratic, both the iterate and accumulator stay unchanged until a nonzero pulse occurs. At scheduled times τ_n, allow symmetric noise of fixed magnitude with probability $q/(n+1)^2$, and switch off all noise after the first pulse. That pulse creates a fixed displacement, followed by deterministic contraction. Its occurrence probability gives $\mathbb{E}[\Delta_{\tau_n+1}] \ge c_1/(n+1)^2$; see (17). Choosing τ_n so that $r_{\tau_n+1} \le (n+1)^{-4}$ proves the claim. The delay, rather than growing noise magnitude, defeats the rate. Appendix A also verifies asymptotic convergence of this particular instance. □

Remark 3.1. The quantifiers select an instance after the proposed sequence r_T, but then hold it fixed for all evaluation times. Thus the theorem is stronger than excluding every polynomial exponent separately, and is not a horizon-by-horizon change of oracle. The quantifier is $\forall (r_T \to 0)\, \exists \mathcal{I}\, \exists (T_n)$, not the existence of one instance slower than every possible sequence. The constructed instance satisfies $\mathbb{E}[\Delta_t] \to 0$; no general qualitative convergence theorem is asserted. An explicit schedule, independent of a proposed rate, gives

$$\mathbb{E}[\Delta_{T_n}] \ge \frac{c}{(1 + \log\log T_n)^2}, \qquad T_n = \lceil \exp(\exp n) \rceil + 1, \tag{3}$$

where $c > 0$ depends only on L, σ, b_0, η, D. The schedule for an arbitrary r_T need not be this particular one. The obstruction uses delayed, time- and history-dependent noise, not unbounded pulse magnitudes; it does not rule out rates under additional oracle restrictions.

Remark 3.2. The construction uses a fixed objective and bounded pulses. The objective is a strongly convex quadratic with curvature $\ell = \min\{L, b_0/(2\eta)\}$, independent of the proposed rate. Its minimizer, the initialization, the radius, and the pulse magnitude M are also fixed independently of r_T. In fact, the construction satisfies the stronger pathwise bounds $|\xi_t| \le M$ and $|g_t| \le \ell d_0 + M$, with $d_0 \le D$. Only the scheduled pulse times depend on r_T. Thus the obstruction is neither a flattening sequence of objectives nor increasingly large gradients. It arises because elapsed time need not force growth of the adaptive accumulator: the oracle can remain noiseless at a minimizer for an arbitrarily long interval. The uniform moment assumptions control pulse size and probability but do not impose a clock on when stochastic variation must occur.

Remark 3.3. Averaging can still have a rate because it dilutes a late excursion: before the pulse, the objective error is zero and the accumulator stays at b_0; immediately afterward, the last iterate has a fixed error. A late evaluation can therefore detect the entire displacement, whereas uniform averaging dilutes a short late excursion by the horizon. This explains why Theorem 3.5 can give an explicit expected average-iterate rate on the same class. An average-iterate upper bound cannot be converted into a last-iterate rate merely by substituting the final point.

3.2 A high-probability limitation even for the average iterate

We next weaken the noise assumption to finite variance and ask for a high-probability guarantee for the *uniform average*, rather than the last iterate. The following result combines one fixed infinite-horizon instance, this averaged output, and any prescribed bounded-trajectory radius.

Theorem 3.2 (No polynomial average-iterate rate with polylogarithmic confidence)**.** *Fix $L, \sigma, b_0, \eta, D > 0$. There exist a fixed one-dimensional convex L-smooth quadratic and a fixed infinite-horizon oracle satisfying Assumptions 2 and 4, such that both algorithms, initialized at $x_1 = x^\star = 0$, satisfy $|x_t - x^\star| \le D$ almost surely for every $t \ge 1$. There also exist deterministic $T_n \to \infty$, $\delta_n \in (0, 1/2)$, and constants $c_\star, K > 0$ such that*

$$\begin{aligned} \mathbb{P}\{f(\bar{x}_{T_n}) - f^\star \ge c_\star\} &\ge 2\delta_n, \\ \log(1/\delta_n) &\le K(1 + \log T_n)^2. \end{aligned} \tag{4}$$

Consequently, no bound

$$f(\bar{x}_T) - f^\star \le K_0\big[T^{-a} + P(\log(1/\delta))T^{-b}\big], \qquad a, b > 0, \tag{5}$$

can hold with probability at least $1 - \delta$ for every $T \ge 1$ and $\delta \in (0, 1/2)$, with $K_0 < \infty$ and a fixed polynomial P, nonnegative on $[\log 2, \infty)$, independent of T, δ.

Proof sketch. A single large gradient can move the iterate almost η, so it cannot respect an arbitrary radius $D < \eta$. Instead, arrange a finite random cascade whose designated path alternates between 0 and $-d_0$, while the accumulator grows geometrically; see (19). Rare pulses preserve a common conditional variance bound, and every departure switches off the noise. Stopping after an odd number n of pulses leaves a nonzero displacement and an exponentially large accumulator, so the displacement survives for an exponentially long horizon. The designated event has probability at least $\exp(-O(n^2))$, yielding (4). Appendix B checks all paths and the fixed-oracle quantifier. □

Remark 3.4. Our result concerns the averaged output on one fixed bounded instance. The last-iterate event in Chezhegov et al. (2025, Theorem B.4) does not by itself lower-bound $f(\bar{x}_T) - f^\star$. Convexity gives an *upper* bound on $f(\bar{x}_T)$ by the mean of the iterate values, not the lower bound needed here. Our cascade instead creates a displacement that persists for a substantial fraction of the horizon. On the designated event all iterates are nonpositive

and the late ones have magnitude at least $d_0/2$, so their uniform average has magnitude at least $d_0/4$; see (25). The argument directly excludes the stated confidence dependence even for this more forgiving output.

The fixed-instance quantifier is particularly strong here. Once the cascade law has been chosen from L, σ, b_0, η, D, the *same* instance excludes every pair of positive exponents and every fixed polynomial appearing in (5). It is not retuned for a proposed confidence bound. Small oscillations consume normalization while leaving little net displacement; after the last designated pulse, the accumulated normalization prevents rapid recovery. This separates trajectory containment from concentration of the averaged objective error. In contrast, a single large initial pulse need not respect a prescribed radius smaller than η.

Remark 3.5. The obstruction concerns confidence dependence rather than nonconvergence: the levels δ_n shrink with the horizons. The theorem does not say that the averaged error fails to vanish for a fixed δ, nor that averaging fails almost surely on the hard instance. Indeed, Theorem 3.5 and Markov's inequality give, in dimension one, for every $\delta > 0$,

$$\mathbb{P}\left\{ f(\bar{x}_T) - f^\star > \frac{1}{\delta}\left(\frac{2L\Gamma^2 + \Gamma b_0}{T} + \frac{\Gamma\sigma}{\sqrt{T}} \right) \right\} \le \delta, \quad \Gamma = \frac{D^2}{2\eta} + \eta. \tag{6}$$

Its $1/\delta$ dependence is outside the class excluded by (5). Thus the substantive distinction is that *uniform averaging does not upgrade finite variance to a polynomial rate with polylogarithmic confidence dependence*, even on one fixed bounded instance.

3.3 Finite variance permits constant worst-case average error

The previous theorem retains a bounded trajectory and isolates confidence dependence. If that stability condition is removed, finite variance alone does not even supply a vanishing *uniform worst-case* expected average-error bound for fixed algorithm parameters.

