THE CUBE-ROOT BARRIER FOR ALMOST-SURE LAST-ITERATE CONVERGENCE OF SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Almost-sure last-iterate convergence is the natural pathwise guarantee for stochastic gradient descent (SGD), which is typically run once with a deterministic stepsize schedule and returns its last iterate. We determine the fastest polynomial order of almost-sure last-iterate convergence that such a schedule can guarantee for smooth convex objectives under the standard conditional ABC second-moment condition, up to a slowly varying logarithmic correction. We establish a sharp cube-root barrier: the optimal polynomial exponent is $1/3$, with a single schedule achieving every smaller exponent, while no deterministic schedule can guarantee a larger one. We further show that the lower bound already holds for one-dimensional smooth convex problems with fixed i.i.d. finite-variance gradient noise, and that a logarithmically corrected schedule attains this optimal cube-root order up to a polylogarithmic factor.

1 INTRODUCTION

Stochastic gradient descent (SGD) is the basic algorithm of large-scale stochastic optimization (Robbins & Monro, 1951). SGD and its variants underlie the training of most machine-learning models (Bottou, 2004). In practice, SGD is run once, with a stepsize schedule fixed in advance, and the model we actually use is the last iterate x_t. A guarantee relevant to practice must therefore control the last iterate along whichever trajectory the run produces. Typical convergence analyses in expectation or with high probability do not offer such a guarantee. A bound in expectation averages over hypothetical reruns of the algorithm. A high-probability bound at a prescribed horizon permits the exceptional event to change with the horizon: even when the failure probability is small at every fixed time, no single realization need satisfy the bound at all late times. An almost-sure last-iterate rate, on the other hand, provides one probability-one event on which every sufficiently late iterate obeys the same error bound.

Such rates are delicate to obtain because a finite conditional second-moment bound still permits rare, large gradient noise. Khaled & Richtárik (2023) proposed a unifying assumption on stochastic gradients for general nonconvex objectives, referred to as the ABC condition. In our smooth convex setting, the ABC condition reduces to a bound of the form $\mathbb{E}_t \|g_t\|^2 \le D\big(f(x_t) - f^*\big) + C$ for finite constants $D, C \ge 0$. Such a bound controls the second moment but not the tails: events that are unlikely at any fixed time can recur infinitely often along one trajectory, and each recurrence can throw the last iterate far from the minimizer. We therefore ask a basic question:

> *What is the fastest polynomial decay that one fixed deterministic schedule can guarantee for the last iterate along almost every trajectory?*

Liu & Yuan (2022, 2024) proved that, for each $\varepsilon \in (0, 1/3)$, the schedule $\alpha_t = \Theta(t^{-(2/3+\varepsilon)})$ yields $f(x_t) - f^* = O(t^{-(1/3-\varepsilon)})$ almost surely. This result points to the somewhat peculiar exponent $1/3$: rates arbitrarily close to $t^{-1/3}$ are attainable, but only by retuning the schedule as the target exponent changes. It is therefore unclear whether $1/3$ reflects a fundamental barrier for the last iterate or merely an artifact of the particular schedules and analysis. This leaves two basic questions open:

1. Does the critical schedule $\alpha_t \asymp t^{-2/3}$ attain every exponent $q < 1/3$ without retuning?
2. More fundamentally, can any deterministic schedule, possibly highly irregular, break the $1/3$ barrier?

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.

Existing minimax bounds in expectation and almost-sure rates for averaged or otherwise modified iterates do not answer these questions: they do not characterize the pathwise rate of the last iterate produced by a single deterministic schedule (Agarwal et al., 2012, Sebbouh et al., 2021).

In this paper, we answer both questions sharply. More specifically, we show the following:

1. **A single critical schedule attains every subcritical rate.** The schedule $\alpha_t = (t+1)^{-2/3}$ ensures that, for every finite-dimensional instance satisfying the conditional ABC assumptions and every $q < 1/3$,
$$(t+1)^q\big(f(x_t) - f^*\big) \longrightarrow 0 \quad \text{almost surely.}$$
2. **No deterministic schedule can break the $1/3$ barrier.** For every positive deterministic schedule and every $r > 1/3$, there exists a one-dimensional smooth convex instance with a fixed, time-homogeneous, i.i.d. finite-variance oracle such that
$$\limsup_{t\to\infty}(t+1)^r\big(f(x_t) - f^*\big) = +\infty \quad \text{almost surely.}$$

Hence $1/3$ is the sharp supremal polynomial exponent for almost-sure last-iterate convergence: the seemingly unusual cube-root rate is a fundamental barrier, not an artifact of a particular stepsize family or proof technique. The exact polynomial rate appears to be more delicate. We do not establish a universal $O(t^{-1/3})$ guarantee. The schedule in Proposition 2.4 attains this rate up to a slowly varying logarithmic factor, while Section 4 shows that, under the broader conditional noise model, even the critical schedule can fail to achieve an exact $O(t^{-1/3})$ rate.

To illustrate why this cube-root barrier occurs, we identify two mechanisms that determine the threshold for a power-law schedule of the form $\alpha_t \asymp t^{-p}$. First, recurrent finite-variance shocks limit the exponent to $q \le 2p - 1$: a shock at time t is scaled by α_t, and sufficiently large shocks recur infinitely often. Second, slow progress on arbitrarily flat objectives limits the exponent to $q \le 1 - p$, since the cumulative step budget satisfies $\sum_{s<t} \alpha_s \asymp t^{1-p}$. The two bounds meet at $(p, q) = (2/3, 1/3)$, as shown in Figure 1.

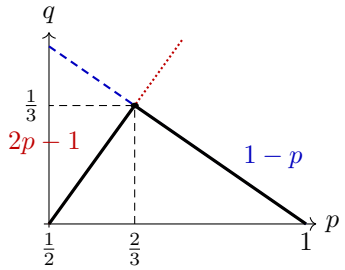

Figure 1: Tradeoff between stepsize decay and convergence rate. For $\alpha_t \asymp t^{-p}$, noise limits the achievable rate to $q \le 2p - 1$, while slow progress limits it to $q \le 1 - p$. The solid line shows the resulting bound $\min\{1 - p, 2p - 1\}$; the dotted and dashed lines show the noise and progress bounds, respectively. The best balance occurs at $(p, q) = (2/3, 1/3)$.

The analysis for power-law schedules leaves open whether an irregular schedule can evade both mechanisms. To rule this out, we count the iterations vulnerable to shocks below a given energy. If this count is sufficiently large, we construct a time-homogeneous i.i.d. finite-variance oracle whose damaging shocks recur infinitely often. Otherwise, the same counting argument bounds the cumulative movement of the iterate, and a sufficiently flat deterministic objective forces slow progress. These alternatives exhaust all positive deterministic schedules and show that no irregular schedule can evade both mechanisms. Our analysis thus establishes that the cube-root barrier is a property of almost-sure last-iterate convergence itself, not an artifact of a particular stepsize family or proof technique. This conclusion is somewhat surprising: to the best of our knowledge, no prior work has identified $1/3$ as a fundamental limit for almost-sure last-iterate convergence.

Related work and positioning. Classical stochastic approximation established almost-sure convergence through the Robbins–Monro recursion and almost-supermartingale arguments (Robbins &

Monro, 1951, Robbins & Siegmund, 1971). Qualitative pathwise convergence for gradient methods and nonconvex SGD was developed further by Bertsekas & Tsitsiklis (2000), Mertikopoulos et al. (2020), while rate results for stochastic approximation and averaged SGD under local strong convexity were obtained in Pelletier (1998), Godichon-Baggioni (2019). Subsequent extensions treat SGD and stochastic heavy-ball methods, composite or proximal updates, biased or growing noise, and gradient domination (Sebbouh et al., 2021, Liang et al., 2024, Liang & Xu, 2024, Karandikar & Vidyasagar, 2024, Weissmann et al., 2025). The closest upper-bound results rely on more structured models. Lei et al. (2018) obtain the envelope $\min\{1-p, 2p-1\}$ for online gradient descent in a reproducing kernel Hilbert space under a structured sample-loss model, while Hudiani (2025) derives related exponents for SGD and stochastic heavy ball in an i.i.d. parametric model, including a stopped-trajectory result in the convex case. Liu & Yuan (2022, 2024) systematically study almost-sure convergence rates of stochastic gradient methods and, in particular, analyze almost-sure last-iterate convergence under the smooth ABC framework used here, but leave the sharp rate for general convex objectives unresolved. In particular, these results leave open whether the $1/3$ exponent is intrinsic or can be improved by a different deterministic stepsize schedule. Our lower bound resolves this question by ruling out *every* exponent above $1/3$ for *every* positive deterministic schedule.

A related line of work studies last-iterate convergence at a prescribed terminal horizon. Standard schedules can lose logarithmic factors for nonsmooth convex problems (Shamir & Zhang, 2013, Harvey et al., 2019), while the horizon-dependent construction of Jain et al. (2019) recovers the information-theoretic rate. More recent work gives unified fixed-horizon guarantees and sharper smooth stochastic bounds under different assumptions (Liu & Zhou, 2024, Garrigos et al., 2025). These results concern expectation or fixed confidence, may tune the schedule using the terminal time, and often impose samplewise structure absent from conditional ABC. They therefore do not provide an almost-sure rate guarantee for a single deterministic schedule run indefinitely.

Work on deterministic schedules that do not depend on a terminal horizon is also closely related. Kornowski & Shamir (2026) prove that such schedules cannot remove the last-iterate logarithmic loss for nonsmooth convex gradient descent. For smooth convex gradient descent, Zhang et al. (2025), Tsai et al. (2026) study acceleration using stepsize schedules fixed in advance and its limitations, while Ma & Chen (2026) gives a complementary lower bound for predetermined schedules at a fixed horizon. These results concern deterministic worst-case optimization without gradient noise. To our knowledge, no prior result covers every positive deterministic schedule and shows, for each $r > 1/3$, that a one-dimensional smooth convex problem with a fixed time-homogeneous i.i.d. finite-variance oracle can violate the rate t^{-r} almost surely. Establishing this lower bound is the main technical contribution of the paper.

