THE HEAVY TAIL BEHAVIOR OF ADAM: SHARP RATES AND A PHASE TRANSITION AT $\mathfrak{p} = \frac{6}{5}$

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient noise in deep learning produces more extremes than models with light tails predict; finite $\mathfrak{p}$-moment models permit infinite variance ($1 < \mathfrak{p} < 2$), beyond classical theory. Despite Adam's observed advantage over SGD, whose degradation is partly attributed to heavy tails, analyses rely on Clip-SGD analogies or modified methods, leaving unmodified coordinatewise Adam's convergence a recurring open problem. The difficulty is that conditional second moment proxies control dependence from sample reuse at $\mathfrak{p} = 2$, but may fail below $\mathfrak{p} = 2$. We take an important step toward resolving this open problem. Specifically, we establish convergence rates for Adam under heavy-tailed noise and a classical parameter setting based on the total iteration budget T. We derive an explicit upper bound for every constant base stepsize using chord cancellation and an exact identity for an auxiliary iterate. These charge normalization bias and drift in adaptive stepsizes to realized second moment increments, avoiding noise variances and direct momentum recursion. For L-smooth objectives bounded below, with fixed $L > 0$ and adapted noise having finite conditional $\mathfrak{p}$-moments ($1 < \mathfrak{p} \leq 2$) bounded by $\sigma^{\mathfrak{p}}$, with $\sigma > 0$ fixed, we select a stepsize yielding the expected average squared gradient norm bound $\mathcal{O}\big(\min\big\{1, [\log(eT)]^2 T^{\max\{3/\mathfrak{p}-5/2, -(\mathfrak{p}-1)/\mathfrak{p}\}}\big\}\big)$. Lower bounds for $L = 1$ over all constant stepsizes establish stepsize minimax sharpness up to logarithms, with a gap of one logarithm for $\mathfrak{p} \leq \frac{6}{5}$. Thus uniform polynomial stationarity decay is achievable iff $\mathfrak{p} > \frac{6}{5}$, recovering the polynomial exponent $-\frac{1}{2}$ at $\mathfrak{p} = 2$.

1 INTRODUCTION

Stochastic gradient noise in modern deep learning is often markedly heavy-tailed, with extreme gradients occurring far more frequently than classical light-tailed models predict. Non-Gaussian or heavy-tailed gradient fluctuations have been observed in image classification, BERT pretraining, and policy gradient tasks; related theory of stochastic recursions further shows that multiplicative sampling noise can induce heavy-tailed optimization trajectories (Şimşekli et al., 2019; Zhang et al., 2020; Garg et al., 2021; Battash et al., 2024; Gürbüzbalaban et al., 2021; Hodgkinson & Mahoney, 2021). Such noise is commonly modeled by a finite conditional $\mathfrak{p}$-th moment, where $1 < \mathfrak{p} \leq 2$. For $1 < \mathfrak{p} < 2$, this assumption guarantees that the noise mean exists while allowing its variance to be infinite, placing the problem beyond classical stochastic optimization theory under finite variance.

Existing studies suggest that Adam often exhibits better empirical convergence than SGD on such heavy-tailed problems. In particular, Zhang et al. (2020) observed a clear advantage of Adam over well tuned SGD in BERT pretraining and identified heavy-tailed gradient noise as a possible cause of SGD's degradation. This phenomenon is commonly attributed to the adaptive suppression of outlying gradients by Adam's coordinatewise second moment normalization. These observations motivate a direct convergence theory for Adam under finite conditional $\mathfrak{p}$-moments, a question raised for both Adam and its variant AdamW with decoupled weight decay (Liu & Zhou, 2025; Chezhegov et al., 2025; Yu et al., 2026b; Liu, 2026a; Yu et al., 2026a).

A central difficulty in resolving this problem is the dependence between Adam's numerator and denominator. Let $g_t = \nabla f(\mathbf{x}_t) + \xi_t$ denote the current stochastic gradient. The same random vector enters both the first moment numerator and, through $g_t^{\odot 2}$, the coordinatewise second moment de-

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.

nominator. Consequently, even when $\mathbb{E}[\xi_t \mid \mathscr{F}_{t-1}] = 0$, the normalized stochastic direction can be conditionally biased. In the finite variance case $\mathfrak{p} = 2$, representative analyses construct a proxy denominator independent of the current sample, using a conditional second moment or variance bound and then use conditional unbiasedness to extract descent (Défossez et al., 2022; Wang et al., 2023). This argument need not extend to $\mathfrak{p} < 2$: the conditional second moment may be infinite, invalidating both the proxy denominator and its error control; directly expanding the coupling between the numerator and denominator merely reintroduces the nonexistent second moment.

We take an important step toward resolving this open problem. Specifically, we analyze Adam with heavy-tailed gradient noise under the classical parameter setting (Défossez et al., 2022; Wang et al., 2023; Hong & Lin, 2024) based on the total iteration budget T (Section 2.2), without clipping or additional normalization. For L-smooth objectives under a conditional $\mathfrak{p}$-moment budget for the full vector, our chord estimate bounds the expected average squared gradient norm along the actual iterates. Hard instance trajectories yield matching polynomial lower rates across $1 < \mathfrak{p} \leq 2$, up to logarithmic factors.