Theorem 3.3 (Constant worst-case error under bounded variance)**.** *Fix $\sigma, b_0, \eta > 0$. For every sufficiently large integer T, there exist a one-dimensional convex 1-smooth function f_T with $f_T^\star = 0$ and an oracle with independent identically distributed additive noise satisfying Assumptions 2 and 4, such that both algorithms, initialized at $x_1 = 1$ with $x^\star = 0$, satisfy*

$$\mathbb{E}[f_T(\bar{x}_T) - f_T^\star] \ge \frac{\eta\sigma^2}{256\sqrt{b_0^2 + \sigma^2/16}}. \tag{7}$$

The objective and noise law may depend on T, but the initial distance, algorithm parameters, variance bound, and displayed lower-bound constant do not.

Proof sketch. Use a smooth hinge with slope $a = \sigma/(4\sqrt{T})$. In its linear region, the oracle returns $-a$ with probability $1 - 1/T$; a rare large positive correction preserves unbiasedness and bounded variance. With probability at least $1/4$, no correction occurs before T. On that event the accumulator stays bounded, the iterates move away from the minimizer, and their average displacement is of order $\eta a T$. The objective gap is therefore of order $\eta a^2 T$, a positive constant in T. Equations (29)–(30) in Appendix C make the constants explicit. □

Remark 3.6. This horizon-wise lower bound concerns uniform worst-case convergence, not a failure of convergence on one fixed instance. Unlike the preceding pulse constructions, it uses independent identically distributed additive noise on every fixed instance. The initial distance remains one across the family; the obstruction cannot be attributed to a growing initialization error.

3.4 A logarithmic average-error lower bound under sub-Gaussian noise

The finite-variance counterexamples allow increasingly large gradients. The next construction shows that a loss persists even with uniformly bounded noise. A single scalar, fixed-magnitude oracle proves the result for both algorithms, without an auxiliary coordinate.

Theorem 3.4 (A scalar logarithmic average-error lower bound)**.** *Fix $L, \sigma, b_0, \eta > 0$. For every sufficiently large integer T, there exist a convex L-smooth function $f_T : \mathbb{R} \to \mathbb{R}$ with $f_T^\star = 0$ and an oracle satisfying Assumptions 2 and 3, such that both AdaGrad-Norm and AdaGrad, initialized at $x_1 = x^\star = 0$, satisfy*

$$\mathbb{E}[f_T(\bar{x}_T) - f_T^\star] \geq c\eta\sigma\frac{\log T}{\sqrt{T}}, \tag{8}$$

where $c > 0$ is universal. The objective and oracle may depend on T, but L, σ, b_0, η do not. No horizon-independent bounded-iterate radius is assumed.

Proof sketch. Choose a smooth hinge with slope $a \asymp \sigma\sqrt{\log T/T}$ and an unbiased oracle with $g_t \in \{-\gamma, \gamma\}$, where $\gamma = \sigma/2$. Then $b_t^2 = b_0^2 + t\gamma^2$ is deterministic. The average splits into a mean-zero martingale term and a nonnegative drift. The squared martingale weights sum to order $\eta^2 \log T$, so second- and fourth-moment estimates give a positive fluctuation of order $\eta\sqrt{\log T}$; see (37)–(40). Choosing the slope constant small makes the drift smaller than this fluctuation. Multiplication by a yields (8). Appendix D verifies the oracle and smoothing conditions. □

Remark 3.7. The scope of the logarithmic lower bound is smooth-convex average function value, rather than the nonconvex stationarity metric of Jiang et al. (2025). It does not exclude horizon-dependent parameter tuning, alternative outputs, or modifications such as projection. In particular, it does not establish that every logarithm in every AdaGrad upper bound is necessary.

Remark 3.8. The same construction has a consequence for inverse-square-root SGD: since $g_t^2 = \gamma^2$ on this oracle, AdaGrad is exactly unprojected SGD with deterministic stepsizes $\eta/\sqrt{b_0^2 + t\gamma^2}$. More generally, fix $L, \sigma, \lambda, \tau > 0$ and consider

$$x_{t+1} = x_t - \frac{\lambda}{\sqrt{t+\tau}}g_t. \tag{9}$$

For every sufficiently large T, there exists a convex L-smooth scalar objective f_T, with $x_1 = x^\star = 0$, for which even the independent additive-noise oracle $g_t = f_T'(x_t) + \xi_t$, $\mathbb{P}(\xi_t = \sigma/2) = \mathbb{P}(\xi_t = -\sigma/2) = 1/2$, satisfies

$$\mathbb{E}[f_T(\bar{x}_T) - f_T^\star] \geq c_{\mathrm{SGD}}\lambda\sigma^2\frac{\log T}{\sqrt{T}}, \tag{10}$$

where $c_{\mathrm{SGD}} > 0$ is universal. The final part of Appendix D gives the full reduction and checks the independent-noise version. This is a lower bound for a smooth-convex *average iterate*, not the nonsmooth last-iterate lower bound of Harvey et al. (2019). We are not aware of an earlier result for this exact combination of assumptions and schedule. The qualification on the schedule is essential: the claim does not extend to all SGD stepsizes, suffix averages, or projected variants.

Remark 3.9. The stepsize schedule matters and cannot be omitted from the statement. For example, for SGD with a constant step $0 < \alpha \leq 1/(2L)$ on a fixed horizon and a conditionally unbiased oracle of variance at most σ^2, the squared-distance recursion and $\|\nabla f(x)\|^2 \leq 2L(f(x) - f^\star)$ give

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{\|x_1 - x^\star\|^2}{\alpha T} + \alpha\sigma^2. \tag{11}$$

Indeed, one update decreases expected squared distance by at least $\alpha\mathbb{E}[\Delta_t] - \alpha^2\sigma^2$; summation and convexity prove the display. A horizon-dependent choice $\alpha \asymp T^{-1/2}$ therefore gives $O(T^{-1/2})$. On the lower-bound instance $x_1 = x^\star$, even the initial-distance term vanishes. The logarithm in (10) is a limitation of the fixed inverse-square-root schedule with uniform averaging, not of the stochastic oracle problem itself. The final part of Appendix D also verifies the integrability and distance calculation behind (11).

3.5 Noise-adaptive upper bounds under the ABC condition

The preceding results distinguish instability, heavy-tail confidence dependence, and last-iterate behavior. None prevents a log-free expected average-error bound under bounded iterates. The final theorem makes this complementary guarantee explicit using the classical regret and self-bounding argument (Duchi et al., 2011; Levy et al., 2018; Vaswani et al., 2020; Liu et al., 2025).

Theorem 3.5 (Noise-adaptive average-iterate upper bounds)**.** *Under Assumptions 1, 2, and 5, let $Q = A + 2LB$. For every integer $T \geq 1$:*

1. ***AdaGrad-Norm.*** *If $\|x_t - x^\star\| \leq D$ almost surely for all t, set $\Gamma = D^2/(2\eta) + \eta$. Then*

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{\Gamma^2 Q + \Gamma b_0}{T} + \Gamma\sqrt{\frac{C}{T}}. \tag{12}$$

2. ***AdaGrad.*** *If $|x_{t,i} - x_i^\star| \leq D_i$ almost surely for all t, i, set $\Gamma_i = D_i^2/(2\eta) + \eta$, $\Lambda = (\sum_i \Gamma_i^2)^{1/2}$, and $\beta_0 = (\sum_i b_{0,i}^2)^{1/2}$. Then*

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{\Lambda^2 Q + \Lambda \beta_0}{T} + \Lambda\sqrt{\frac{C}{T}}. \tag{13}$$

The radii are nonnegative, finite, and independent of t. In particular, $C = 0$ gives $O(T^{-1})$; bounded variance permits $Q = 2L$, $C = \sigma^2$ and gives $O(T^{-1} + \sigma T^{-1/2})$.