2 Problem Setup and Main Results

A schedule is a positive deterministic sequence $\alpha = (\alpha_t)_{t\geq 0}$ fixed before the objective and stochastic gradient oracle are chosen. For a fixed schedule α, an admissible SGD instance under α consists of a finite dimension d, a complete filtered probability space $(\Omega, \mathcal{A}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$, a differentiable convex objective $f : \mathbb{R}^d \to \mathbb{R}$, an $\mathcal{F}_0$-measurable initialization x_0, and stochastic gradients $(g_t)_{t\geq 0}$ generating the iterates according to

$$x_{t+1} = x_t - \alpha_t g_t. \tag{1}$$

We require x_t to be $\mathcal{F}_t$-measurable and g_t to be $\mathcal{F}_{t+1}$-measurable. The objective has a nonempty minimizer set $S := \operatorname{argmin} f$ and an L-Lipschitz gradient for some finite $L > 0$. We write $f^* := \min f$ and $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. For some finite $A, B, C \geq 0$, the stochastic gradients satisfy

$$\mathbb{E}_t g_t = \nabla f(x_t), \tag{2}$$

$$\mathbb{E}_t \|g_t\|^2 \leq A\big(f(x_t) - f^*\big) + B\|\nabla f(x_t)\|^2 + C. \tag{3}$$

We assume $\mathbb{E}\|x_0\|^2 < \infty$. The finite-time square-integrability lemma (Lemma A.3 in Appendix A) then gives square-integrability of x_t and g_t at every finite time. Let $\mathfrak{C}_{\mathrm{ABC}}(\alpha)$ denote the collection of all admissible SGD instances under the schedule α. Across instances, the dimension, filtered probability space, objective, initialization, oracle, and finite constants L, A, B, C may vary.

Asymptotic notation. All asymptotic statements are understood as $t \to \infty$. For positive deterministic sequences (a_t) and (b_t), we write $a_t \asymp b_t$ if there exist constants $0 < c \leq C < \infty$ and

$T < \infty$ such that $cb_t \le a_t \le Cb_t$ for all $t \ge T$. For a random sequence (X_t), $X_t = o(a_t)$ almost surely means that $X_t/a_t \to 0$ almost surely. We write $X_t = O(a_t)$ almost surely if there exist almost surely finite random variables K and T such that $|X_t| \le Ka_t$ for all integers $t \ge T$. Unless stated otherwise, deterministic implied constants may depend on the fixed instance and on fixed parameters such as p, q, r, but not on t.

Define

$$\rho_{\mathrm{a.s.}} := \sup\Big\{ q > 0 : \exists \text{ one positive deterministic schedule } \alpha \text{ such that, for every } \mathcal{I} \in \mathfrak{C}_{\mathrm{ABC}}(\alpha), \\ (t+1)^q \big(f_{\mathcal{I}}(x_{\mathcal{I},t}) - f^*_{\mathcal{I}}\big) \longrightarrow 0 \quad \mathbb{P}_{\mathcal{I}}\text{-almost surely} \Big\}. \tag{4}$$

Here the schedule is fixed before the SGD instance is chosen. The probability-one event and the constants and starting indices in the asymptotic statements may depend on the instance. Once α and $\mathcal{I}$ are fixed, we suppress these indices.

Theorem 2.1 (Achievable rates for polynomial schedules)**.** *Fix a positive deterministic schedule (α_t) for which there exist $p \in (1/2, 1)$, $0 < a_- \le a_+ < \infty$, and $t_0 \in \mathbb{N}$ such that*

$$a_-(t+1)^{-p} \le \alpha_t \le a_+(t+1)^{-p} \qquad (t \ge t_0). \tag{5}$$

Then, for every $\mathcal{I} \in \mathfrak{C}_{\mathrm{ABC}}(\alpha)$ and every

$$0 < q \le 1 - p, \qquad q < 2p - 1, \tag{6}$$

we have

$$(t+1)^q \big(f(x_t) - f^*\big) \longrightarrow 0 \quad \textit{almost surely}. \tag{7}$$

Moreover, x_t converges almost surely to an S-valued random vector.

The two restrictions on q are balanced at $p = 2/3$. Hence the single schedule

$$\alpha^*_t := (t+1)^{-2/3} \tag{8}$$

satisfies (7) for every $q < 1/3$ and every $\mathcal{I} \in \mathfrak{C}_{\mathrm{ABC}}(\alpha^*)$. Importantly, the schedule does not depend on the target exponent q: the same choice attains all polynomial rates strictly below $t^{-1/3}$. This strengthens the earlier result of Liu & Yuan (2022, 2024), where approaching the $1/3$ exponent requires retuning the stepsize decay exponent.

Theorem 2.2 (Cube-root barrier for arbitrary deterministic schedules)**.** *Let $(\alpha_t)_{t \ge 0}$ be any deterministic sequence of positive real numbers and let $r > 1/3$. There exist*

- *a one-dimensional convex C^1 function f with $f(0) = \min f = 0$ and a 1-Lipschitz derivative,*
- *a deterministic initialization $x_0 = 1/2$, and*
- *a time-homogeneous i.i.d. unbiased oracle satisfying $\mathbb{E}\,|g(x, Z)|^2 \le |f'(x)|^2 + 1$ for every $x \in \mathbb{R}$,*

such that $x_{t+1} = x_t - \alpha_t g(x_t, Z_t)$ obeys

$$\limsup_{t \to \infty} (t+1)^r f(x_t) = +\infty \quad \textit{almost surely}. \tag{9}$$

The objective and oracle law may depend on the schedule and on r. Once selected, the objective is fixed and the oracle law is time-homogeneous. Every construction lies in the fixed normalized class $L = 1, A = 0, B = 1, C = 1$.

Theorem 2.2 shows that the lower-bound construction can be remarkably simple: the problem is one-dimensional, the objective is convex with a 1-Lipschitz derivative, the initialization is deterministic, and the oracle is a single fixed i.i.d. law rather than an adversarial or time-varying noise process. Thus the failure of rates beyond the $1/3$ threshold is not caused by high dimension, nonsmoothness, or noise adapted to the trajectory; it is intrinsic to almost-sure last-iterate convergence under deterministic schedules fixed in advance. Whether adaptive stepsizes chosen from observations along the trajectory can improve this threshold remains open.

Corollary 2.3 (Sharp exponent and polynomially decaying stepsizes)**.** *The supremal exponent in (4) is*

$$\rho_{\mathrm{a.s.}} = \frac{1}{3}.$$

Moreover, fix $p \in (1/2, 1)$ and a positive deterministic schedule satisfying $\alpha_t \asymp t^{-p}$. The supremum of the exponents $q > 0$ for which (7) holds for every $\mathcal{I} \in \mathfrak{C}_{\mathrm{ABC}}(\alpha)$ is $\min\{1 - p, 2p - 1\}$.

Corollary 2.3 also shows how the guaranteed rate depends on the stepsize decay exponent p. The best choice is $p = 2/3$; moving away from it decreases the achievable exponent according to $\min\{1-p, 2p-1\}$, as illustrated in Figure 1. At $p = 2/3$ and $q = 1/3$, the weighted-noise series $\sum_t (t+1)^q \alpha_t^2$ becomes harmonic and therefore diverges. A slowly varying logarithmic correction to the stepsize makes this series summable while preserving the divergent progress series, which gives the following near-$t^{-1/3}$ guarantee.

Proposition 2.4 (Near-cube-root rate with logarithmic correction)**.** *Fix $\delta, \eta > 0$ and, for all sufficiently large integers t, set*

$$\ell(t)^3 := \log t \, (\log\log t)^{1+\delta}, \qquad \alpha_t := \frac{\eta}{t^{2/3}\ell(t)}, \tag{10}$$

with arbitrary positive values at the preceding times. Then every $\mathcal{I} \in \mathfrak{C}_{\mathrm{ABC}}(\alpha)$ satisfies

$$f(x_t) - f^* = o\left(\frac{[\log t (\log\log t)^{1+\delta}]^{1/3}}{t^{1/3}}\right) \quad \textit{almost surely}. \tag{11}$$

Complete proofs of Theorem 2.1 and Proposition 2.4 appear in Appendix B, and the proof of Theorem 2.2 appears in Appendix E. The following section provides an overview of the main proof ideas.

3 Proof Overview

We now explain why the $1/3$ threshold appears and how the two main results are proved. For the upper bound, the key step is to combine a recursion for the distance to the solution set with a recursion for the objective gap. Their interaction upgrades weighted summability from Robbins–Siegmund to an almost-sure rate for the last iterate. For the lower bound, we show that every deterministic schedule must fail in one of two ways: either sufficiently large noise shocks recur often enough to prevent faster decay, or the cumulative steps are too small to make sufficient progress on a sufficiently flat objective. A counting argument shows that these two mechanisms cover all positive deterministic schedules.

3.1 Upper bound: from weighted summability to last-iterate decay

Fix an admissible SGD instance and a deterministic minimizer $z \in S$. Write

$$F_t := f(x_t) - f^*, \qquad R_t := \|x_t - z\|^2.$$

Smooth convexity gives $\|\nabla f(x_t)\|^2 \le 2LF_t$, so the ABC condition reduces to

$$\mathbb{E}_t \|g_t\|^2 \le DF_t + C, \qquad D := A + 2BL.$$

Using this bound in the SGD update gives two conditional recursions:

$$\mathbb{E}_t R_{t+1} \le R_t - (2\alpha_t - D\alpha_t^2)F_t + C\alpha_t^2, \tag{12}$$

$$\mathbb{E}_t F_{t+1} \le (1 + u\alpha_t^2)F_t + v\alpha_t^2, \tag{13}$$

where $u := LD/2$ and $v := LC/2$.

The first recursion controls cumulative progress. Since $\alpha_t \to 0$, we eventually have $D\alpha_t \le 1$, and hence

$$\mathbb{E}_t R_{t+1} \le R_t - \alpha_t F_t + C\alpha_t^2. \tag{14}$$

For the polynomially decaying stepsizes in Theorem 2.1, $p > 1/2$ implies $\sum_t \alpha_t^2 < \infty$. Therefore the Robbins–Siegmund argument stated in Lemma A.4 in Appendix A applies to (14). It shows that R_t converges and, more importantly,

$$\sum_{t=0}^{\infty} \alpha_t F_t < \infty \quad \text{almost surely}. \tag{15}$$

This weighted summability is the first key ingredient of the upper bound (see also Lemma A.8 in Appendix A).

Weighted summability alone, however, does not imply convergence of the last iterate. For example, if $\alpha_t = 1/t$ and $y_t = \mathbf{1}\{t = 2^k \text{ for some } k\}$, then $\sum_t \alpha_t y_t < \infty$ even though y_t does not converge to zero. Thus (15) by itself does not rule out sparse spikes in F_t. The second recursion, (13), provides the additional control needed to exclude such spikes.