To address this coupling, we develop *chord cancellation*. Conditional on the past, we represent each coordinate's stochastic gradient law exactly as a mixture of two-point laws whose endpoints bracket the true gradient coordinate and whose means equal it. Conditional unbiasedness then yields an exact weighted cancellation between the chord endpoints. For Adam's normalization map, the residual bias is a difference of secant slopes, controlled by the same sample's squared gradient increment in the second moment recursion. The conditional $\mathfrak{p}$-moment enters the inequality for one step as a single local factor, charging normalization to an increment already recorded by the algorithm. This estimate accommodates infinite variance; the same moment assumption controls the noise energy when removing preconditioning.

Specifically, under the classical finite horizon parameter setting in Section 2.2, we establish stepsize minimax stationarity rates, with upper and lower bounds matching up to polylogarithmic factors:

- **A new chord cancellation method.** We develop a chord cancellation inequality that controls Adam's normalization bias using only the available conditional $\mathfrak{p}$-moment. Exact cancellation of centered linear terms converts the adverse bias into a cost controlled by realized deposits from the current squared gradients, bypassing the potentially infinite second moment in finite variance proxy arguments. Together with an exact identity for an auxiliary iterate, this yields an energy estimate compatible with adapted noise and arbitrary coordinate dependence.

- **Upper bound.** Building on the chord method above, we obtain the following guarantee. In any finite dimension d fixed independently of T, let f be bounded below, L-smooth for fixed $L > 0$, and possibly nonseparable, and let the stochastic gradient noise be adapted, conditionally unbiased, and satisfy a fixed conditional $\mathfrak{p}$-moment bound $\sigma^{\mathfrak{p}}$ on the full vector for $1 < \mathfrak{p} \leq 2$, with arbitrarily dependent coordinates. For coordinatewise Adam, Theorem 1 gives an upper bound with explicit stepsize dependence. Optimizing its explicit form yields Proposition 2 and a suitable constant stepsize for each horizon, independently of the particular objective and oracle policy with the prescribed parameters and initial gap, for which uniformly

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] = \mathcal{O}\Big(\min\Big\{1, [\log(eT)]^2\, T^{\max\{\frac{3}{\mathfrak{p}}-\frac{5}{2}, -\frac{\mathfrak{p}-1}{\mathfrak{p}}\}}\Big\}\Big).$$

- **Matching lower bounds.** For $L = 1$, fixed $\sigma > 0$, and every constant stepsize, we construct an admissible hard instance, obtaining matching lower bounds up to logarithms with constants independent of the stepsize. The uniform upper and lower guarantees differ by one logarithm for $1 < \mathfrak{p} \leq \frac{6}{5}$ and have the displayed polynomial rate for $\mathfrak{p} > \frac{6}{5}$, recovering the polynomial exponent $-\frac{1}{2}$ at $\mathfrak{p} = 2$. To our knowledge, these are the first matching bounds for direct coordinatewise Adam across $1 < \mathfrak{p} \leq 2$, without clipping, randomized stepsizes, or online-to-nonconvex conversion.

2 Problem Setup

We consider the stochastic optimization problem

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

where $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. The objective is bounded below and L-smooth for some $L > 0$, as specified in Assumptions 1 and 2 below. Let $1 < \mathfrak{p} \leq 2$, let $d \geq 1$ and $T \geq 1$ be integers, and let $\sigma \geq 0$ denote the noise scale. The objective is selected before the run and is independent of the realized oracle path.

Starting from an arbitrary point $\mathbf{x}_0 \in \mathbb{R}^d$, we access ∇f through a stochastic first-order oracle and study the resulting iterates $\{\mathbf{x}_t\}_{t=0}^T$. Let $\mathscr{F}_{t-1}$ denote the history available before the oracle query at clock t, with $\mathscr{F}_{-1}$ containing the initial information. The iterate $\mathbf{x}_t$ is $\mathscr{F}_{t-1}$-measurable. At clock t, a rule fixed before the run selects a conditional noise row using this history, draws ξ_t, and returns the stochastic gradient

$$g_t = \nabla f(\mathbf{x}_t) + \xi_t.$$

The history $\mathscr{F}_t$ includes this oracle draw. Assumption 3 states the conditional unbiasedness and $\mathfrak{p}$-moment budget for the full vector.

2.1 Adaptive Moment Estimation (Adam)

Algorithm 1 describes coordinatewise Adam. At each iteration, it queries the stochastic oracle, updates the first and second moment states $\mathbf{m}_t$ and $\mathbf{v}_t$, and uses the current second moment state to normalize the update of $\mathbf{x}_t$.

Algorithm 1 Adam

1: **Input:** Stochastic oracle satisfying Assumption 3; initial point $\mathbf{x}_0 \in \mathbb{R}^d$; stepsize $\gamma > 0$; coefficients $\beta_1, \beta_2 \in [0, 1)$; $v > 0$, $\epsilon \geq 0$; number of iterations T.
2: **Initialize:** $\mathbf{x}_{-1} = \mathbf{x}_0$, $\mathbf{m}_{-1} = 0$, $\mathbf{v}_{-1} = v\mathbf{1}$.
3: **for** $t = 0$ **to** $T - 1$ **do**
4: Draw ξ_t from the selected conditional row and compute $g_t \leftarrow \nabla f(\mathbf{x}_t) + \xi_t$.
5: Update the first moment state: $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) g_t$.
6: Update the second moment state: $\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2)(g_t \odot g_t)$.
7: Update the iterate: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \gamma \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t + \epsilon\mathbf{1}}}$.
8: **end for**
9: **Output:** Final iterate $\mathbf{x}_T$.