Proof sketch. The key is to retain the objective gap in the second-moment bound instead of replacing it by a uniform constant. For $S_T = \sum_{t=1}^T \mathbb{E}[\Delta_t]$, smoothness and the ABC condition give $\sum_{t=1}^T \mathbb{E}[\|g_t\|^2] \leq QS_T + CT$. A pathwise regret inequality removes the correlated denominator before taking expectations, yielding $S_T \leq \Gamma\sqrt{b_0^2 + QS_T + CT}$ for AdaGrad-Norm. For AdaGrad, coordinate-wise regret and Cauchy–Schwarz give the same inequality with (Γ, b_0) replaced by (Λ, β_0); see (45) and (49). Solving these quadratic inequalities produces the noise-adaptive bounds. Appendix E supplies separate proofs for the two algorithms. □

Remark 3.10. Noise adaptation here is conditional on stability, although the algorithms need not know A, B, C, L or the stability radii. For a common coordinate initialization, $\beta_0 = \sqrt{d}\, b_0$. The case $C = 0$ permits state-dependent stochastic noise; it does not require a deterministic oracle. The root-form refinements and the two separate algorithm proofs are in Appendix E. The theorem assumes stability and therefore does not contradict the lower bounds in Sections 3.3 and 3.4, where a common trajectory radius is absent.

4 Conclusion

Our first result rules out any prescribed uniform vanishing expected last-iterate rate for AdaGrad-Norm and AdaGrad under conditional sub-Gaussian oracles, even within an arbitrarily small trajectory radius. One fixed instance violates the proposed rate along a subsequence while still converging; only the pulse schedule depends on that rate.

Under finite variance, a fixed infinite-horizon instance also defeats polynomial average-iterate rates with polylogarithmic confidence dependence within any prescribed radius. A persistent one-sided displacement survives averaging, but the conclusion remains compatible with fixed-confidence convergence and Markov bounds.

Without a common radius, horizon-dependent families give constant worst-case expected average error under finite variance or a logarithmic lower bound under bounded noise. The latter also applies to fixed inverse-square-root SGD, not every stepsize or output. Under bounded iterates, classical regret analysis gives noise-adaptive ABC upper bounds. Identifying sampling assumptions that restore uniform last-iterate rates for the unchanged updates remains an important direction.

AI USE STATEMENT

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

Reproducibility statement

All assumptions and theoretical claims are stated in the main paper, with complete proofs provided in the appendix.

References

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

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.

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

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Savelii Chezhegov, Yaroslav Klyukin, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves Adam-norm and AdaGrad-norm when the noise is heavy-tailed. In *International Conference on Machine Learning*, pp. 10269–10333. PMLR, 2025.

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

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

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 *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

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

Guillaume Garrigos, Daniel Cortild, Lucas Ketels, and Juan Peypouquet. Last-iterate complexity of SGD for convex and smooth stochastic problems. *arXiv preprint arXiv:2507.14122*, 2025.

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

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958v1*, 2021.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

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

Yusu Hong and Junhong Lin. Revisiting convergence of AdaGrad with relaxed assumptions. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pp. 1727–1750, 2024b.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of AdaGrad over SGD: Upper and lower bounds in stochastic non-convex optimization. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 3124–3158. PMLR, 2025.

Ruinan Jin and Xiaoyu Wang. Long-run dynamics of AdaGrad-Norm: Trajectory stability and asymptotic stationarity. *arXiv preprint arXiv:2601.01853v2*, 2026. Version 2, August 30, 2026.

Ruinan Jin, Yu Xing, and Xingkang He. On the convergence of mSGD and AdaGrad for stochastic optimization. *arXiv preprint arXiv:2201.11204*, 2022.

Ali Kavis, Kfir Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *10th International Conference on Learning Representations (ICLR)*, 2022.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Kfir Y. Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. In *Workshop on Beyond First Order Methods in ML Systems at ICML'20*, 2020.

Xiaoyu Li, Mingrui Liu, and Francesco Orabona. On the last iterate convergence of momentum methods. In *International Conference on Algorithmic Learning Theory*, pp. 699–717. PMLR, 2022.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Conference on Learning Theory*, pp. 2963–2983. PMLR, 2022.

Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Forty-first International Conference on Machine Learning*, 2024.

Yuxing Liu, Rui Pan, and Tong Zhang. AdaGrad under anisotropic smoothness. In *13th International Conference on Learning Representations, ICLR 2025*. International Conference on Learning Representations, ICLR, 2025.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. *arXiv preprint arXiv:2312.08531*, 2023.

Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. On the convergence of AdaGrad (norm) on $\mathbb{R}^d$: Beyond convexity, non-asymptotic rate and acceleration. In *International Conference on Learning Representations*. International Conference on Learning Representations, 2023a.

Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International conference on machine learning*, pp. 21884–21914. PMLR, 2023b.

H. Brendan McMahan and Matthew J. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, pp. 244–256, 2010.

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

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

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

Francesco Orabona. Almost sure convergence of SGD on smooth non-convex functions. Blog post, 2020a. URL https://parameterfree.com/2020/10/05/almost-sure-convergence-of-sgd-on-smooth-non-convex-functions/.

Francesco Orabona. Last iterate of SGD converges (even in unbounded domains). Blog post, 2020b. URL https://parameterfree.com/2020/08/07/last-iterate-of-sgd-converges-even-in-unbounded-domains/.

Margarita Preobrazhenskaia, Makar Sidorov, Igor Preobrazhenskii, and Eduard Gorbunov. Last iterate convergence of AdaGrad-norm for convex non-smooth optimization. *arXiv preprint arXiv:2604.10728*, 2026.

Maxim Raginsky and Alexander Rakhlin. Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10): 7036–7056, 2011.

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

El Mehdi Saad, Wei-Cheng Lee, and Francesco Orabona. New lower bounds for stochastic non-convex optimization through divergence decomposition. *arXiv preprint arXiv:2502.14060*, 2025.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013.

Sharan Vaswani, Issam Laradji, Frederik Kunstner, Si Yi Meng, Mark Schmidt, and Simon Lacoste-Julien. Adaptive gradient methods converge faster with over-parameterization (but you should do a line-search). *arXiv preprint arXiv:2006.06835*, 2020.

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

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

Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned SGD and the power of adaptive methods. *Advances in Neural Information Processing Systems*, 36:74257–74288, 2023.

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

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116, 2004.

APPENDIX

The proofs follow the original order of the main results. Appendix A first proves the absence of a universal last-iterate rate on a fixed instance. Appendix B treats the uniformly averaged output at shrinking confidence levels, still on a fixed infinite-horizon bounded instance. Appendices C and D then construct horizon-dependent expected average-error lower bounds, and the final part of Appendix D proves the SGD consequence. Appendix E provides separate proofs of the two noise-adaptive upper bounds. Each lower-bound construction for AdaGrad is scalar, so the two algorithms have the same trajectory.