To obtain a target rate, introduce a positive deterministic sequence b_t and aim to prove $b_t F_t \to 0$. Suppose that, eventually,

$$b_{t+1} - b_t \le \kappa_1 \alpha_t, \qquad b_{t+1}\alpha_t \le \kappa_2, \tag{16}$$

and that

$$\sum_t b_{t+1}\alpha_t^2 < \infty, \qquad \sum_t \frac{\alpha_t}{b_t} = \infty. \tag{17}$$

The first two conditions ensure that the growth of b_t is compatible with the stepsizes, while the two series conditions balance the accumulated noise against the progress needed to force the rescaled objective gap to zero.

Set $\kappa := \kappa_1 + u\kappa_2$. Multiplying (13) by b_{t+1} and using (16) gives

$$\mathbb{E}_t[b_{t+1}F_{t+1}] \le b_t F_t + \kappa\alpha_t F_t + v b_{t+1}\alpha_t^2. \tag{18}$$

On the other hand, multiplying (14) by κ gives

$$\mathbb{E}_t[\kappa R_{t+1}] \le \kappa R_t - \kappa\alpha_t F_t + \kappa C\alpha_t^2. \tag{19}$$

The terms involving $\kappa\alpha_t F_t$ cancel when the two inequalities are added. Therefore, for

$$W_t := b_t F_t + \kappa R_t,$$

we obtain

$$\mathbb{E}_t W_{t+1} \le W_t + (v b_{t+1} + \kappa C)\alpha_t^2. \tag{20}$$

The error term is summable by (17), so another Robbins–Siegmund argument shows that W_t converges almost surely. Since R_t also converges, $b_t F_t = W_t - \kappa R_t$ has a finite nonnegative limit, say Z.

It remains to show that this limit is zero. If $Z > 0$, then eventually $F_t \ge Z/(2b_t)$, which together with (17) gives

$$\sum_t \alpha_t F_t \ge \frac{Z}{2}\sum_t \frac{\alpha_t}{b_t} = \infty.$$

This contradicts (15). Hence $Z = 0$ and

$$b_t F_t \longrightarrow 0 \quad \text{almost surely.}$$

The role of the second recursion is therefore to turn the weighted summability of (15) into a genuine last-iterate rate.

For the polynomially decaying stepsizes $\alpha_t \asymp t^{-p}$, take $b_t = (t+1)^q$. If $q \le 1 - p$, then

$$b_{t+1} - b_t = O(t^{q-1}) = O(\alpha_t), \qquad \frac{\alpha_t}{b_t} \asymp t^{-(p+q)},$$

and $\sum_t \alpha_t/b_t$ diverges because $p + q \le 1$. The weighted-noise condition becomes

$$\sum_t b_{t+1}\alpha_t^2 \le K\sum_t (t+1)^{q-2p} < \infty$$

for some finite K, which requires $q < 2p - 1$. Finally, $b_{t+1}\alpha_t = O(t^{q-p})$ is bounded because $q \le 1 - p < p$ when $p > 1/2$. Thus the argument gives exactly the range

$$q \le 1 - p, \qquad q < 2p - 1.$$

The strict inequality in the second condition comes from the noise term: when $q = 2p - 1$, $b_{t+1}\alpha_t^2$ is of order $1/t$, so the corresponding series is harmonic and diverges.

3.2 LOWER BOUND: RECURRENT SHOCKS VERSUS SLOW PROGRESS

Fix a positive deterministic schedule and a target exponent $r \in (1/3, 1)$. The goal is to show that the rate t^{-r} cannot be guaranteed. The proof asks how large a noise shock is needed at time t to make the rescaled objective gap $t^r f(x_t)$ non-negligible.

Consider first the quadratic objective $f(x) = x^2/2$. A noise shock with squared amplitude H is multiplied by α_t^2 in the objective, so its contribution to the rescaled gap is of order $t^r \alpha_t^2 H$. Define

$$\tau_t := \frac{t^{-r}}{\alpha_t^2}, \qquad N(s) := \#\{t \geq 1 : \tau_t \leq s\}. \tag{21}$$

Here τ_t is the squared shock amplitude needed at time t to produce a gap of order t^{-r}, while $N(s)$ counts how many iterations can be disrupted by shocks of squared amplitude at most s. The behaviour of $N(s)$ leads to three exhaustive cases, summarized in Figure 2.

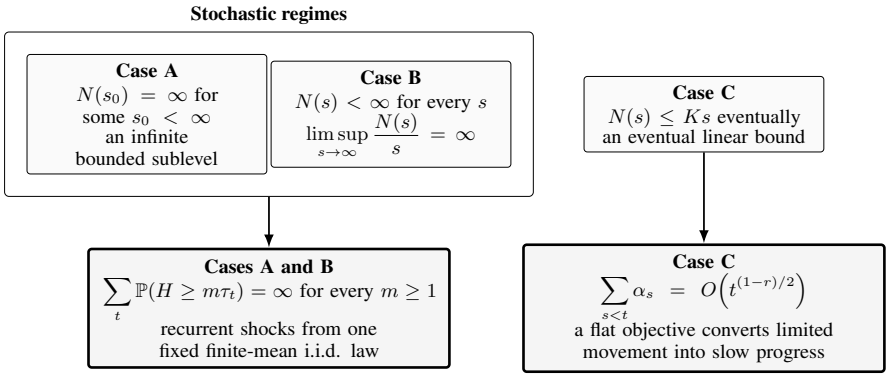

Figure 2: The three cases in the lower-bound argument. In Cases A and B, sufficiently many iterations are vulnerable to noise, allowing the construction of a fixed i.i.d. finite-variance oracle whose damaging shocks occur infinitely often. In Case C, vulnerable iterations are sparse enough to imply a bound on the cumulative steps, and a sufficiently flat objective then forces slow progress.

Case A: infinitely many iterations are vulnerable at a fixed energy level. Suppose $N(s_0) = \infty$ for some finite s_0. Then infinitely many iterations can be disrupted by shocks whose squared amplitude is bounded by s_0. Choose an unbounded nonnegative random variable H with finite mean and rescale it so that $\mathbb{E}H \leq 1$. For every integer $m \geq 1$, $\mathbb{P}(H \geq ms_0) > 0$. Since $\tau_t \leq s_0$ at infinitely many indices,

$$\sum_t \mathbb{P}(H \geq m\tau_t) = \infty \qquad \text{for every } m \geq 1.$$

Thus even a fixed finite-mean distribution produces shocks of arbitrarily large rescaled effect infinitely often.

Case B: the number of vulnerable iterations grows faster than the energy level. Suppose every $N(s)$ is finite, but $N(s)/s$ is unbounded for arbitrarily large s. Choose $v_k \uparrow \infty$ such that

$$N(v_k) \geq k2^k v_k.$$

We construct a single shock-energy distribution, fixed for the entire run, by setting

$$h_k := kv_k, \qquad \mathbb{P}(H = h_k) := \frac{2^{-k}}{kv_k}, \tag{22}$$

and placing the remaining probability mass at zero. This distribution has finite mean, $\mathbb{E}H = \sum_k 2^{-k} = 1$. Moreover, for every integer $m \geq 1$, Tonelli's theorem and the definition of v_k give

$$\sum_t \mathbb{P}(H \geq m\tau_t) = \mathbb{E}N(H/m) \geq \sum_{k\geq m} \frac{2^{-k}}{kv_k} N(v_k) \geq \sum_{k\geq m} 1 = \infty. \tag{23}$$

Hence Cases A and B both yield a fixed finite-mean shock distribution for which shocks exceeding $m\tau_t$ occur often enough, for every m.

To turn these shock energies into an SGD oracle, take independent copies H_t and independent Rademacher signs ε_t, and define

$$g(x, H_t, \varepsilon_t) = x + \varepsilon_t \sqrt{H_t}.$$

For $f(x) = x^2/2$, this oracle is unbiased and satisfies

$$\mathbb{E}\,|g(x, H, \varepsilon)|^2 = x^2 + \mathbb{E}H \le |f'(x)|^2 + 1.$$

Conditional on the past, one of the two signs moves the next iterate away from zero, and that sign has probability $1/2$. Therefore (23), together with the conditional Borel–Cantelli argument in Lemma A.5 in Appendix A, implies that such favourable shocks occur infinitely often almost surely. On each of these events,

$$|x_{t+1}| \ge \alpha_t \sqrt{H_t}, \qquad (t+2)^r f(x_{t+1}) \ge t^r f(x_{t+1}) \ge \frac{t^r \alpha_t^2 H_t}{2} = \frac{H_t}{2\tau_t} \ge \frac{m}{2}. \tag{24}$$

Since this holds infinitely often for every $m \in \mathbb{N}$, the rescaled objective gap has infinite limsup, as required by Theorem 2.2.

Case C: few vulnerable iterations imply limited cumulative movement. It remains to consider the case in which

$$N(s) \le Ks$$

for all sufficiently large s. Here large steps cannot occur too frequently. To make this precise, let $c_1 \le \cdots \le c_k$ be the nonexceptional values of τ_t in any finite set of indices. Since at least j values satisfy $\tau_t \le c_j$,

$$j \le N(c_j) \le Kc_j,$$

and therefore

$$\sum_{j=1}^{k} c_j^{-1/2} \le \sqrt{K} \sum_{j=1}^{k} j^{-1/2} \le 2\sqrt{Kk}. \tag{25}$$

Because

$$\alpha_t = t^{-r/2} \tau_t^{-1/2},$$

applying this estimate on dyadic blocks yields

$$\sum_{s<t} \alpha_s = O\big(t^{(1-r)/2}\big). \tag{26}$$

Thus, when the schedule is not sufficiently vulnerable to recurrent shocks, its cumulative movement is necessarily limited. The complete derivation appears in Lemmas E.4 and E.5 in Appendix E.