Notation. Unless otherwise stated, $\|\cdot\|$ denotes the Euclidean norm.

All arithmetic operations are coordinatewise: $\odot$ denotes the Hadamard product, and square roots and division are applied componentwise. For a vector indexed by time, a second subscript denotes its coordinate, as in $g_{t,i}$. Division by a zero denominator is defined as zero. We write $a_+ = \max\{a, 0\}$ for $a \in \mathbb{R}$, and $\mathbf{1}_A$ for the indicator of an event A.

2.2 Parameter settings

We study Adam with heavy-tailed gradient noise under a classical, widely used parameter setting based on the total iteration budget T (Défossez et al., 2022; Wang et al., 2023; Hong & Lin, 2024). Specifically, we take

$$\beta_1 \in [0, 1), \qquad \beta_2 = 1 - \frac{1}{T},$$

with β_1 fixed independently of (T, γ) and the instance. Both coefficients remain constant within each run, with second moment memory of order T. Asymptotic analyses use time varying $\beta_{2,t} \to 1$ (Jin et al., 2025), including $\beta_{2,t} = 1 - 1/t$ for $t \geq 2$ with suitable diminishing stepsizes (Zou et al., 2019;

Liang et al., 2025), extending the same principle of growing memory without a prescribed horizon. For $T \geq 2$, $\beta_2^T \geq 1/4$ retains past squared gradient mass and yields (8), while $-T \log \beta_2 \leq 2$ controls denominator decay in the bounds on stepsize variation of Lemma 1.

Uniform convergence requires suitable parameter restrictions even with finite second moments. Counterexamples with bounded gradients include i.i.d. convex failure with $\epsilon = 0$ for every fixed $0 \leq \beta_1 < \sqrt{\beta_2} < 1$ (Reddi et al., 2018, Theorem 3), and stationarity failure on a compact interval with $\beta_1 = 0$ and fixed β_2, stepsize, and damping (Savarese et al., 2021, Appendix B). Smooth convex finite-sum divergence also occurs in a region of decay parameters under random reshuffling (Zhang et al., 2022).

2.3 Technical Assumptions

The objective assumptions specify the lower bound and smoothness used in the descent argument. The oracle assumption imposes a conditional $\mathfrak{p}$-moment bound on the noise around the true gradient.

Assumption 1 (Objective bounded below)**.** *We have $f \in C^1(\mathbb{R}^d)$ and $f_* := \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$.*

We write $\Delta_f := f(\mathbf{x}_0) - f_*$ for the initial objective gap.

Assumption 2 (L-smoothness)**.** *There exists $L > 0$ such that $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

This condition permits the standard descent inequality. No convexity, separability, or boundedness of gradients or trajectories is assumed.

Assumption 3 (Adapted moment budget for the full vector)**.** *There exists $\mathfrak{p} \in (1, 2]$ such that, for every $0 \leq t < T$, after every predictable selection of the conditional noise row, $\mathbb{E}[\xi_t \mid \mathscr{F}_{t-1}] = 0$ and $\mathbb{E}[\|\xi_t\|^{\mathfrak{p}} \mid \mathscr{F}_{t-1}] \leq \sigma^{\mathfrak{p}}$ almost surely.*

The rule for selecting the conditional distribution is $\mathscr{F}_{t-1}$-measurable and fixed before the run. Coordinates may be arbitrarily dependent, and the noise may be adapted and time-inhomogeneous. An oblivious time-inhomogeneous oracle, including an iid oracle independent of the trajectory, is the special case in which the selection rule ignores the realized history. For $1 < \mathfrak{p} < 2$ and $\sigma > 0$, this assumption permits infinite noise variance. The budget σ is common to the admissible class and does not depend on the horizon, stepsize, initial gap, or realized path. The case $\sigma = 0$ forces $\xi_t = 0$ almost surely at every clock.

Performance measure. For the iterates of Algorithm 1, we study the expected average squared gradient norm

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big]$$

under the adapted oracle in Assumption 3. Uniform guarantees hold over all objectives and oracle policies satisfying Assumptions 1–3 with the prescribed smoothness, moment budget, and initial gap. A common stepsize may depend on these parameters and the horizon, but not on the particular objective or oracle policy. Asymptotic rates keep these parameters and Δ_f fixed as T increases. In the lower bounds, the objective and complete oracle policy may depend on (T, γ), but both are fixed before the run; the policy may use the realized history as specified in Assumption 3. All guarantees apply at any prescribed $\mathbf{x}_0$, by the translation argument in Appendix A.

3 Main Results

We first explain the difficulty created by normalization using the same sample and how chord cancellation resolves it. We then present the upper bound in Section 3.2 and the lower bound in Section 3.3. Write

$$\rho_T = 1 + \sigma^2 T^{(\frac{2}{\mathfrak{p}} - \frac{3}{2})_+}, \qquad r_*(\mathfrak{p}) = \max\left\{\frac{3}{\mathfrak{p}} - \frac{5}{2}, -\frac{\mathfrak{p}-1}{\mathfrak{p}}\right\}. \tag{1}$$

The upper bound is uniform over the admissible objectives and predictable oracle policies satisfying Assumptions 1–3. The lower bound uses objectives and complete oracle policies selected for each announced (T, γ), before the run. Both directions hold in each fixed dimension.