A Proof of Theorem 3.1

We begin with the last-iterate limitation, for which the entire oracle must remain fixed while the evaluation times tend to infinity. The construction below separates the timing of stochastic variation from its magnitude, so a bounded pulse can reveal slow convergence at an arbitrarily late time.

Proof. The construction starts at the minimizer and keeps the accumulator unchanged until the first nonzero pulse. Choose the objective curvature and pulse size so that a pulse respects the stability radius and every later noise-free update contracts. Specifically, set

$$\ell = \min\{L, b_0/(2\eta)\}, \qquad d_0 = \min\{D/2, \eta/2\}, \qquad M = \frac{b_0 d_0}{\sqrt{\eta^2 - d_0^2}}, \qquad f(x) = \frac{\ell}{2}x^2. \tag{14}$$

Initialize at $x_1 = x^\star = 0$, and let

$$q = \min\left\{\frac{1}{4}, \frac{e-1}{\exp(M^2/\sigma^2) - 1}\right\}, \qquad q_n = \frac{q}{(n+1)^2}.$$

Given a deterministic strictly increasing sequence $(\tau_n)_{n\geq 1}$ of positive integers, allow noise at time $t = \tau_n$ only if no earlier nonzero pulse has occurred. At such an active time, draw $\xi_t = M, -M, 0$ with probabilities $q_n/2, q_n/2, 1 - q_n$, using fresh randomness. At all other times, including every time after the first nonzero pulse, set $\xi_t = 0$. Define $g_t = \ell x_t + \xi_t$.

The noise is conditionally symmetric and hence unbiased. At active times,

$$\mathbb{E}[e^{\xi_t^2/\sigma^2} \mid \mathcal{F}_{t-1}] = 1 + q_n(e^{M^2/\sigma^2} - 1) \leq e; \tag{15}$$

at inactive times the conditional expectation is 1. Thus one fixed infinite-horizon oracle satisfies the required moment bound for any choice of the schedule.

With the oracle specified, we next verify stability and quantify the displacement. Before the first pulse, $x_t = 0$ and $b_{t-1} = b_0$, regardless of how much time has elapsed. If the first pulse occurs at τ_n, the next iterate satisfies

$$b_{\tau_n} = \sqrt{b_0^2 + M^2}, \qquad |x_{\tau_n+1}| = \frac{\eta M}{\sqrt{b_0^2 + M^2}} = d_0.$$

Afterward the noise is zero and

$$x_{t+1} = (1 - \eta\ell/b_t)x_t, \qquad 0 < \eta\ell/b_t \leq 1/2. \tag{16}$$

Consequently $|x_t| \leq d_0 \leq D$ on every path.

Let A_n be the event that the first nonzero pulse occurs at τ_n. Since $\sum_{j\geq 1} q_j < q \leq 1/4$,

$$\mathbb{P}(A_n) = q_n \prod_{j<n}(1 - q_j) \geq \frac{3}{4}q_n.$$

With $T_n = \tau_n + 1$, the fixed displacement on A_n therefore yields

$$\mathbb{E}[\Delta_{T_n}] \geq \frac{3\ell d_0^2 q}{8(n+1)^2}. \tag{17}$$

For the explicit schedule $\tau_n = \lceil \exp(\exp n) \rceil$, we have $\log \log T_n \geq n$. Hence (3) follows with $c = 3\ell d_0^2 q/8$.

For an arbitrary prescribed sequence $r_T > 0$ tending to zero, choose the times recursively so that $r_{\tau_n+1} \leq (n+1)^{-4}$. Such increasing times exist because $r_T \to 0$. Equation (17) then gives

$$\frac{\mathbb{E}[\Delta_{T_n}]}{r_{T_n}} \geq \frac{3\ell d_0^2 q}{8}(n+1)^2 \longrightarrow \infty.$$

Once this schedule has been chosen, both the objective and the entire oracle are fixed. Only the evaluation times change along the subsequence.

To complete the argument, we verify that the lower bound concerns speed, not failure of convergence. Indeed, every path has at most one nonzero pulse. After it occurs, $b_t^2 \leq b_0^2 + M^2 + \ell^2 d_0^2 t$, so $\sum_t \eta\ell/b_t = \infty$. Equation (16) and $1 - u \leq e^{-u}$ imply $x_t \to 0$ on that path. A path with no pulse remains at zero. Since $0 \leq \Delta_t \leq \ell d_0^2/2$, bounded convergence also gives $\mathbb{E}[\Delta_t] \to 0$. This verifies convergence only for the constructed instance; no general fixed-instance convergence theorem is used or asserted.

□

B Proof of Theorem 3.2

The preceding proof exploits an accumulator that remains small; we now study the opposite mechanism for the averaged output. A finite cascade increases normalization without leaving a prescribed radius, and the resulting displacement persists long enough to survive uniform averaging.

Proof. A single sufficiently large gradient can cause a displacement close to η, so it need not respect a prescribed radius $D < \eta$. We instead prescribe alternating small displacements and fix their law independently of the eventual evaluation horizons. Let

$$d_0 = \min\{D/4, \eta/4\}, \qquad \ell = \min\{L, b_0/(2\eta)\}, \qquad f(x) = \frac{\ell}{2}x^2, \qquad x_1 = x^\star = 0. \tag{18}$$

To keep each designated displacement equal to d_0, define

$$r = \frac{d_0}{\sqrt{\eta^2 - d_0^2}}, \qquad R = \sqrt{1 + r^2} > 1, \qquad M_j = r b_0 R^{j-1}, \qquad p_j = \min\{1/4, \sigma^2/(4M_j^2)\}.$$

If the first j gradients follow the designated signs $g_j = (-1)^{j+1} M_j$, then

$$b_j = b_0 R^j, \qquad \frac{\eta M_j}{b_j} = d_0, \qquad x_{j+1} = \begin{cases} -d_0, & j \text{ odd}, \\ 0, & j \text{ even}. \end{cases} \tag{19}$$

These identities follow by induction from $M_j = r b_{j-1}$ on that path and $\eta r/R = d_0$.

We now embed this path in an unbiased oracle. Draw an internal finite odd integer N with $\mathbb{P}(N = 2k+1) = 2^{-(k+1)}$ for $k \geq 0$, independently of the fresh random choices used below. The oracle is active at time j only if $j \leq N$ and every preceding gradient followed its designated sign. Write $h_j = \ell x_j$. On an active branch, use

$$g_j = \begin{cases} +M_j, & \text{with probability } \frac{p_j}{2}(1 + h_j/M_j), \\ -M_j, & \text{with probability } \frac{p_j}{2}(1 - h_j/M_j), \\ h_j, & \text{with probability } 1 - p_j. \end{cases} \tag{20}$$

After an outcome different from the designated sign, or after time N, the oracle permanently returns $g_j = h_j$. Thus one law specifies the oracle for the entire infinite trajectory; no later choice of T_n changes it.

The designated path explains the mechanism, but the theorem requires all outcomes to satisfy the oracle and stability assumptions. We therefore check the active branch and every possible way of leaving it.