We exploit this bound with a sufficiently flat objective. Choose an even integer

$$m > \frac{4r}{3r-1}$$

and a globally convex function with a 1-Lipschitz derivative that agrees with $\lambda\,|x|^m/m$ on $[-1, 1]$; Appendix E gives an explicit smooth continuation outside this interval. Starting from $x_0 = 1/2$ and choosing λ sufficiently small, deterministic gradient descent remains in $(0, 1)$ and satisfies

$$x_{t+1} = x_t\big(1 - \lambda \alpha_t x_t^{m-2}\big).$$

Introduce $Z_t := x_t^{-(m-2)}$. A single iteration changes Z_t by an amount comparable to α_t. Consequently, (26) implies

$$Z_t = O\big(1 + t^{(1-r)/2}\big).$$

Since $f(x_t) = \lambda x_t^m/m$, this gives, for some $c > 0$ and all sufficiently large t,

$$f(x_t) = \frac{\lambda}{m} Z_t^{-m/(m-2)} \ge c\, t^{-\frac{1-r}{2}\frac{m}{m-2}}. \tag{27}$$

The choice $m > 4r/(3r-1)$ is precisely what ensures

$$\frac{1-r}{2}\frac{m}{m-2} < r.$$

Hence $(t+1)^r f(x_t) \to \infty$ deterministically. In this final case, no stochastic construction is needed: the oracle is simply the exact gradient, viewed as a degenerate i.i.d. oracle, and it satisfies the same normalized ABC bound as in the stochastic cases.

The three cases exhaust all positive deterministic schedules. Either the schedule leaves enough iterations vulnerable for recurrent finite-variance shocks to destroy the target rate, or it accumulates too little movement to make sufficiently fast progress on a flat objective. This dichotomy is the main mechanism behind the cube-root lower bound.

4 Further discussions

4.1 Exact $t^{-1/3}$ rate and the role of the noise assumptions

The upper and lower bounds hold under different levels of generality for the stochastic gradient oracle. Theorem 2.1 allows history-dependent stochastic gradients, provided they are conditionally unbiased and satisfy the ABC second-moment bound. By contrast, the lower bound in Theorem 2.2 already holds in a much simpler setting: for every fixed deterministic schedule and $r > 1/3$, the counterexample can be chosen to be one-dimensional with a fixed time-homogeneous i.i.d. oracle and normalized constants $L = 1$, $A = 0$, $B = 1$, and $C = 1$.

The stochastic counterexamples have finite variance, but they may have heavy tails. The ABC condition controls the conditional second moment without imposing a uniform bound on tail probabilities. In Case B of the construction, the shock distribution is chosen so that the large number of vulnerable iterations compensates for the small probability of large shocks. In particular,

$$\sum_t \mathbb{P}(H \geq m\tau_t) = \infty \quad \text{for every } m \geq 1, \qquad \mathbb{E}H = 1.$$

This construction is therefore excluded by stronger assumptions such as bounded noise or a uniform sub-Gaussian tail bound. Whether such assumptions permit faster almost-sure rates remains open.

Our results also leave open whether the exact $t^{-1/3}$ rate can be guaranteed under a fixed time-homogeneous i.i.d. oracle. Proposition 2.4 reaches this rate up to a slowly varying logarithmic factor. Under the more general conditional-noise model, however, Proposition C.1 in Appendix C shows that the exact $O(t^{-1/3})$ rate can fail even for $\alpha_t = (t+1)^{-2/3}$.

Beyond the exact-rate question, several directions remain open. In particular, the best slowly varying correction under a fixed i.i.d. oracle is unknown. Our lower bound also assumes that the deterministic schedule is fixed in advance. Whether adaptive stepsizes chosen from observations along the SGD trajectory can improve the $1/3$ threshold remains open, as does the behaviour of schedules designed for a prescribed terminal horizon.

5 Conclusion

This paper identifies the fastest almost-sure last-iterate polynomial order, up to a slowly varying logarithmic correction, for smooth convex SGD under the conditional ABC condition. The critical schedule $\alpha_t = (t+1)^{-2/3}$ attains every exponent below $1/3$, while a logarithmically corrected schedule has the optimal cube-root polynomial order. Conversely, no positive deterministic schedule fixed in advance can guarantee a higher polynomial exponent. Thus the cube-root barrier is not a consequence of restricting attention to regularly decaying stepsizes.

The proof also explains the origin of this threshold. The upper bound combines two conditional recursions to convert weighted progress into a last-iterate rate. The lower bound shows that every deterministic schedule is limited either by recurrent finite-variance shocks or by insufficient progress on a sufficiently flat objective. Whether the exact $t^{-1/3}$ rate can be attained under a fixed i.i.d. oracle, whether stronger tail assumptions permit faster rates, and whether adaptive stepsizes can improve the $1/3$ threshold remain open.

Reproducibility Statement

Complete proofs and explicit constructions are provided in the appendix. The anonymized supplementary material accompanying this submission contains the Lean 4 source formalizing Theorem 2.2, the pinned toolchain, and instructions for auditing the assumptions.

AI Use Statement

OpenAI generative-AI tools, including Codex and ChatGPT, and Anthropic's Claude were used to assist with editing and reorganizing the manuscript, checking notation and exposition, refining mathematical statements, and developing and debugging the Lean 4 formalization. They were not used to formulate the research questions, develop the theoretical framework, or determine the research methodology. The authors reviewed and revised all AI-assisted materials and take full responsibility for the final submission.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012. doi: 10.1109/TIT.2011.2182178.

Dimitri P. Bertsekas and John N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000. doi: 10.1137/S1052623497331063.

Léon Bottou. Stochastic learning. In *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pp. 146–168. Springer, 2004. doi: 10.1007/978-3-540-28650-9_7.

Guillaume Garrigos, Daniel Cortild, Lucas Ketels, and Juan Peypouquet. Last-iterate complexity of SGD for convex and smooth stochastic problems, 2025. URL https://arxiv.org/abs/2507.14122.

Antoine Godichon-Baggioni. l^p and almost sure rates of convergence of averaged stochastic gradient algorithms: Locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873, 2019. doi: 10.1051/ps/2019013.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1579–1613, 2019. URL https://proceedings.mlr.press/v99/harvey19a.html.

Marcel Hudiani. Convergence rate for the last iterate of stochastic gradient descent schemes, 2025. URL https://arxiv.org/abs/2507.07281.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1752–1755, 2019. URL https://proceedings.mlr.press/v99/jain19a.html.

Rajeeva Laxman Karandikar and Mathukumalli Vidyasagar. Convergence rates for stochastic approximation: Biased noise with unbounded variance, and applications. *Journal of Optimization Theory and Applications*, 203(3):2412–2450, 2024. doi: 10.1007/s10957-024-02547-7.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=AU4qHN2VkS.

Guy Kornowski and Ohad Shamir. Gradient descent's last iterate is often (slightly) suboptimal, 2026. URL https://arxiv.org/abs/2604.13870.

Yunwen Lei, Lei Shi, and Zheng-Chu Guo. Convergence of unregularized online learning algorithms. *Journal of Machine Learning Research*, 18(171):1–33, 2018. URL https://www.jmlr.org/papers/v18/17-457.html.

Yuqing Liang and Dongpo Xu. Almost sure convergence rates of stochastic proximal gradient descent algorithm. *Optimization*, 73(8):2413–2446, 2024. doi: 10.1080/02331934.2023.2230976.

Yuqing Liang, Dongpo Xu, Naimin Zhang, and Danilo P. Mandic. Almost sure convergence of stochastic composite objective mirror descent for non-convex non-smooth optimization. *Optimization Letters*, 18(9):2113–2131, 2024. doi: 10.1007/s11590-023-01972-3.

Jun Liu and Ye Yuan. On almost sure convergence rates of stochastic gradient methods. In *Proceedings of the Thirty-Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 2963–2983, 2022. URL https://proceedings.mlr.press/v178/liu22d.html.

Jun Liu and Ye Yuan. Almost sure convergence rates analysis and saddle avoidance of stochastic gradient methods. *Journal of Machine Learning Research*, 25(271):1–40, 2024. URL https://www.jmlr.org/papers/v25/23-1436.html.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xxaEhwC1I4.

Jianhao Ma and Yuxin Chen. A lower bound for stepsize-based acceleration of gradient descent, 2026. URL https://arxiv.org/abs/2608.10418.

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

Mariane Pelletier. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic Processes and Their Applications*, 78(2):217–244, 1998. doi: 10.1016/S0304-4149(98)00029-5.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Jagdish S. Rustagi (ed.), *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, 1971. doi: 10.1016/B978-0-12-604550-5.50015-8.

Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3935–3971, 2021. URL https://proceedings.mlr.press/v134/sebbouh21a.html.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 71–79, 2013. URL https://proceedings.mlr.press/v28/shamir13.html.

Chung-En Tsai, Ilyas Fatkhullin, Liang Zhang, and Niao He. Lower bounds for anytime acceleration of gradient descent, 2026. URL https://arxiv.org/abs/2607.02053.

Simon Weissmann, Sara Klein, Waïss Azizian, and Leif Döring. Almost sure convergence of stochastic gradient methods under gradient domination. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=OTwnNBxZFB.

Zihan Zhang, Jason D. Lee, Simon S. Du, and Yuxin Chen. Anytime acceleration of gradient descent. In *Proceedings of the Thirty-Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 5991–6013, 2025. URL https://proceedings.mlr.press/v291/zhang25a.html.

This appendix contains the complete upper- and lower-bound proofs, the logarithmic boundary calculation, an endpoint example for time-varying conditional noise, and the abstract rate-transfer principle. It uses the notation and assumptions introduced in Section 2.

A PRELIMINARY RESULTS

This section collects the supporting results used later. We first record the required smooth convex inequalities and probabilistic lemmas. We then derive two conditional SGD recursions and show how they convert weighted summability into last-iterate decay.

A.1 SMOOTH CONVEX INEQUALITIES

Lemma A.1 (First-order and smoothness inequalities)**.** *For every $x \in \mathbb{R}^d$ and $z \in S$,*

$$f(x) - f^* \geq 0, \tag{S1}$$
$$\langle \nabla f(x), x - z \rangle \geq f(x) - f^*, \tag{S2}$$
$$\|\nabla f(x)\|^2 \leq 2L\big(f(x) - f^*\big), \tag{S3}$$
$$f(x) - f^* \leq \frac{L}{2}\|x - z\|^2. \tag{S4}$$

Proof. Equation (S1) is the definition of f^*. Convexity gives $f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle$, which rearranges to (S2).