3.1 THE CHORD CANCELLATION METHOD

Finite variance analyses compare the denominator containing the current sample with a predictable proxy controlled by conditional second moments (Défossez et al., 2022; Wang et al., 2023). For $1 < \mathfrak{p} < 2$, Assumption 3 allows $\mathbb{E}\left[\|\xi_t\|^2 \,\middle|\, \mathscr{F}_{t-1}\right]$ to be infinite, so that comparison and a bound on the quadratic remainder in the untransformed noise need not be finite. The chord estimate below instead controls normalization bias through the available $\mathfrak{p}$-moment and the realized deposits from the current squared gradients.

The main obstacle in the upper bound occurs within a single normalized update. Under the parameters of Theorem 1, with $T \geq 2$, write $h = (\nabla f(\mathbf{x}_t))_i$, condition on $\mathscr{F}_{t-1}$, and let $g = h + \xi$, where $\mathbb{E}[\xi \mid \mathscr{F}_{t-1}] = 0$. With

$$D(z) = \sqrt{\beta_2 v_{t-1,i} + \frac{z^2}{T}} + \epsilon, \qquad W(z) = \frac{z}{D(z)}, \qquad Q(z) = \frac{z^2}{T\beta_2 v_{t-1,i} + z^2}.$$

the ideal normalized direction is $W(h)$, while the term containing the current gradient in the recursion for the auxiliary iterate uses $W(g)$. Centering gives $\mathbb{E}[g \mid \mathscr{F}_{t-1}] = h$, but it does not imply $\mathbb{E}[W(g) \mid \mathscr{F}_{t-1}] = W(h)$, because the same g appears in the numerator and the denominator. We control this bias directly through the available $\mathfrak{p}$-moment and realized deposits from the current squared gradients. If $h < 0$, reflect (h, g, ξ) to $(-h, -g, -\xi)$. It is therefore enough below to describe the adverse gap for $h \geq 0$; for $h < 0$, the reflected certificate controls $\left[\mathbb{E}[W(g) \mid \mathscr{F}_{t-1}] - W(h)\right]_+$, which is exactly the sign that can make the normalized alignment bias adverse.

Exact reduction to two-point chords. For $h \geq 0$, take two values $\ell < h < u$, write $r = h - \ell$ and $s = u - h$, and put probabilities $s/(r+s)$ and $r/(r+s)$ on ℓ and u. This two-point law has mean h. Define its left and right secant slopes by

$$S_- = \frac{W(h) - W(\ell)}{r}, \qquad S_+ = \frac{W(u) - W(h)}{s}.$$

Its adverse Jensen gap is exactly

$$\left[W(h) - \mathbb{E}[W(g)]\right]_+ = \frac{rs}{r+s}(S_- - S_+)_+.$$

Every scalar law with mean h admits an exact mixture representation by such two-point laws and a possible degenerate law at h: pair the mass below h, weighted by $h - \ell$, with the mass above h, weighted by $u - h$. Consequently, it suffices to control one chord, and no symmetry or coordinate independence is required.

Figure 1 gives a geometric illustration of the chord cancellation method.

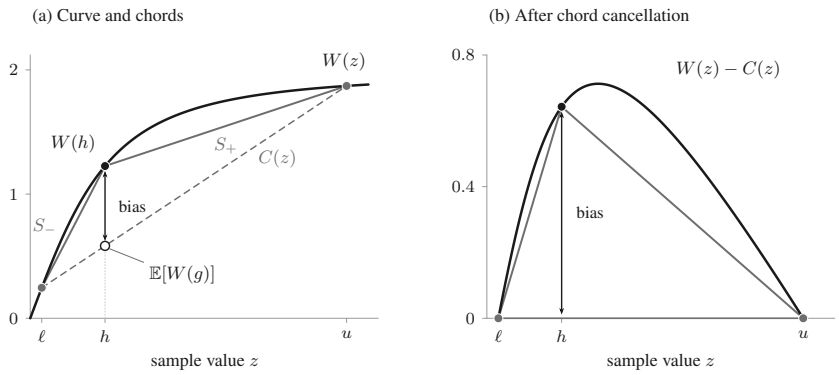


Figure 1: Chord cancellation for a two-point law with mean $h = \mathbb{E}[g]$. (a) The two solid chords have slopes S_- and S_+; the dashed chord C satisfies $C(h) = \mathbb{E}[W(g)]$. (b) Subtracting C isolates the bias $W(h) - C(h)$, since $\mathbb{E}[C(g)] = C(h)$.