On an active branch, (19) gives $|x_j| \le d_0$, and

$$|h_j| \le \ell d_0 \le \frac{b_0 d_0}{2\eta} \le \frac{M_1}{2} \le \frac{M_j}{2}.$$

The probabilities in (20) are consequently valid. Let $\mathcal{G}_{j-1} = \sigma(N, g_1, \dots, g_{j-1})$, which also reveals whether the oracle is active. On an active branch, direct calculation gives

$$\mathbb{E}[g_j \mid \mathcal{G}_{j-1}] = h_j, \qquad \mathbb{E}[(g_j - h_j)^2 \mid \mathcal{G}_{j-1}] = p_j(M_j^2 - h_j^2) \le \frac{\sigma^2}{4}. \tag{21}$$

On an inactive branch, the mean remains h_j and the variance is zero. Conditioning these statements on $\mathcal{F}_{j-1} \subseteq \mathcal{G}_{j-1}$ proves Assumptions 2 and 4 for the natural filtration. The internal draw of N need not be observable to the algorithm.

For stability, the only undesignated sign sends 0 to d_0 or $-d_0$ to $-2d_0$. A deterministic-gradient outcome instead contracts toward zero, since $0 \le \eta\ell/b_j \le 1/2$. After deactivation every update is such a contraction. Thus

$$|x_j - x^\star| \le 2d_0 \le D \qquad \text{for every } j \ge 1 \text{ on every path.} \tag{22}$$

No relation such as $D \ge \eta$ is required.

Having verified the assumptions globally, we can focus on rare paths on which the cascade lasts long enough to make later updates very small. For odd n, let E_n be the event that $N = n$ and the first n gradients all follow the designated signs. Each such sign has conditional probability at least $p_j/4$. Set

$$p_\star = \min\{1/4, \sigma^2/(4r^2 b_0^2)\} > 0, \qquad \pi_n = 2^{-(n+1)/2}(p_\star/4)^n R^{-n(n-1)}.$$

Since $p_j \ge p_\star R^{-2(j-1)}$,

$$\mathbb{P}(E_n) \ge 2^{-(n+1)/2} \prod_{j=1}^{n} (p_j/4) \ge \pi_n. \tag{23}$$

On E_n, $x_{n+1} = -d_0$, $b_n = b_0 R^n$, and all subsequent noise vanishes. For sufficiently large odd n, choose

$$T_n = \left\lfloor \sqrt{b_0 R^n} \right\rfloor. \tag{24}$$

Then $T_n \ge 2(n+1)$ and $\eta\ell T_n/(b_0 R^n) \le 1/2$. Monotonicity of the accumulator and Bernoulli's inequality imply, for $n+1 \le t \le T_n$,

$$|x_t| \ge d_0 \left(1 - \frac{\eta\ell}{b_0 R^n}\right)^{t-n-1} \ge d_0 \left(1 - \frac{\eta\ell T_n}{b_0 R^n}\right) \ge \frac{d_0}{2}.$$

All earlier iterates on this event are also nonpositive. Consequently,

$$|\bar{x}_{T_n}| \ge \frac{T_n - n}{T_n} \frac{d_0}{2} \ge \frac{d_0}{4}, \qquad f(\bar{x}_{T_n}) - f^\star \ge \frac{\ell d_0^2}{32} =: c_\star. \tag{25}$$

Taking $\delta_n = \pi_n/2$ proves the probability statement in (4). To compare confidence and horizon, observe that

$$\log(1/\delta_n) = \log 2 + \frac{n+1}{2}\log 2 + n\log(4/p_\star) + n(n-1)\log R = O(n^2),$$
$$\log T_n = \frac{n}{2}\log R + O(1).$$

This gives the second statement in (4) with a fixed parameter-dependent K. The sequence over sufficiently large odd n is eventually strictly increasing; relabeling it yields the sequence asserted in the theorem.

Finally, for every fixed polynomial P and $a, b > 0$, $T_n^{-a} + P(\log(1/\delta_n))T_n^{-b} \to 0$. Thus the right-hand side of (5) is eventually smaller than $c_\star$, while its failure probability is at least $2\delta_n > \delta_n$. This proves the claimed impossibility on one fixed bounded instance.

For clarity, the same rate formula and constants are required for every pair (T, δ), but the theorem does not require a single probability event holding simultaneously over all times. This quantitative requirement should not be confused with qualitative convergence at a fixed confidence level. In this construction the internal odd integer N is finite almost surely, and noise is permanently switched off no later than time $N+1$. Thereafter $x_{t+1} = (1 - \eta\ell/b_t)x_t$, with $0 < \eta\ell/b_t \le 1/2$. On each path, boundedness gives $b_t^2 \le K(\omega) + 4\ell^2 d_0^2 t$ for a finite path-dependent constant $K(\omega)$, so $\sum_t \eta\ell/b_t = \infty$. The contraction product tends to zero, hence $x_t \to 0$ and $\bar{x}_T \to 0$ almost surely. Bounded convergence also gives $\mathbb{E}[f(\bar{x}_T) - f^\star] \to 0$. What fails is precisely the rate/confidence combination in (5); the rare event probabilities are not bounded away from zero as $n \to \infty$.

□

C Proof of Theorem 3.3

Appendix B retained a uniform trajectory radius and isolated the dependence on confidence. Removing that radius allows a different obstruction: a horizon-dependent family with constant expected average error, even though its additive noise is independent and identically distributed on each chosen instance.

Proof. We use rare corrections to make a usually misleading gradient unbiased, while keeping the probability of no correction bounded below over the chosen horizon. Fix a sufficiently large integer $T \ge 2$ and set

$$q = \frac{1}{T}, \qquad a = \frac{\sigma\sqrt{q}}{4}, \qquad B_0 = \sqrt{b_0^2 + \sigma^2/16}. \tag{26}$$

Take T large enough that $a \le 1$, and define

$$f_T(x) = \begin{cases} 0, & x \le 0, \\ x^2/2, & 0 < x < a, \\ ax - a^2/2, & x \ge a. \end{cases} \tag{27}$$

Its derivative is continuous, nondecreasing, and 1-Lipschitz. Thus f_T is convex and 1-smooth, with $f_T^\star = 0$. Choose $x^\star = 0$ and initialize at $x_1 = 1$.

For this fixed instance, draw independent identically distributed noise at every time with law

$$\xi_t = \begin{cases} -2a, & \text{with probability } 1-q, \\ 2a(1-q)/q, & \text{with probability } q. \end{cases} \tag{28}$$

The probabilities and values do not vary with time. Direct calculation gives

$$\mathbb{E}[\xi_t] = 0, \qquad \mathbb{E}[\xi_t^2] = \frac{4a^2(1-q)}{q} \le \frac{\sigma^2}{4}.$$

Therefore $g_t = f_T'(x_t) + \xi_t$ satisfies the conditional unbiasedness and variance assumptions for all $t \ge 1$. The objective and noise law depend on the chosen horizon, but the variance bound is common to the entire family.

The noise law is unbiased only after averaging over its rare correction. To obtain a lower bound, we isolate the complementary event on which that correction never occurs. Let $\mathcal{E}_T$ be the event that the first T noise values are all $-2a$. Then $\mathbb{P}(\mathcal{E}_T) = (1 - 1/T)^T \ge 1/4$.