The descent lemma for an L-Lipschitz gradient states

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

Taking $y = x - L^{-1}\nabla f(x)$ yields

$$f^* \leq f(y) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2,$$

which is (S3). Finally, differentiability and optimality give $\nabla f(z) = 0$. Apply the descent lemma with base point z and endpoint x to obtain (S4). □

Define once and for all

$$D := A + 2BL, \qquad u := \frac{LD}{2}, \qquad v := \frac{LC}{2}. \tag{S5}$$

The ABC condition (3) of Khaled & Richtárik (2023) then reduces as follows.

Corollary A.2 (Reduced ABC bound)**.** *Almost surely,*

$$\mathbb{E}_t\|g_t\|^2 \leq D\big(f(x_t) - f^*\big) + C. \tag{S6}$$

Proof. Substitute (S3) into (3). □

Lemma A.3 (Finite-time square integrability)**.** *If $\mathbb{E}\|x_0\|^2 < \infty$, then x_t and g_t are square-integrable for every finite t.*

Proof. We argue by induction. If $x_t \in L^2$, then (S4) gives

$$0 \leq f(x_t) - f^* \leq \frac{L}{2}\|x_t - z\|^2$$

for any fixed $z \in S$, so the objective gap is integrable. Taking expectations in (S6) yields $\mathbb{E}\|g_t\|^2 < \infty$. The update $x_{t+1} = x_t - \alpha_t g_t$ then gives $x_{t+1} \in L^2$. The base case is the assumed square integrability of x_0. □

A.2 PROBABILISTIC LEMMAS

We use the following deterministic-error form of the Robbins–Siegmund almost-supermartingale theorem (Robbins & Siegmund, 1971) for the upper bound. The adaptive-escape lemma is used in the stochastic lower bound.

Lemma A.4 (Deterministic-error almost-supermartingale)**.** *Let Y_t, U_t be nonnegative integrable $\mathcal{F}_t$-measurable random variables. Let $c_t \geq 0$ be deterministic and $\sum_t c_t < \infty$. If*

$$\mathbb{E}_t Y_{t+1} \leq Y_t - U_t + c_t \tag{S7}$$

almost surely for every t, then Y_t converges almost surely to a finite random variable and

$$\sum_{t=0}^{\infty} U_t < \infty \qquad \text{almost surely.} \tag{S8}$$

Proof. Let $r_t := \sum_{s=t}^{\infty} c_s$ and $Z_t := Y_t + r_t$. Since $r_t = c_t + r_{t+1}$, (S7) gives

$$\mathbb{E}_t Z_{t+1} \leq Z_t - U_t \leq Z_t.$$

Thus (Z_t) is a nonnegative supermartingale and converges almost surely to a finite random variable Z_∞. Since $r_t \to 0$ and $Y_t = Z_t - r_t$,

$$Y_t \longrightarrow Z_\infty \qquad \text{almost surely.}$$

Taking ordinary expectations in (S7) and summing from $t = 0$ to n gives

$$\sum_{t=0}^{n} \mathbb{E}U_t \leq \mathbb{E}Y_0 - \mathbb{E}Y_{n+1} + \sum_{t=0}^{n} c_t \leq \mathbb{E}Y_0 + \sum_{t=0}^{\infty} c_t.$$

Set $S_n := \sum_{t=0}^{n} U_t$ and $S := \sum_{t=0}^{\infty} U_t \in [0, \infty]$. Since $S_n \uparrow S$, the monotone convergence theorem yields

$$\mathbb{E}S = \lim_{n\to\infty} \mathbb{E}S_n \leq \mathbb{E}Y_0 + \sum_{t=0}^{\infty} c_t < \infty.$$

Therefore $\mathbb{P}(S = \infty) = 0$, which proves (S8). □

We also use the following conditional Borel–Cantelli lemma.

Lemma A.5 (Adaptive escape)**.** *Let $E_t \in \mathcal{F}_{t+1}$ and let $q_t \in [0, 1]$ be deterministic. If*

$$\mathbb{P}(E_t \mid \mathcal{F}_t) \geq q_t \quad \text{almost surely,} \qquad \sum_t q_t = \infty, \tag{S9}$$

then E_t occurs infinitely often almost surely.

Proof. Fix $n \geq 0$ and define $A_{n,N} := \bigcap_{t=n}^{N} E_t^c$ for $N \geq n$. Since $A_{n,N-1} \in \mathcal{F}_N$,

$$\begin{aligned}\mathbb{P}(A_{n,N}) &= \mathbb{E}\big[\mathbf{1}_{A_{n,N-1}} \mathbb{E}(\mathbf{1}_{E_N^c} \mid \mathcal{F}_N)\big] \\ &\leq (1 - q_N)\mathbb{P}(A_{n,N-1}).\end{aligned}$$

Iteration gives

$$\mathbb{P}(A_{n,N}) \leq \prod_{t=n}^{N}(1 - q_t) \leq \exp\left(-\sum_{t=n}^{N} q_t\right) \longrightarrow 0.$$

The events $A_{n,N}$ decrease to $A_n := \bigcap_{t=n}^{\infty} E_t^c$, so continuity from above gives $\mathbb{P}(A_n) = 0$. Finally,

$$\{E_t \text{ occurs only finitely often}\} = \bigcup_{n=0}^{\infty} A_n,$$

a countable union of null events. □

A.3 TWO CONDITIONAL RECURSIONS FOR SGD

Fix a deterministic minimizer $z \in S$.

Lemma A.6 (Squared-distance recursion)**.** *Almost surely,*

$$\mathbb{E}_t \|x_{t+1} - z\|^2 \le \|x_t - z\|^2 - (2\alpha_t - D\alpha_t^2)\big(f(x_t) - f^*\big) + C\alpha_t^2. \tag{S10}$$

Proof. Expanding the SGD update $x_{t+1} = x_t - \alpha_t g_t$ gives

$$\|x_{t+1} - z\|^2 = \|x_t - z\|^2 - 2\alpha_t \langle g_t, x_t - z\rangle + \alpha_t^2 \|g_t\|^2 .$$

The vector $x_t - z$ is $\mathcal{F}_t$-measurable. Conditional unbiasedness therefore yields

$$\mathbb{E}_t \langle g_t, x_t - z\rangle = \langle \nabla f(x_t), x_t - z\rangle .$$

Use (S2) and (S6) to obtain (S10). □

Lemma A.7 (Objective-gap recursion)**.** *Almost surely,*

$$\mathbb{E}_t\big[f(x_{t+1}) - f^*\big] \le (1 + u\alpha_t^2)\big(f(x_t) - f^*\big) + v\alpha_t^2. \tag{S11}$$

Proof. The descent lemma and the SGD update $x_{t+1} = x_t - \alpha_t g_t$ give pathwise

$$f(x_{t+1}) - f^* \le f(x_t) - f^* - \alpha_t \langle \nabla f(x_t), g_t\rangle + \frac{L\alpha_t^2}{2} \|g_t\|^2 .$$

Take conditional expectation. By (2), the middle conditional expectation is $\|\nabla f(x_t)\|^2$. Hence (S6) gives

$$\mathbb{E}_t\big[f(x_{t+1}) - f^*\big] \le f(x_t) - f^* - \alpha_t \|\nabla f(x_t)\|^2 + \frac{L\alpha_t^2}{2}\big(D(f(x_t) - f^*) + C\big).$$

Discard the nonpositive gradient term and use (S5). □

Lemma A.8 (Weighted gap summability)**.** *Suppose* $\sum_t \alpha_t^2 < \infty$*. Then, for every deterministic* $z \in S$*,*

$$\|x_t - z\|^2 \text{ converges}, \qquad \sum_{t=0}^{\infty} \alpha_t\big(f(x_t) - f^*\big) < \infty \quad \text{almost surely}. \tag{S12}$$

Proof. Square summability implies $\alpha_t \to 0$. Choose T_0 so that $D\alpha_t \le 1$ for every $t \ge T_0$. Then $2\alpha_t - D\alpha_t^2 \ge \alpha_t$, and (S10) gives

$$\mathbb{E}_t \|x_{t+1} - z\|^2 \le \|x_t - z\|^2 - \alpha_t\big(f(x_t) - f^*\big) + C\alpha_t^2, \qquad t \ge T_0.$$

Apply Lemma A.4 to the shifted filtration $\widetilde{\mathcal{F}}_k := \mathcal{F}_{T_0+k}$ and the shifted processes

$$Y_k := \|x_{T_0+k} - z\|^2 , \quad U_k := \alpha_{T_0+k}\big(f(x_{T_0+k}) - f^*\big), \quad c_k := C\alpha_{T_0+k}^2.$$

The error series is summable, so the distance converges on the tail and

$$\sum_{t=T_0}^{\infty} \alpha_t\big(f(x_t) - f^*\big) < \infty \qquad \text{almost surely.}$$

Restoring finitely many initial terms does not change convergence, and the omitted finite sum is almost surely finite. This proves both conclusions. □

A.4 LAST-ITERATE DECAY CRITERION

Combining the objective-gap and squared-distance recursions with Lemma A.8 yields the following last-iterate criterion.

Lemma A.9 (Two-recursion criterion for last-iterate decay)**.** *Assume the hypotheses of Lemma A.8. Let $b_t > 0$ be deterministic. Suppose there exist $\kappa_1, \kappa_2 \geq 0$ and T_0 such that, for $t \geq T_0$,*

$$b_{t+1} - b_t \leq \kappa_1 \alpha_t, \tag{S13}$$

$$b_{t+1}\alpha_t \leq \kappa_2, \tag{S14}$$

and

$$\sum_t b_{t+1}\alpha_t^2 < \infty, \tag{S15}$$

$$\sum_t \frac{\alpha_t}{b_t} = \infty. \tag{S16}$$

Then

$$b_t\big(f(x_t) - f^*\big) \longrightarrow 0 \qquad \textit{almost surely}. \tag{S17}$$

Proof. Fix $z \in S$, and write $F_t := f(x_t) - f^*$ and $R_t := \|x_t - z\|^2$. Put

$$\kappa := \kappa_1 + u\kappa_2, \qquad W_t := b_t F_t + \kappa R_t.$$

First multiply the objective-gap recursion (S11) by the positive deterministic weight b_{t+1}. Using (S13) and (S14),

$$\begin{aligned}\mathbb{E}_t[b_{t+1}F_{t+1}] &\leq b_t F_t + \big(b_{t+1} - b_t + ub_{t+1}\alpha_t^2\big)F_t + vb_{t+1}\alpha_t^2 \\ &\leq b_t F_t + \kappa\alpha_t F_t + vb_{t+1}\alpha_t^2.\end{aligned}$$

Since $\alpha_t \to 0$, after increasing T_0 if necessary we may also assume $D\alpha_t \leq 1$ for $t \geq T_0$. Hence $2\alpha_t - D\alpha_t^2 \geq \alpha_t$, and the squared-distance recursion (S10) implies

$$\mathbb{E}_t R_{t+1} \leq R_t - \alpha_t F_t + C\alpha_t^2, \qquad t \geq T_0.$$

Multiply this inequality by κ and add it to the preceding weighted objective-gap inequality. The terms $+\kappa\alpha_t F_t$ and $-\kappa\alpha_t F_t$ cancel exactly, giving

$$\mathbb{E}_t W_{t+1} \leq W_t + (vb_{t+1} + \kappa C)\alpha_t^2. \tag{S18}$$

The error series is finite by (S15) and $\sum_t \alpha_t^2 < \infty$. Lemma A.4, with $U_t = 0$, makes W_t converge. Lemma A.8 makes R_t converge. Therefore $b_t F_t = W_t - \kappa R_t$ has a finite nonnegative limit, say ζ.