The geometric core is the following secant inequality, proved in Lemma 3:

$$D(h)(S_- - S_+)_+^2 \le C(r^{\mathfrak{p}-1} + s^{\mathfrak{p}-1}) \left\{ \frac{|\ell|^{2-\mathfrak{p}} Q(\ell)}{r} + \frac{|u|^{2-\mathfrak{p}} Q(u)}{s} \right\}. \tag{2}$$

At $\mathfrak{p} = 2$, every product $|z|^{2-\mathfrak{p}} Q(z)$ below is interpreted as $Q(z)$, including at $z = 0$. The first factor on the right is precisely the centered $\mathfrak{p}$-moment price of the chord. The second factor is its realized cost from the current squared gradient. Integrating (2) over the exact chord mixture and using Cauchy–Schwarz therefore gives the certificate for one step

$$\left[W(h) - \mathbb{E}\left[W(g) \mid \mathscr{F}_{t-1}\right]\right]_+^2 \le \frac{C\,\mathbb{E}\left[|\xi|^{\mathfrak{p}} \mid \mathscr{F}_{t-1}\right]}{D(h)} \mathbb{E}\left[|g|^{2-\mathfrak{p}} Q(g) \mid \mathscr{F}_{t-1}\right]. \tag{3}$$

The certificate (3) asks only for the assumed $\mathfrak{p}$-moment. It uses the fraction $Q(g)$ contributed by the current squared gradient to control the bias. The positive initial accumulator supplies the denominator floor. This conversion of cancellation into deposits is the independent methodological contribution of our analysis of the upper bound.

3.2 Upper bound

For any initial point $\mathbf{x}_0 \in \mathbb{R}^d$, fix $1 < \mathfrak{p} \le 2$, an integer $d \ge 1$, $L > 0$, $v > 0$, and $\epsilon, \sigma, \Delta_f \ge 0$, independently of T, γ, and the instance. Under Assumptions 1–3, run Algorithm 1 with the moment coefficients in Section 2.2. The implied constants and horizon thresholds below depend only on these fixed parameters and β_1.

Theorem 1 (Upper bound for arbitrary base stepsizes)**.** *With ρ_T from (1), for all sufficiently large T and every $\gamma > 0$, uniformly over admissible objectives and oracle policies,*

$$\begin{aligned} &\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \\ &\quad = \mathcal{O}\Bigg(\left(\frac{1}{\sqrt{T}} + \frac{\sigma}{T^{1-\frac{1}{\mathfrak{p}}}} \right) \left(\rho_T + L\gamma\sqrt{T} + \frac{1}{L\gamma\sqrt{T}} \right) \log^2\!\big(eT(1 + L\gamma\sqrt{T})\big) \\ &\qquad + \frac{1}{T} \left(\rho_T + L\gamma\sqrt{T} + \frac{1}{L\gamma\sqrt{T}} \right)^2 \log^4\!\big(eT(1 + L\gamma\sqrt{T})\big) \Bigg). \end{aligned} \tag{4}$$

The formal statement is Theorem 4 in Appendix C, which gives the explicit bound (25) and its proof.

Proposition 2 (Stepsize optimization and minimax upper bound)**.** *Under the conditions of Theorem 1, a constant stepsize $\gamma = \gamma(T) > 0$ can be chosen independently of the particular instance so that, for all sufficiently large T, uniformly over admissible instances, with $r_*(\mathfrak{p})$ as defined in (1),*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] = \mathcal{O}\Big(\min\left\{ 1, [\log(eT)]^2 T^{r_*(\mathfrak{p})} \right\} \Big). \tag{5}$$

The formal statement and proof are given in Proposition 6 and Appendix D. They also characterize the stepsize minimizing the explicit upper bound (25).

Corollary 1 (A simple stepsize)**.** *Under the conditions of Theorem 1, the choice $\gamma = 1/(L\sqrt{T})$ gives, uniformly over admissible instances and with $r_*(\mathfrak{p})$ from (1),*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] = \mathcal{O}\Big([\log(eT)]^2 T^{r_*(\mathfrak{p})} \Big). \tag{6}$$

The formal statement and proof are given in Corollary 2 in Appendix D.4. For $\mathfrak{p} > \frac{6}{5}$, this stepsize gives the same polynomial exponent as (5). The optimized stepsize gives a strictly smaller value of the explicit bound in (25). The squared logarithmic factor is exact for the minimized expression at fixed $\sigma > 0$, as shown in Appendix D.2. For $1 < \mathfrak{p} \le \frac{6}{5}$, the bounded guarantee in (5) follows from the choice of a small stepsize.

3.3 Lower bound

Theorem 3 (Lower bounds for arbitrary base stepsizes). *Fix $L = 1$, $\Delta_f = 1$, $1 < \mathfrak{p} \le 2$, an integer $d \ge 1$, $v > 0$, $\epsilon \ge 0$, and $\sigma > 0$. Use the moment coefficients in Section 2.2 and the exponent $r_*(\mathfrak{p})$ from (1). For all sufficiently large T and every $\gamma > 0$, there exist an objective and a predictable oracle policy satisfying Assumptions 1–3 such that Algorithm 1, started from $\mathbf{x}_0 = 0$, satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] = \begin{cases} \Omega\left(\max\left\{\min\left\{1, \dfrac{1}{\gamma T}\right\}, \dfrac{1}{\log(eT)}\right\}\right), & 1 < \mathfrak{p} \le \frac{6}{5}, \\ \Omega\left(\max\left\{\min\left\{1, \dfrac{1}{\gamma T}\right\}, T^{r_*(\mathfrak{p})}\right\}\right), & \frac{6}{5} < \mathfrak{p} \le 2. \end{cases} \tag{7}$$

The implied constant and horizon threshold depend only on the fixed parameters, uniformly over $\gamma > 0$. The objective and complete oracle policy may depend on (T, γ), but are fixed before the run.