On $\mathcal{E}_T$, induction shows that $x_t \ge 1 \ge a$ and $g_t = -a$ through time T. The updates thus remain in the linear region and move to the right. From (26), for $t \le T$,

$$b_t^2 = b_0^2 + ta^2 \le B_0^2, \qquad x_t \ge 1 + \frac{\eta a}{B_0}(t-1), \qquad \bar{x}_T \ge 1 + \frac{\eta a(T-1)}{2B_0}. \tag{29}$$

Because $a - a^2/2 \ge 0$, evaluating (27) on this event gives

$$f_T(\bar{x}_T) \ge \frac{\eta a^2(T-1)}{2B_0} = \frac{\eta\sigma^2(T-1)}{32B_0T} \ge \frac{\eta\sigma^2}{64B_0}. \tag{30}$$

The objective is nonnegative on the complement of $\mathcal{E}_T$. Multiplying the last bound by $\mathbb{P}(\mathcal{E}_T) \geq 1/4$ proves (7). This argument fixes an instance for each T; it does not produce nonconvergence of one instance as $T \to \infty$.

□

D Proof of Theorem 3.4

The previous lower bound used rare large corrections under a variance constraint; we now ask what remains possible when the noise is uniformly bounded. A constant-magnitude scalar oracle exposes harmonic accumulation of fluctuations, yielding the logarithmic average-error lower bound for both algorithms.

Proof. Fix a sufficiently large horizon T. We choose a shallow linear region so that the restoring drift is smaller than the accumulated fluctuations, and then smooth its junction with the minimum region. Set $\gamma = \sigma/2$, choose a universal constant $c_0 > 0$ below, and define

$$a = c_0\gamma\sqrt{\frac{\log T}{T}}, \qquad f_T(x) = \begin{cases} 0, & x \leq 0, \\ Lx^2/2, & 0 < x < a/L, \\ ax - a^2/(2L), & x \geq a/L. \end{cases} \tag{31}$$

The derivative is continuous, nondecreasing, and L-Lipschitz, so f_T is convex and L-smooth. Its minimum is zero, and we initialize at $x_1 = x^\star = 0$.

For large enough T, $a \leq \gamma/2$. Write $h_t = f_T'(x_t) \in [0, a]$ and, using fresh randomness at each query, let

$$\mathbb{P}(g_t = \gamma \mid \mathcal{F}_{t-1}) = \frac{1}{2}(1 + h_t/\gamma), \qquad \mathbb{P}(g_t = -\gamma \mid \mathcal{F}_{t-1}) = \frac{1}{2}(1 - h_t/\gamma). \tag{32}$$

These probabilities are valid and give $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = h_t$. For $\xi_t = g_t - h_t$,

$$|\xi_t| \leq \gamma + a \leq \frac{3\sigma}{4}, \qquad \mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}] = \gamma^2 - h_t^2 \geq \frac{3}{4}\gamma^2. \tag{33}$$

The pathwise noise bound implies Assumption 3. The oracle is defined for every time $t \geq 1$, not just for the first T queries. Moreover, $g_t^2 = \gamma^2$ on every path, so both algorithms have the deterministic accumulator

$$b_t^2 = b_0^2 + t\gamma^2. \tag{34}$$

For a fixed chosen T, the oracle depends on the current point through h_t, rather than on an externally specified time schedule.

With the oracle verified, the deterministic denominator allows us to expand the recursion with nonrandom weights. This is the step that isolates the source of the logarithmic factor. Define

$$\bar{x}_T = Y_T - \mathcal{D}_T, \qquad Y_T = -\sum_{s=1}^{T-1} \alpha_s \xi_s, \qquad \mathcal{D}_T = \sum_{s=1}^{T-1} \alpha_s h_s, \qquad \alpha_s = \frac{\eta(T-s)}{Tb_s}. \tag{35}$$

Here Y_T is a sum of martingale differences and has mean zero. The drift is random, but it admits the deterministic bound

$$0 \leq \mathcal{D}_T \leq \frac{\eta a}{\gamma} \sum_{s=1}^{T-1} s^{-1/2} \leq 2c_0\eta\sqrt{\log T} =: d_T^{\max}. \tag{36}$$

We use this pathwise upper bound when comparing positive parts; the random drift itself is never moved outside an expectation.

To quantify the fluctuations, put $W_T = \gamma^2 \sum_{s=1}^{T-1} \alpha_s^2$. Orthogonality of martingale differences and (33) imply

$$\mathbb{E}[Y_T] = 0, \qquad \frac{3}{4}W_T \leq \mathbb{E}[Y_T^2] \leq W_T. \tag{37}$$

Let $s_0 = \max\{1, \lceil b_0^2/\gamma^2 \rceil\}$. For $s_0 \le s \le \lfloor T/2 \rfloor$, we have $b_s^2 \le 2s\gamma^2$ and $(T-s)/T \ge 1/2$. Consequently, for all sufficiently large T,

$$\frac{\eta^2}{16}\log T \le \frac{\eta^2}{8}\sum_{s=s_0}^{\lfloor T/2 \rfloor}\frac{1}{s} \le W_T \le \eta^2\sum_{s=1}^{T-1}\frac{1}{s} \le \eta^2(1+\log T). \tag{38}$$

The lower estimate holds once the finite initial segment up to s_0 is negligible relative to $\log T$.

The second-moment estimate must now be converted into a positive fluctuation. A variance lower bound alone would allow that fluctuation to be concentrated on very rare events. The bounded martingale increments give a fourth-moment bound that prevents this, and hence a lower bound on the expected positive part.

Set $X_s = -\alpha_s\xi_s$ and $d_s = 3\gamma\alpha_s/2$. Then $\mathbb{E}[X_s \mid \mathcal{F}_{s-1}] = 0$ and $|X_s| \le d_s$, with d_s deterministic. Convexity of the exponential on $[-d_s, d_s]$ gives, for any $u \in \mathbb{R}$,

$$\mathbb{E}[e^{uX_s} \mid \mathcal{F}_{s-1}] \le \cosh(ud_s) \le e^{u^2 d_s^2/2}.$$

Iterated conditioning and Chernoff's inequality therefore yield

$$\mathbb{P}(|Y_T| \ge v) \le 2e^{-v^2/(2V_T)}, \qquad V_T = \sum_{s=1}^{T-1} d_s^2 = \frac{9}{4}W_T.$$

Integrating the tail proves

$$\mathbb{E}[Y_T^4] \le \int_0^\infty 8v^3 e^{-v^2/(2V_T)}\,dv = 16V_T^2 = 81W_T^2. \tag{39}$$

Hölder's inequality and (37) now give

$$\mathbb{E}[|Y_T|] \ge \frac{(\mathbb{E}[Y_T^2])^{3/2}}{(\mathbb{E}[Y_T^4])^{1/2}} \ge \frac{\sqrt{3}}{24}\sqrt{W_T}.$$

Since Y_T is integrable and mean zero, $\mathbb{E}[(Y_T)_+] = \mathbb{E}[|Y_T|]/2$; symmetry of its law is not needed. Combining this identity with (38), we obtain

$$\mathbb{E}[(Y_T)_+] \ge k\eta\sqrt{\log T}, \qquad k = \frac{\sqrt{3}}{192}. \tag{40}$$

We can now convert this fluctuation into objective error. It remains to choose the slope so that neither the restoring drift nor the smoothing interval absorbs the positive fluctuation. Set $c_0 = k/8$ and increase the threshold on T so that $a/L \le (k/4)\eta\sqrt{\log T}$. Using $(y-z)_+ \ge y_+ - z$ for $z \ge 0$, together with (35), (36), and (40), gives

$$\mathbb{E}\left[\left(\bar{x}_T - \frac{a}{L}\right)_+\right] \ge \mathbb{E}[(Y_T)_+] - d_T^{\max} - \frac{a}{L} \ge \frac{k}{2}\eta\sqrt{\log T}. \tag{41}$$

For every $x \in \mathbb{R}$, (31) implies $f_T(x) \ge a(x - a/L)_+$. On $x < a/L$ this follows from nonnegativity, and on $x \ge a/L$ it follows from the linear formula. Hence

$$\mathbb{E}[f_T(\bar{x}_T) - f_T^\star] \ge \frac{ak\eta}{2}\sqrt{\log T} = \frac{k^2}{32}\eta\sigma\frac{\log T}{\sqrt{T}}.$$

This proves (8) with the universal constant $c = k^2/32$. Every lower threshold on T depends only on L, σ, b_0, η. The same scalar function and oracle prove the statement for both algorithms.