If $\zeta > 0$, then eventually $F_t \geq \zeta/(2b_t)$, and hence

$$\sum_t \alpha_t F_t \geq \frac{\zeta}{2}\sum_t \frac{\alpha_t}{b_t} = \infty,$$

contrary to (S12). Formally, apply the argument on each event $\{\zeta \geq 1/m\}$ and take their countable union. Thus $\zeta = 0$ almost surely. □

Remark A.10 (Role of the objective recursion). Lemma A.8 alone permits sparse recurrent spikes of the objective gap. The second recursion forces the rescaled gap $b_t(f(x_t) - f^*)$ to have a limit. If that limit were positive, then the gap would remain of order b_t^{-1} eventually, contradicting weighted summability and (S16). The cancellation therefore converts cumulative progress into a genuine last-iterate rate by excluding spikes at the target scale.

B Proof of Theorem 2.1

B.1 Polynomial schedules

Proof of Theorem 2.1. All conditions below are unchanged by deleting finitely many terms, so we work on the tail $t \geq t_0$ where (5) holds. Since $p > 1/2$,

$$\sum_t \alpha_t^2 \leq a_+^2 \sum_t (t+1)^{-2p} < \infty.$$

Lemma A.8 applies. The divergent progress series needed later follows from $q \leq 1-p$ and is checked explicitly in (S22).

Fix q satisfying (6) and let $b_t = (t+1)^q$. Since $q \leq 1-p < 1$, concavity of $s \mapsto s^q$ gives

$$b_{t+1} - b_t = (t+2)^q - (t+1)^q \leq q(t+1)^{q-1} \leq \frac{q}{a_-}\alpha_t. \tag{S19}$$

Moreover, $q \leq 1-p < p$, and therefore

$$b_{t+1}\alpha_t \leq 2^q a_+ (t+1)^{q-p} \leq 2^q a_+. \tag{S20}$$

The two comparison series are

$$\sum_t b_{t+1}\alpha_t^2 \leq 2^q a_+^2 \sum_t (t+1)^{q-2p} < \infty, \tag{S21}$$

$$\sum_t \frac{\alpha_t}{b_t} \geq a_- \sum_t (t+1)^{-(p+q)} = \infty. \tag{S22}$$

The first assertion uses $q < 2p-1$. The second uses $q \leq 1-p$. All hypotheses of Lemma A.9 are now verified, and it yields (7).

It remains to prove iterate convergence. The set S is closed and separable, so choose a countable dense subset $Q \subseteq S$. Apply Lemma A.8 separately to every deterministic $z \in Q$ and intersect the resulting countably many probability-one events. On that event, $\|x_t - z\|$ has a limit for every $z \in Q$. In particular (x_t) is bounded. Since $f(x_t) - f^* \to 0$ and f is continuous, every cluster point belongs to S.

Let $x_{t_j} \to x_\infty \in S$ be one convergent subsequence. Approximate x_∞ by $z_n \in Q$. The reverse triangle inequality gives

$$\left| \|x_t - x_\infty\| - \|x_t - z_n\| \right| \leq \|x_\infty - z_n\|.$$

Because the distance to each z_n converges, sending $n \to \infty$ shows that $\|x_t - x_\infty\|$ has a limit. Along t_j that limit is zero, so the entire sequence converges to x_∞. □

Corollary B.1 (Decay exponents for polynomially decaying stepsizes)**.** *For $p \in (1/2, 1)$, every*

$$0 < q < \min\{1-p, 2p-1\}$$

is admissible. If $p > 2/3$, the endpoint $q = 1-p$ is also admissible. In particular, $p = 2/3$ gives

$$f(x_t) - f^* = o(t^{-q}) \quad \textit{almost surely for every } q < 1/3. \tag{S23}$$

Proof. Only simultaneous quantification needs comment. Choose rationals $q_n \uparrow 1/3$, intersect the probability-one events on which $(t+1)^{q_n}(f(x_t) - f^*) \to 0$, and fix a real $q < 1/3$. Select n with $q < q_n$. Then

$$(t+1)^q \big(f(x_t) - f^*\big) = (t+1)^{q-q_n}\big((t+1)^{q_n}(f(x_t) - f^*)\big) \to 0.$$

This avoids an uncountable intersection of probability-one events. □

B.2 Logarithmic boundary refinement

The exact $O(t^{-1/3})$ endpoint is not asserted. The following explicit schedule reaches it up to a slowly varying factor.

Proof of Proposition 2.4. The hypotheses of Lemmas A.8 and A.9 are eventual, while their series conditions are unchanged by a finite prefix. Hence the arbitrary positive steps preceding the displayed definition play no role in the argument. Put $b_t = t^{1/3}/\ell(t)$. For the continuous extension $b(x)$,

$$\frac{d}{dx}\log b(x) = \frac{1}{3x}\left(1 - \frac{1}{\log x} - \frac{1+\delta}{\log x \log\log x}\right). \tag{S24}$$

Increase the starting time until the parenthesis lies in $[0,1]$. The mean value theorem, after one further finite enlargement, gives

$$b_{t+1} \le 2b_t, \qquad 0 \le b_{t+1} - b_t \le \frac{b_t}{t} = \frac{\alpha_t}{\eta}. \tag{S25}$$

Also, since $\ell(t) \ge 1$ eventually,

$$b_{t+1}\alpha_t \le 2\eta t^{-1/3}\ell(t)^{-2} \le 2\eta. \tag{S26}$$

The noise comparison is

$$b_{t+1}\alpha_t^2 \le \frac{2\eta^2}{t\ell(t)^3} = \frac{2\eta^2}{t\log t(\log\log t)^{1+\delta}}. \tag{S27}$$

Its series converges by the integral substitution $u = \log\log x$:

$$\int_{t_0}^{\infty} \frac{dx}{x\log x(\log\log x)^{1+\delta}} = \frac{1}{\delta(\log\log t_0)^{\delta}} < \infty.$$

Since $\ell(t) \ge 1$ eventually, $\sum_t \alpha_t^2 < \infty$ by comparison with $\sum_t t^{-4/3}$. Finally,

$$\frac{\alpha_t}{b_t} = \frac{\eta}{t}, \qquad \sum_t \frac{\alpha_t}{b_t} = \infty. \tag{S28}$$

Lemma A.9 applies and gives $b_t(f(x_t) - f^*) \to 0$, which is (11). □

C Time-varying noise at the boundary

The main theorem leaves the exact endpoint open for a fixed i.i.d. oracle. Under the broader conditional formulation, however, the critical schedule need not satisfy an exact $O(t^{-1/3})$ bound.

Proposition C.1 (Time-varying endpoint shocks)**.** *Let $f(x) = x^2/2$, $x_0 = 1/2$, and $\alpha_t = (t+1)^{-2/3}$ for every $t \ge 0$. There is a time-varying stochastic-gradient oracle driven by independent innovations such that*

$$\mathbb{E}_t g_t = f'(x_t), \qquad \mathbb{E}_t |g_t|^2 = |f'(x_t)|^2 + 1,$$

and the resulting SGD iterates satisfy

$$\limsup_{t\to\infty} (t+1)^{1/3} f(x_t) = +\infty \qquad \textit{almost surely}.$$

Proof. For $t \ge 0$, set $a_t := (t+2)\log(t+2)$. Since $a_t \ge 2\log 2 > 1$, we may define independent innovations by

$$\xi_t = \begin{cases} +\sqrt{a_t}, & \text{with probability } (2a_t)^{-1}, \\ -\sqrt{a_t}, & \text{with probability } (2a_t)^{-1}, \\ 0, & \text{with probability } 1 - a_t^{-1}. \end{cases}$$

Let $g_t = x_t + \xi_t$, and let $\mathcal{F}_t$ be the completion of $\sigma(x_0, \xi_s : 0 \le s < t)$. Then x_t is $\mathcal{F}_t$-measurable and g_t is $\mathcal{F}_{t+1}$-measurable. Moreover, $\mathbb{E}\xi_t = 0$ and $\mathbb{E}\xi_t^2 = 1$. Independence of ξ_t from $\mathcal{F}_t$ therefore gives

$$\mathbb{E}_t g_t = x_t = f'(x_t), \qquad \mathbb{E}_t |g_t|^2 = x_t^2 + 1 = |f'(x_t)|^2 + 1.$$

Moreover, $\xi_t \in L^2$, so induction through the update gives $x_t, g_t \in L^2$ at every finite time.