The formal statement is Theorem 8 in Appendix E.8; Appendix E gives the proof. The constructions use a deterministic plateau, a rare shock, a rare shock followed by a ramp with adjustable memory, and a rescaled pump with a return phase. Together they cover every base stepsize $\gamma > 0$.

Comparison of the upper and lower bounds. For $L = 1$, $\Delta_f = 1$, fixed $\sigma > 0$, and $\mathfrak{p} > \frac{6}{5}$, the upper and lower bounds give matching polynomial rates, up to logarithms, for optimization over all constant base stepsizes. Their gap is a factor of $[\log(eT)]^2$. At $\mathfrak{p} = \frac{4}{3}$, the exponent is $-\frac{1}{4}$, and at $\mathfrak{p} = 2$ it is $-\frac{1}{2}$. For $1 < \mathfrak{p} \le \frac{6}{5}$, every stepsize admits an instance whose expected average squared gradient norm is $\Omega(1/\log(eT))$, while $\mathcal{O}(1)$ is uniformly achievable. These bounds exclude uniform polynomial decay, with a gap of one logarithm between the upper bound of constant order and the lower bound.

Figure 2 summarizes the common polynomial exponent and representative rate profiles.

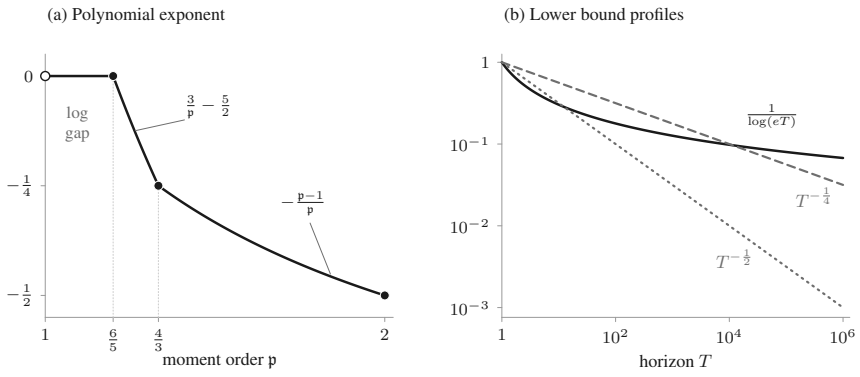


Figure 2: Stepsize minimax rates for $L = 1$, $\Delta_f = 1$, and fixed $\sigma > 0$. (a) The common polynomial exponent $\min\{0, r_*(\mathfrak{p})\}$, with $r_*(\mathfrak{p})$ from (1). (b) The rates in the lower bound $1/\log(eT)$, $T^{-1/4}$, and $T^{-1/2}$ for $\mathfrak{p} = \frac{6}{5}, \frac{4}{3}, 2$, respectively, on logarithmic axes. For $1 < \mathfrak{p} \le \frac{6}{5}$, the lower profile and the upper bound of constant order differ by one logarithm.

4 Proof Sketch

4.1 Upper bound

Step 1: Chord cancellation. Condition on the history and decompose the law of each stochastic gradient coordinate into two-point laws with the same mean. Linear cancellation leaves a difference of secant slopes controlled by the conditional $\mathfrak{p}$-moment and deposit from the current squared gradient, yielding (3) (Lemma 3, Appendix C.1).

Step 2: Stopped descent. For $\mathbf{z}_t, \boldsymbol{\lambda}_t$ in (10) and (9), (11) isolates momentum. Under the setup of Theorem 1, let $T \ge 10$, $\gamma > 0$, and $H \ge \max\{1, L\Delta_f\}$. Define $\tau_H = \inf\{1 \le t \le T :$

$L(f(\mathbf{x}_t) - \inf f) > H\}$, with $\inf \varnothing = T+1$. Pathwise for all $0 \le n \le T$, with $C = C(\beta_1, d) > 0$,

$$L\big(f(\mathbf{z}_{n\wedge\tau_H}) - f(\mathbf{x}_0)\big) \le -L\gamma \sum_{t=0}^{n\wedge\tau_H-1} \sum_{i=1}^{d} \frac{(\nabla f(\mathbf{x}_t))_i \, g_{t,i}}{\sqrt{v_{t,i}} + \epsilon} + C\left(L\gamma\sqrt{T(H+1)} + L^2\gamma^2 T\right)\left(1 + \sum_{t=0}^{n\wedge\tau_H-1} \sum_{i=1}^{d} \frac{g_{t,i}^2}{T v_{t,i}}\right).$$

Lemma 2, Appendix B, proves the formal bound (21), including the update that causes the first crossing.

Step 3: Fluctuations and removal of stopping. The chord estimate controls conditional means under arbitrary coordinate dependence. Stopped centered fluctuations are scalar martingale differences: normalization bounds gradient ratios by $\sqrt{T}$, stopping bounds true gradients, and the accumulator potential controls conditional variances and momentum errors without noise second moments. With H chosen by the failure probability, bounds for all prefixes and smoothness comparison of $\mathbf{x}_t, \mathbf{z}_t$ exclude a first crossing. Tail integration gives moments $1, 2, \mathfrak{p}/(\mathfrak{p}-1)$ of $\mathcal{A}_T$ in (33) (Proposition 5, Appendix C.2).