□

We now verify (10), including the independent additive-noise version. This consequence is separate from the AdaGrad oracle construction: constant gradient magnitude is needed to make the adaptive denominator deterministic, but is unnecessary when SGD already uses deterministic stepsizes.

Proof of the SGD consequence. Fix $L, \sigma, \lambda, \tau > 0$, and put $\gamma = \sigma/2$, $\widetilde{\eta} = \lambda\gamma$, and $\widetilde{b}_0 = \gamma\sqrt{\tau}$. Then

$$\frac{\widetilde{\eta}}{\sqrt{\widetilde{b}_0^2 + t\gamma^2}} = \frac{\lambda}{\sqrt{t+\tau}}.$$

With the constant-magnitude oracle (32), this identity makes the SGD and scalar AdaGrad trajectories exactly equal. Substituting $\eta = \widetilde{\eta}$ in (8) already proves the conditional-oracle version of (10).

For the stronger independent-noise statement, use the same objective (31) with $a = c_0\gamma\sqrt{\log T/T}$, $c_0 = k/8$, and $k = \sqrt{3}/192$, but draw independent Rademacher noise $\xi_t \in \{-\gamma, \gamma\}$. Let $g_t = f_T'(x_t) + \xi_t$. Conditional unbiasedness holds, and $\mathbb{E}[\exp(\xi_t^2/\sigma^2) \mid \mathcal{F}_{t-1}] = e^{1/4} \le e$. All these noise laws are fixed independently of the evaluation horizon.

Write $h_t = f_T'(x_t) \in [0, a]$ and

$$w_s = \frac{\lambda(T-s)}{T\sqrt{s+\tau}}, \qquad Y_T = -\sum_{s<T} w_s\xi_s, \qquad \mathcal{D}_T = \sum_{s<T} w_s h_s, \qquad W_T = \gamma^2 \sum_{s<T} w_s^2.$$

The SGD recursion gives $\bar{x}_T = Y_T - \mathcal{D}_T$, and

$$0 \le \mathcal{D}_T \le 2c_0\widetilde{\eta}\sqrt{\log T}, \qquad \mathbb{E}[Y_T] = 0, \qquad \mathbb{E}[Y_T^2] = W_T.$$

The harmonic estimates (38) apply with $\eta = \widetilde{\eta}$, $b_0 = \widetilde{b}_0$, since w_s is exactly the previous deterministic weight. Independence also gives the elementary fourth-moment identity

$$\mathbb{E}[Y_T^4] = \gamma^4 \sum_{s<T} w_s^4 + 6\gamma^4 \sum_{r<s<T} w_r^2 w_s^2 = 3W_T^2 - 2\gamma^4 \sum_{s<T} w_s^4 \le 3W_T^2.$$

Thus the weaker estimates $\mathbb{E}[Y_T^2] \ge 3W_T/4$ and $\mathbb{E}[Y_T^4] \le 81W_T^2$ used in (40) remain valid. They yield $\mathbb{E}[(Y_T)_+] \ge k\widetilde{\eta}\sqrt{\log T}$. The same drift and smoothing comparison (41) therefore gives

$$\mathbb{E}[f_T(\bar{x}_T) - f_T^\star] \ge \frac{k^2}{32}\widetilde{\eta}\sigma\frac{\log T}{\sqrt{T}} = \frac{k^2}{64}\lambda\sigma^2\frac{\log T}{\sqrt{T}}.$$

This proves (10) with $c_{\mathrm{SGD}} = k^2/64$. Every sufficiently-large-T threshold depends only on L, σ, λ, τ. Only the hard objective changes with T; the stepsize schedule and the independent noise distribution do not. □

The finite-horizon comparison. For completeness, let SGD use a constant $0 < \alpha \le 1/(2L)$ and put $z_t = x_t - x^\star$. Conditional unbiasedness, convexity, and the smooth self-bound give

$$\mathbb{E}[\|z_{t+1}\|^2 \mid \mathcal{F}_{t-1}] \le \|z_t\|^2 - 2\alpha(1 - L\alpha)\Delta_t + \alpha^2\sigma^2 \le \|z_t\|^2 - \alpha\Delta_t + \alpha^2\sigma^2.$$

All finite-time moments here exist: $\|\nabla f(x_t)\| \le L\|z_t\|$ and the variance bound imply, recursively, a finite second moment for z_t from deterministic x_1. Taking expectations, summing over $t \le T$, dropping the nonnegative terminal squared distance, and using convexity of f proves (11). This comparison is a direct distance-recursion calculation, not a claim that every SGD schedule avoids logarithmic losses.

E Proof of Theorem 3.5

The lower bounds above isolate failures that averaging, tail assumptions, and stability do not individually remove. We now combine expected averaging with bounded iterates and the conditional ABC condition. The proof follows the classical regret and self-bounding approach (Duchi et al., 2011; Levy et al., 2018; Vaswani et al., 2020), while retaining the initialization and noise coefficients. After establishing the shared estimates, we give a separate proof for each algorithm.

E.1 Shared self-bounding and integrability estimates

Both proofs close the same quadratic inequality, but use different pathwise regret bounds. We first record the common second-moment estimate and justify the expectations used later. Smoothness applied at $x - \nabla f(x)/L$ gives

$$f^\star \le f\left(x - \frac{\nabla f(x)}{L}\right) \le f(x) - \frac{\|\nabla f(x)\|^2}{2L}.$$

Therefore

$$\|\nabla f(x)\|^2 \le 2L(f(x) - f^\star). \tag{42}$$

Writing $S_T = \sum_{t=1}^T \mathbb{E}[\Delta_t]$ and using $Q = A + 2LB$, the ABC condition implies

$$\sum_{t=1}^T \mathbb{E}[\|g_t\|^2] \le QS_T + CT. \tag{43}$$

All terms are finite. Indeed, smoothness at $x^\star$, where $\nabla f(x^\star) = 0$, bounds Δ_t by $LD^2/2$ for AdaGrad-Norm and by $(L/2)\sum_i D_i^2$ for AdaGrad. The ABC condition then gives uniform conditional second-moment bounds. In particular, the accumulators have finite first moments at each finite time, and the inner products with bounded iterates are integrable. These observations justify both the martingale cancellation and the applications of Jensen's inequality below.

E.2 AdaGrad-Norm

Scalar normalization allows a single Euclidean distance identity. The important order of operations is to sum this identity pathwise before taking expectations, so the correlation between g_t and b_t never needs to be discarded.