Put $h_t := (1-\alpha_t)x_t$ and define the predictable sign

$$s_t := \begin{cases} -1, & h_t \geq 0, \\ +1, & h_t < 0. \end{cases}$$

For $E_t := \{\xi_t = s_t\sqrt{a_t}\}$, independence gives

$$\mathbb{P}(E_t \mid \mathcal{F}_t) = \frac{1}{2a_t} = \frac{1}{2(t+2)\log(t+2)}.$$

The sum of these probabilities diverges, so Lemma A.5 implies that E_t occurs infinitely often almost surely. On E_t, the selected sign points away from zero, and hence

$$|x_{t+1}| = |h_t - \alpha_t s_t \sqrt{a_t}| = |h_t| + \alpha_t\sqrt{a_t} \geq \alpha_t\sqrt{a_t}.$$

Consequently,

$$(t+2)^{1/3} f(x_{t+1}) \geq \frac{1}{2}(t+2)^{1/3}\alpha_t^2 a_t = \frac{1}{2}\left(\frac{t+2}{t+1}\right)^{4/3}\log(t+2) \geq \frac{1}{2}\log(t+2).$$

The right-hand side tends to infinity along the infinitely many favorable times, which proves the claimed limsup. □

D Rate-transfer principle

Proposition D.1 (Rate transfer for two Lyapunov recursions). *Let Y_t, R_t be nonnegative integrable $\mathcal{F}_t$-measurable random variables, let $\alpha_t, b_t > 0$ be deterministic, and let $c_0, c_1, c_2 \geq 0$. Suppose*

$$\mathbb{E}_t R_{t+1} \leq R_t - \alpha_t Y_t + c_0\alpha_t^2, \tag{S28a}$$

$$\mathbb{E}_t Y_{t+1} \leq (1 + c_1\alpha_t^2)Y_t + c_2\alpha_t^2. \tag{S28b}$$

Assume $\sum_t \alpha_t^2 < \infty$ and that, eventually, for some $k_1, k_2 \geq 0$,

$$b_{t+1} - b_t \leq k_1\alpha_t, \qquad b_{t+1}\alpha_t \leq k_2, \qquad \sum_t b_{t+1}\alpha_t^2 < \infty, \qquad \sum_t \frac{\alpha_t}{b_t} = \infty.$$

Then $b_t Y_t \to 0$ almost surely.

Proof. Lemma A.4 applied to (S28a) gives convergence of R_t and $\sum_t \alpha_t Y_t < \infty$ almost surely. Put

$$k := k_1 + c_1 k_2, \qquad W_t := b_t Y_t + k R_t.$$

Multiplying (S28b) by b_{t+1} and using the two eventual scale bounds gives

$$\mathbb{E}_t[b_{t+1}Y_{t+1}] \leq b_t Y_t + k\alpha_t Y_t + c_2 b_{t+1}\alpha_t^2.$$

Add k times (S28a). The terms $k\alpha_t Y_t$ cancel, leaving

$$\mathbb{E}_t W_{t+1} \leq W_t + (c_2 b_{t+1} + k c_0)\alpha_t^2.$$

The error is summable, so Lemma A.4 makes W_t converge. Since R_t also converges, $b_t Y_t$ has a finite nonnegative limit ζ. On $\{\zeta > 0\}$, eventually $Y_t \geq \zeta/(2b_t)$, contradicting $\sum_t \alpha_t Y_t < \infty$ and $\sum_t \alpha_t/b_t = \infty$. Use the countable events $\{\zeta \geq 1/m\}$ to make the last step literal. □

E Proof of Theorem 2.2

We now prove Theorem 2.2. The counting function below splits every deterministic stepsize schedule into three exhaustive cases. Cases (A) and (B) are handled by i.i.d. shocks, while case (C) is handled by a flat deterministic objective.

To prove Theorem 2.2, it suffices to treat $r \in (1/3, 1)$. Indeed, for a target $r \geq 1$, choose any $r_0 \in (1/3, 1)$. A construction with $\limsup_t (t+1)^{r_0} f(x_t) = +\infty$ also has $\limsup_t (t+1)^r f(x_t) = +\infty$, since $(t+1)^{r-r_0} \geq 1$. The shock construction below will also be used to prove Corollary 2.3 for exponents below $1/3$, and its argument requires only $r > 0$. We therefore fix $r > 0$ through Proposition E.3.

E.1 COUNTING-FUNCTION TRICHOTOMY

Fix a positive deterministic schedule $(\alpha_t)_{t\geq 0}$. On its positive-time tail, define

$$\tau_t := \frac{t^{-r}}{\alpha_t^2}, \qquad N(s) := \#\{t \geq 1 : \tau_t \leq s\} \in \mathbb{N} \cup \{\infty\}. \tag{S29}$$

Exactly one of the following regimes occurs, and each is sufficient for our construction:

(A) some bounded sublevel set is infinite: $N(s_0) = \infty$ for some $s_0 < \infty$

(B) every bounded sublevel set is finite, but N has no eventual linear upper bound

(C) there exist $K, s_0 > 0$ such that $N(s) \leq Ks$ for every $s \geq s_0$.

If (A) fails, $N(s)$ is finite for every finite s. The negation of (C) then says that for every $K, U > 0$ there exists $s \geq U$ with $N(s) > Ks$, which is (B). Hence the trichotomy is exhaustive.

Lemma E.1 (An amplitude law in case (A))**.** *In case (A), there is a nonnegative random variable Y with unbounded support and $\mathbb{E}Y \leq 1$ such that, for every integer $m \geq 1$,*

$$\sum_{t=1}^{\infty} \mathbb{P}(Y \geq m\tau_t) = \infty. \tag{S30}$$

Proof. For example, put

$$\mathbb{P}(Y = 2^k) = 2^{-2k-1} \quad (k \geq 1)$$

and place the remaining mass at zero. Then $\mathbb{E}Y = \sum_{k\geq 1} 2^{-k-1} = 1/2$. If $\tau_t \leq s_0$ for infinitely many t, then, for fixed m, each corresponding probability is at least the positive number $\mathbb{P}(Y \geq ms_0)$. □

Lemma E.2 (An amplitude law in case (B))**.** *In case (B), there is a nonnegative random variable Y with $\mathbb{E}Y = 1$ such that, for every integer $m \geq 1$, (S30) holds.*

Proof. Recursively choose a strictly increasing sequence $v_k \geq 1$ such that

$$N(v_k) \geq k2^k v_k. \tag{S31}$$

The witnesses can be chosen arbitrarily large by case (B). Let

$$y_k := kv_k, \qquad p_k := \frac{2^{-k}}{kv_k} \quad (k \geq 1).$$

Since $v_k \geq 1$, $\sum_k p_k \leq \sum_k 2^{-k} = 1$. Place the remaining probability at zero. Moreover

$$\mathbb{E}Y = \sum_{k\geq 1} p_k y_k = \sum_{k\geq 1} 2^{-k} = 1. \tag{S32}$$

Because $\tau_t > 0$, $N(0) = 0$. Tonelli's theorem and monotonicity of N give, for every $m \geq 1$,

$$\begin{aligned}\sum_{t=1}^{\infty} \mathbb{P}(Y \geq m\tau_t) &= \mathbb{E}\left[\sum_{t=1}^{\infty} \mathbf{1}_{\{\tau_t \leq Y/m\}}\right] = \mathbb{E}N(Y/m) \\ &= \sum_{k\geq 1} p_k N(y_k/m) \\ &\geq \sum_{k\geq m} p_k N(v_k) \geq \sum_{k\geq m} 1 = \infty. \end{aligned} \tag{S33}$$

□

E.2 I.I.D. SHOCK CONSTRUCTION

Proposition E.3 (Cases (A) and (B)). *In either case (A) or case (B), the conclusion (9) holds with $f(x) = x^2/2$ and one time-homogeneous i.i.d. finite-variance oracle law.*

Proof. Let Y be furnished by Lemma E.1 or Lemma E.2. The preceding lemmas already give $\mathbb{E}Y \le 1$. On a countable product probability space, take i.i.d. copies Y_t and independent i.i.d. Rademacher signs ε_t. Put

$$\xi_t := \varepsilon_t \sqrt{Y_t}, \qquad g(x, (Y_t, \varepsilon_t)) := x + \xi_t.$$

Set $x_0 = 1/2$ and generate $x_{t+1} = x_t - \alpha_t g(x_t, (Y_t, \varepsilon_t))$. Then $\mathbb{E}\xi_t = 0$ (Cauchy–Schwarz gives integrability) and $\mathbb{E}\xi_t^2 = \mathbb{E}Y \le 1$. Therefore

$$\mathbb{E}[g(x, Z)] = x = f'(x), \qquad \mathbb{E}\,|g(x, Z)|^2 = x^2 + \mathbb{E}Y \le |f'(x)|^2 + 1. \tag{S34}$$

Since $\xi_t \in L^2$ and every deterministic α_t is finite, induction through the update gives $x_t, g(x_t, Z_t) \in L^2$ at every finite time. Independence of the fresh innovation from $\mathcal{F}_t$ turns the two static identities above into conditional unbiasedness and the conditional ABC bound.

Let $\mathcal{F}_t$ contain the initialization and innovations strictly before time t. Set $h_t := (1 - \alpha_t)x_t$, which is $\mathcal{F}_t$-measurable, and define the predictable sign

$$s_t := \begin{cases} -1, & h_t \ge 0, \\ +1, & h_t < 0. \end{cases}$$

For a fixed integer $m \ge 1$, let

$$E_t^{(m)} := \{Y_t \ge m\tau_t,\ \varepsilon_t = s_t\}.$$

Independence of the new amplitude and sign from $\mathcal{F}_t$ gives

$$\mathbb{P}(E_t^{(m)} \mid \mathcal{F}_t) = \frac{1}{2}\mathbb{P}(Y \ge m\tau_t) =: q_t^{(m)}. \tag{S35}$$

The sum of the deterministic $q_t^{(m)}$ diverges by (S30). Lemma A.5 says that $E_t^{(m)}$ occurs infinitely often almost surely.

On this event the chosen sign points away from zero, so

$$|x_{t+1}| = \left|h_t - \alpha_t \varepsilon_t \sqrt{Y_t}\right| = |h_t| + \alpha_t \sqrt{Y_t} \ge \alpha_t \sqrt{Y_t}.$$

Using (S29),

$$(t+2)^r f(x_{t+1}) \ge t^r f(x_{t+1}) \ge \frac{1}{2} t^r \alpha_t^2 Y_t = \frac{Y_t}{2\tau_t} \ge \frac{m}{2}. \tag{S36}$$

Intersect the probability-one events over the countable set $m \in \mathbb{N}_{\ge 1}$. For every height M and starting time T, choose $m > 2M$ and then an occurrence of $E_t^{(m)}$ after T. This is precisely the assertion that the limsup in (9) is infinite. □

E.3 CUMULATIVE STEP BOUND

For the remainder of the proof of Theorem 2.2, we now assume $r \in (1/3, 1)$.