Step 4: Recover ordinary gradient norms. Subadditivity and the moment budget give

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=0}^{T-1} \|\xi_t\|^2\right)^{\frac{\mathfrak{p}}{2}}\right] \le \sigma^{\mathfrak{p}} T^{1-\frac{\mathfrak{p}}{2}}.$$

Together with the pathwise conversion (59), Hölder and the first two energy moments yield (25) (Lemma 4, Appendix C.3).

Step 5: Optimize the stepsize. The stepsize $\gamma = 1/(L\sqrt{T})$ gives Corollary 1; the unique minimizer of the right side of (25) strictly improves its value at this stepsize. For fixed parameters and initial gap, the quadratic term's extra $T^{-1/2}$ leaves a squared logarithm. The alternative of a small positive stepsize independent of the instance completes Proposition 2 (Appendix D).

4.2 Lower bound

For sufficiently large T, four admissible scalar constructions cover all constant stepsizes; each objective and oracle policy is fixed before the run.

Step 1: Limited travel on a plateau. For small stepsizes, use the plateau in Figure 3, with a, ℓ from (77) (Lemma 5). Zero noise and $f' = a > 0$ give nonnegative momentum and leftward motion; the denominator floor prevents escape before T:

$$\sum_{t=0}^{T-1} |x_{t+1} - x_t| \le \frac{\gamma a T}{\epsilon + \frac{\sqrt{v}}{2}} \le \ell, \qquad \frac{1}{T}\sum_{t=0}^{T-1} |f'(x_t)|^2 = a^2.$$

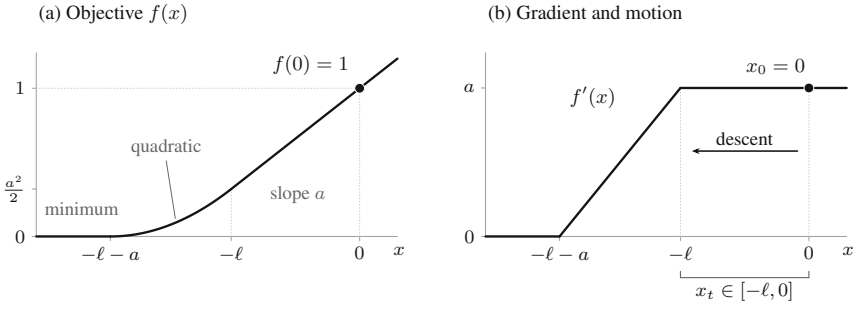


Figure 3: Deterministic plateau: noiseless iterates move left within $[-\ell, 0]$, retaining gradient a.

Step 2: A shock delays recovery. On $f(x) = x^2/2$, centered noise has rare atoms $\pm M$, with $M \asymp T^{1/\mathfrak{p}}$, total probability $q \asymp 1/T$, and $qM^{\mathfrak{p}} \le \sigma^{\mathfrak{p}}$. With constant probability, an early

isolated positive shock creates positive momentum, sends the iterate left by $\asymp \gamma\sqrt{T}$, and deposits $v_t \asymp M^2/T$. Residual momentum can push farther left; negative objective gradients oppose it, but the large denominator delays recovery. In the occupation regime, $|x_t| \asymp \gamma\sqrt{T}$ persists for $\asymp \min\{T, M/(\gamma\sqrt{T})\}$ updates. For very large stepsizes, the initial displacement suffices (Lemma 7).

Step 3: Growth and residence on a ramp. For intermediate stepsizes and $\mathfrak{p} < 2$, a rare negative shock launches the iterate rightward into $f'(x) = -x$. Subsequent noiseless gradients reinforce negative momentum, so position and gradient magnitude grow. The deposited memory limits the effective stepsize; tuning the shock balances the time to reach a band with fixed gradient magnitude against prolonged residence there (Lemma 8).

Step 4: Uphill charging and downhill residence. For the remaining stepsizes when $\mathfrak{p} \leq 4/3$, frequent negative noise overcomes a plateau with positive gradient, generating negative momentum and rightward uphill motion; rare positive compensators ensure centering. Upon reaching a launch point in a flat interval, updates with zero gradient let momentum decay geometrically. A rare negative shock then launches the iterate farther right into a descending ramp and deposits $v_t \asymp T^{2/\mathfrak{p}-1}$. Noiseless descent increases the gradient magnitude, while the small effective stepsize prolongs residence at the target scale:

$$\underbrace{\gamma T^{\frac{2}{\mathfrak{p}}-1}}_{\text{squared gradient}} \quad \underbrace{\frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma}}_{\text{residence time}} \quad \frac{1}{T} = T^{\frac{3}{\mathfrak{p}}-\frac{5}{2}}.$$

The phases fit within T, with success probability bounded below independently of (T, γ) (Lemma 9). Splicing ranges, completing the gap outside the reachable region, and embedding in one coordinate proves Theorem 3 (Proposition 7, Appendix E.8): matching polynomial rates for $\mathfrak{p} > 6/5$, and no uniform polynomial decay for $1 < \mathfrak{p} \leq 6/5$.