Proof of Theorem 3.5, part 1. Write $z_t = x_t - x^\star$. The actual current-gradient update gives

$$\langle g_t, z_t\rangle = \frac{b_t}{2\eta}(\|z_t\|^2 - \|z_{t+1}\|^2) + \frac{\eta\|g_t\|^2}{2b_t}.$$

Summation by parts and $\|z_t\| \le D$ yield

$$\begin{aligned}\sum_{t=1}^T b_t(\|z_t\|^2 - \|z_{t+1}\|^2) &= b_1\|z_1\|^2 + \sum_{t=2}^T (b_t - b_{t-1})\|z_t\|^2 - b_T\|z_{T+1}\|^2 \\ &\le D^2 b_T.\end{aligned}$$

Also,

$$\frac{\|g_t\|^2}{b_t} = \frac{(b_t - b_{t-1})(b_t + b_{t-1})}{b_t} \le 2(b_t - b_{t-1}).$$

Consequently,

$$\sum_{t=1}^T \langle g_t, z_t\rangle \le \left(\frac{D^2}{2\eta} + \eta\right) b_T - \eta b_0 \le \Gamma b_T. \tag{44}$$

Because z_t is bounded and $\mathcal{F}_{t-1}$-measurable, conditional unbiasedness gives $\mathbb{E}[\langle \xi_t, z_t\rangle] = 0$. Convexity and (44) thus imply

$$S_T \le \sum_{t=1}^T \mathbb{E}[\langle \nabla f(x_t), z_t\rangle] = \sum_{t=1}^T \mathbb{E}[\langle g_t, z_t\rangle] \le \Gamma \mathbb{E}[b_T].$$

Applying Jensen's inequality only now, and then (43), gives

$$S_T \le \Gamma\sqrt{\mathbb{E}[b_T^2]} \le \Gamma\sqrt{b_0^2 + QS_T + CT}. \tag{45}$$

Since $S_T \geq 0$, squaring and solving the quadratic inequality yields

$$S_T \leq \frac{\Gamma^2 Q + \sqrt{\Gamma^4 Q^2 + 4\Gamma^2(b_0^2 + CT)}}{2}.$$

Convexity gives $\mathbb{E}[f(\bar{x}_T) - f^\star] \leq S_T/T$, hence the sharper estimate

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{\Gamma^2 Q + \sqrt{\Gamma^4 Q^2 + 4\Gamma^2(b_0^2 + CT)}}{2T}. \tag{46}$$

Finally,

$$\sqrt{\Gamma^4 Q^2 + 4\Gamma^2(b_0^2 + CT)} \leq \Gamma^2 Q + 2\Gamma\sqrt{b_0^2 + CT} \leq \Gamma^2 Q + 2\Gamma b_0 + 2\Gamma\sqrt{CT},$$

which proves (12) with the stated constants. □

E.3 Coordinate-wise AdaGrad

Each coordinate has its own accumulator, so one first bounds its scalar regret separately. Summing these bounds and applying Cauchy–Schwarz recovers a scalar quadratic inequality without requiring coordinate-wise ABC assumptions.

Proof of Theorem 3.5, part 2. Let $z_{t,i} = x_{t,i} - x_i^\star$. The update in coordinate i gives

$$g_{t,i} z_{t,i} = \frac{b_{t,i}}{2\eta}(z_{t,i}^2 - z_{t+1,i}^2) + \frac{\eta g_{t,i}^2}{2b_{t,i}}.$$

Since $|z_{t,i}| \leq D_i$ and $b_{t,i}$ is nondecreasing,

$$\sum_{t=1}^{T} b_{t,i}(z_{t,i}^2 - z_{t+1,i}^2) \leq D_i^2 b_{T,i}, \qquad \sum_{t=1}^{T} \frac{g_{t,i}^2}{b_{t,i}} \leq 2(b_{T,i} - b_{0,i}).$$

It follows pathwise that

$$\sum_{t=1}^{T} g_{t,i} z_{t,i} \leq \left(\frac{D_i^2}{2\eta} + \eta\right) b_{T,i} - \eta b_{0,i} \leq \Gamma_i b_{T,i}. \tag{47}$$

Summing over i, taking expectations, and using convexity and conditional unbiasedness after this pathwise estimate gives

$$S_T \leq \sum_{i=1}^{d} \Gamma_i \mathbb{E}[b_{T,i}] \leq \sum_{i=1}^{d} \Gamma_i \sqrt{b_{0,i}^2 + \sum_{t=1}^{T} \mathbb{E}[g_{t,i}^2]}. \tag{48}$$

The martingale terms cancel because $z_{t,i}$ is bounded and $\mathcal{F}_{t-1}$-measurable. Cauchy–Schwarz and the global estimate (43) now yield

$$\begin{aligned} S_T &\leq \left(\sum_{i=1}^{d} \Gamma_i^2\right)^{1/2} \left(\sum_{i=1}^{d} b_{0,i}^2 + \sum_{t=1}^{T} \mathbb{E}[\|g_t\|^2]\right)^{1/2} \\ &\leq \Lambda\sqrt{\beta_0^2 + QS_T + CT}. \end{aligned} \tag{49}$$

Squaring and taking the positive root gives

$$S_T \leq \frac{\Lambda^2 Q + \sqrt{\Lambda^4 Q^2 + 4\Lambda^2(\beta_0^2 + CT)}}{2}.$$

Dividing by T and using convexity of f gives the sharper estimate

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \leq \frac{\Lambda^2 Q + \sqrt{\Lambda^4 Q^2 + 4\Lambda^2(\beta_0^2 + CT)}}{2T}. \tag{50}$$

The estimate

$$\sqrt{\Lambda^4 Q^2 + 4\Lambda^2(\beta_0^2 + CT)} \leq \Lambda^2 Q + 2\Lambda\beta_0 + 2\Lambda\sqrt{CT}$$

then gives (13). This argument uses only the global ABC condition. The intermediate bound (48) preserves more detailed coordinate-wise second-moment information when it is available. □

E.4 Noise specializations and the role of stability

The two proofs have the same final dependence on the residual coefficient C, even though their geometric constants differ. When $C = 0$, their simplified bounds are $(\Gamma^2 Q + \Gamma b_0)/T$ and $(\Lambda^2 Q + \Lambda\beta_0)/T$. This case allows nonzero state-dependent noise; it excludes only an additive residual term in the ABC model.

Under conditional unbiasedness and bounded variance,

$$\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_{t-1}] = \|\nabla f(x_t)\|^2 + \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_{t-1}] \le \|\nabla f(x_t)\|^2 + \sigma^2.$$

Thus $(A, B, C) = (0, 1, \sigma^2)$, so $Q = 2L$ and

$$\mathbb{E}[f(\bar{x}_T) - f^\star] \le \frac{2L\Gamma^2 + \Gamma b_0}{T} + \frac{\Gamma\sigma}{\sqrt{T}}$$

for AdaGrad-Norm, with the analogous substitution $(\Gamma, b_0) \mapsto (\Lambda, \beta_0)$ for AdaGrad. These guarantees assume a common almost-sure stability radius. They therefore do not contradict the horizon-dependent lower bounds without such a radius, nor do they imply a last-iterate rate or polynomial high-probability bounds with polylogarithmic confidence dependence. Markov's inequality does give the weaker $1/\delta$ dependence in (6).