Lemma E.4 (Finite reciprocal-square-root sum). *Assume case (C). There are constants $K_0, K_1 < \infty$ such that every finite set $I \subset \mathbb{N}_{\ge 1}$ satisfies*

$$\sum_{t \in I} \tau_t^{-1/2} \le K_0 + K_1 \sqrt{\#I}. \tag{S37}$$

Proof. Only finitely many τ_t lie below s_0. Let K_0 be the sum of their reciprocal square roots. For the remaining elements of I, list their τ-values increasingly as $c_1 \le \cdots \le c_k$. Then at least j indices satisfy $\tau_t \le c_j$, so

$$j \le N(c_j) \le K c_j, \qquad c_j^{-1/2} \le \sqrt{K/j}.$$

Consequently

$$\sum_{j=1}^{k} c_j^{-1/2} \le \sqrt{K} \sum_{j=1}^{k} j^{-1/2} \le 2\sqrt{Kk}.$$

Take $K_1 = 2\sqrt{K}$ and add the exceptional contribution K_0. □

Lemma E.5 (Cumulative step bound). *Under case (C), there is $K_2 < \infty$ such that*

$$S_T := \sum_{t=0}^{T-1} \alpha_t \le K_2\big(1 + T^{(1-r)/2}\big) \qquad (T \ge 1). \tag{S38}$$

Furthermore $\sup_t \alpha_t < \infty$.

Proof. First, $N(1) < \infty$. If $s_0 \le 1$, this follows from $N(1) \le K$. If $s_0 > 1$, then $N(1) \le N(s_0) \le Ks_0$. Hence eventually $\tau_t > 1$, and then

$$\alpha_t = t^{-r/2}\tau_t^{-1/2} \le 1.$$

Only finitely many exceptional steps remain, including α_0, proving their boundedness.

On the dyadic block $D_j = \{2^j, \dots, 2^{j+1} - 1\}$, Lemma E.4 gives

$$\begin{aligned} \sum_{t \in D_j} \alpha_t &= \sum_{t \in D_j} t^{-r/2}\tau_t^{-1/2} \\ &\le 2^{-jr/2}\big(K_0 + K_1 2^{j/2}\big) \le K_3\, 2^{j(1-r)/2} \end{aligned} \tag{S39}$$

after enlarging K_3. Because $(1-r)/2 > 0$, summing this geometric bound over the dyadic blocks up to T and adding α_0 proves (S38). □

E.4 Flat deterministic construction

Let $m > 2$ be an arbitrary even integer. The next two lemmas hold for every such m. For the case-(C) construction, we later specialize m so that

$$m > \max\left\{2, \frac{4r}{3r-1}\right\}. \tag{S40}$$

For $u \ge 0$, define

$$\Psi_m(u) := \begin{cases} u^m/m, & 0 \le u \le 1, \\ 1/m + (u-1) + (m-1)(u-1)^2/2, & u \ge 1. \end{cases} \tag{S41}$$

For a constant $\lambda > 0$, set

$$f_m(x) := \lambda \Psi_m(|x|). \tag{S42}$$

Lemma E.6 (Properties of the flat objective). *The function f_m is convex and C^1, has unique minimizer zero, and*

$$f_m(x) = \frac{\lambda}{m}\,|x|^m \qquad (|x| \le 1). \tag{S43}$$

With the convention $\operatorname{sgn}(0) = 0$*, its derivative is*

$$f_m'(x) = \lambda \operatorname{sgn}(x) \begin{cases} |x|^{m-1}, & |x| \le 1, \\ 1 + (m-1)(|x|-1), & |x| \ge 1, \end{cases} \tag{S44}$$

and is globally $\lambda(m-1)$-Lipschitz.

Proof. The two formulas in (S41) have equal values and first derivatives at $u = 1$. Their derivatives are respectively u^{m-1} and $1 + (m-1)(u-1)$. These are nonnegative, nondecreasing, and join continuously. Composing the resulting increasing radial derivative with the sign gives (S44), an odd nondecreasing function, so f_m is convex and has the unique minimizer zero.

On $[0, 1]$, the derivative of the positive branch has slope $\lambda(m-1)u^{m-2} \le \lambda(m-1)$. Outside it has constant slope $\lambda(m-1)$. The same holds on the negative side, and the pieces join continuously. Applying the mean value theorem on each piece, and splitting a segment at the junctions when necessary, proves the global Lipschitz bound. □

Lemma E.7 (Flat deterministic dynamics). *Let $M_\alpha := \sup_t \alpha_t < \infty$, set $x_0 = 1/2$, and choose*

$$0 < \lambda \le \min\left\{\frac{1}{m-1}, \frac{1}{2M_\alpha x_0^{m-2}}\right\}. \tag{S45}$$

Run deterministic gradient descent $x_{t+1} = x_t - \alpha_t f_m'(x_t)$. Then

$$0 < x_{t+1} \le x_t \le x_0 < 1 \tag{S46}$$

and, with $a := m-2$ and $Z_t := x_t^{-a}$,

$$a\lambda\alpha_t \le Z_{t+1} - Z_t \le a\, 2^{a+1}\lambda\alpha_t. \tag{S47}$$

Proof. As long as $0 < x_t \le x_0$, the local formula gives

$$x_{t+1} = x_t(1-u_t), \qquad u_t := \lambda\alpha_t x_t^{m-2} \le \lambda M_\alpha x_0^{m-2} \le \frac{1}{2}.$$

This proves (S46) by induction.

For $a > 0$ and $u \in [0, 1/2]$, convexity of $(1-u)^{-a}$ at zero and the mean value theorem give

$$au \le (1-u)^{-a} - 1 \le a\, 2^{a+1}u. \tag{S48}$$

Indeed, the derivative $a(1-u)^{-a-1}$ lies between a and $a\, 2^{a+1}$ on this interval. Consequently

$$Z_{t+1} - Z_t = Z_t\big((1-u_t)^{-a} - 1\big).$$

Since $Z_t u_t = x_t^{-a}\lambda\alpha_t x_t^a = \lambda\alpha_t$, inequality (S48) is exactly (S47). □

Proposition E.8 (Case (C)). *Under case (C), choose m as in (S40) and λ as in (S45). Then the deterministic oracle $g(x) = f_m'(x)$ and the objective above satisfy the conclusion (9).*

Proof. The choice (S45) makes the derivative 1-Lipschitz. The deterministic oracle is a degenerate i.i.d. oracle, is unbiased, and satisfies the common ABC bound

$$|g(x)|^2 \le |f_m'(x)|^2 + 1.$$

Sum the upper half of (S47) and use (S38):

$$Z_t \le Z_0 + a\, 2^{a+1}\lambda S_t \le K_4\big(1 + t^{(1-r)/2}\big). \tag{S49}$$

All iterates lie in $(0,1)$, so (S43) and $x_t = Z_t^{-1/a}$ give

$$f_m(x_t) = \frac{\lambda}{m} Z_t^{-m/(m-2)} \ge K_5 t^{-\frac{1-r}{2}\frac{m}{m-2}}. \tag{S50}$$

The algebraic condition (S40) is equivalent to

$$\frac{1-r}{2}\frac{m}{m-2} < r : \tag{S51}$$

indeed, after multiplying by $2(m-2) > 0$, it becomes $(3r-1)m > 4r$. Thus $(t+1)^r f_m(x_t) \to \infty$ deterministically. □

E.5 Completion of the lower-bound proof

The stochastic and deterministic constructions cover all three cases of the counting trichotomy.

Proof of Theorem 2.2. Apply the exhaustive trichotomy following (S29). Proposition E.3 handles (A) and (B), while Proposition E.8 handles (C). All examples are admissible with the common declared constants $L = 1, A = 0, B = 1, C = 1$ and the same initialization $x_0 = 1/2$. □

F Proof of Corollary 2.3

Proof. The critical schedule (8) and Theorem 2.1 give every $q < 1/3$ over $\mathfrak{C}_{\mathrm{ABC}}(\alpha^*)$, while Theorem 2.2 rules out every $r > 1/3$ already in the normalized subclass. Hence $\rho_{\mathrm{a.s.}} = 1/3$.

Now fix $p \in (1/2, 1)$ and a schedule $\alpha_t \asymp t^{-p}$. Theorem 2.1 guarantees every exponent $q < \min\{1-p, 2p-1\}$; it also guarantees the endpoint $q = 1-p$ when $p > 2/3$.

Conversely, fix

$$r > \min\{1-p, 2p-1\}.$$

If $r > 2p-1$, set $\beta := 2p - r < 1$. Then

$$\tau_t = \frac{t^{-r}}{\alpha_t^2} \asymp t^{\beta}.$$

If $\beta \le 0$, some bounded sublevel set of (τ_t) is infinite, so case (A) applies. If $0 < \beta < 1$, then $\tau_t \to \infty$ and

$$N(s) \asymp s^{1/\beta}.$$

Thus every bounded sublevel set is finite, but $N(s)/s \to \infty$, so case (B) applies. Proposition E.3 therefore rules out the exponent r.

It remains to consider $r \le 2p-1$. Since $r > \min\{1-p, 2p-1\}$, we must then have $r > 1-p$. We apply the flat construction directly, without assuming case (C). Because $\alpha_t \asymp t^{-p}$ and $p < 1$,

$$M_\alpha := \sup_t \alpha_t < \infty, \qquad \sum_{s=0}^{t-1} \alpha_s \le K(t+1)^{1-p}$$

for some finite constant K. Choose an even integer $m > 2$ such that

$$\gamma := \frac{(1-p)m}{m-2} < r,$$

which is possible because $r > 1-p$, and choose λ according to (S45). Lemmas E.6 and E.7 apply. With $a := m-2$ and $Z_t := x_t^{-a}$, summing the upper bound in (S47) gives

$$\begin{aligned} Z_t &\le Z_0 + a\, 2^{a+1} \lambda \sum_{s=0}^{t-1} \alpha_s \\ &\le K'(t+1)^{1-p} \end{aligned}$$

for some $K' < \infty$. Since every iterate lies in $(0,1)$, (S43) yields

$$f_m(x_t) = \frac{\lambda}{m} Z_t^{-m/(m-2)} \ge c(t+1)^{-\gamma}$$

for some $c > 0$. Consequently,

$$(t+1)^r f_m(x_t) \ge c(t+1)^{r-\gamma} \longrightarrow +\infty.$$

The exact-gradient oracle is a degenerate time-homogeneous i.i.d. oracle and satisfies the normalized ABC bound. Hence no exponent above $\min\{1-p, 2p-1\}$ can be guaranteed uniformly. Together with the upper bound, this proves that the supremal guaranteed exponent is $\min\{1-p, 2p-1\}$. □