5 Related Work

Foundations include robust stochastic approximation (Nemirovski et al., 2009; Nazin et al., 2019) and adaptive methods (Duchi et al., 2011; Ward et al., 2019). Heavy-tailed analyses cover clipping and normalization (Zhang et al., 2020; Gorbunov et al., 2020; Cutkosky & Mehta, 2021; Nguyen et al., 2023; Liu & Zhou, 2025; Hübler et al., 2025; Sun et al., 2025; Fradin et al., 2026), additional smoothness or noise structure (Liu et al., 2023; Puchkin et al., 2024), unclipped SGD/SGDM (Wang et al., 2021; Fatkhullin et al., 2025; Liu, 2026c), clipped Adam-Norm/AdaGrad-Norm (Chezhegov et al., 2025), AdaGrad for $\mathfrak{p} > \frac{4}{3}$ (Liu, 2026a), normalized preconditioning (Fang et al., 2026), and sign methods (Yu et al., 2026b).

Adam (Kingma & Ba, 2015) has convex failure examples (Reddi et al., 2018), alongside convergence and complexity guarantees under suitable second moment or stronger assumptions (Chen et al., 2019; Zou et al., 2019; Zhou et al., 2024; Zhang et al., 2022; Défossez et al., 2022; Wang et al., 2023; Li et al., 2023; Hong & Lin, 2024; Jin et al., 2025).

Liu (2026b) and Pang (2026) analyze Adam with bias correction and exponential stepsize randomization: coordinatewise with $\beta_1 = \beta_2$ tuned to the horizon, and scalar squared norm accumulation with coefficients tuned to accuracy, respectively. Unclipped bounds on distributionally smoothed ℓ_1 and Euclidean stationarity, respectively, vanish for $\mathfrak{p} > \frac{4}{3}$. Pang covers $1 < \mathfrak{p} \leq 2$ by clipping online outputs at a known radius. Yu et al. (2026a) study weighted stationarity for AdamW.

6 Conclusion

We characterize the heavy tail behavior of unclipped coordinatewise Adam under the finite horizon parameter setting in Section 2.2. For smooth nonconvex objectives and conditional $\mathfrak{p}$-moment noise, $1 < \mathfrak{p} \leq 2$, chord cancellation and an exact momentum identity yield an expected average squared gradient bound for every constant base stepsize. Lower bounds from explicit trajectories match the polynomial rates up to logarithmic factors. With $L = 1$, $\Delta_f = 1$, and fixed $\sigma > 0$, these bounds identify the transition at $\mathfrak{p} = 6/5$: uniform polynomial stationarity decay is achievable precisely when $\mathfrak{p} > 6/5$.

REPRODUCIBILITY STATEMENT

Section 2 specifies the objective and oracle assumptions and the algorithm. Section 2.2 gives the moment coefficients, and Section 3.2 gives the remaining common setup for Theorem 1 and Proposition 2; Theorem 3 states its parameters for the lower bound. The appendix proves the expected upper bound in Theorem 1 through the martingale and stopping arguments, and establishes the stochastic lower bounds through the constructions, trajectory estimates, and gap completion. The Lean 4 and Mathlib formalization in the supplementary material covers selected algebraic identities, centering relations for noise distributions, exponent comparisons, logarithmic bounds, and the unique global minimizer of the explicit scalar profile for the upper bound.

AI USE STATEMENT

Generative AI tools assisted with manuscript writing and revision, literature searches and reference verification, proof writing and checking, and LaTeX typesetting. The authors take responsibility for the final content.

REFERENCES

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4780–4788. PMLR, 2024.

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

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10269–10333. PMLR, 2025.

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

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

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

Yuchen Fang, James Demmel, and Javad Lavaei. Why is normalization preferred? a worst-case complexity theory for stochastically preconditioned SGD under heavy-tailed noise. *arXiv preprint arXiv:2602.13413*, 2026.

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can SGD handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*, 2025.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2548–2556. PMLR, 2026.

Saurabh Garg, Joshua Zhanson, Emilio Parisotto, Adarsh Prasad, J. Zico Kolter, Zachary C. Lipton, Sivaraman Balakrishnan, Ruslan Salakhutdinov, and Pradeep Ravikumar. On proximal policy optimization's heavy-tailed gradients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3610–3619. PMLR, 2021.

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

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3964–3975. PMLR, 2021.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4262–4274. PMLR, 2021.

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

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2413–2421. PMLR, 2025.

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

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

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

Yuqing Liang, Meixuan He, Jinlan Liu, and Dongpo Xu. Convergence of Adam for non-convex objectives: Relaxed hyperparameters and non-ergodic case. *Machine Learning*, 114(3):75, 2025.

Zijian Liu. Can adaptive gradient methods converge under heavy-tailed noise? a case study of AdaGrad. In *International Conference on Machine Learning*, 2026a.

Zijian Liu. Adam converges in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2606.22326*, 2026b.

Zijian Liu. In-expectation convergence of stochastic gradient methods under heavy-tailed noise. *arXiv preprint arXiv:2606.00520*, 2026c.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 2266–2290. PMLR, 2023.

Alexander V. Nazin, Arkadi S. Nemirovsky, Alexandre B. Tsybakov, and Anatoli B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

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.

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

Yijiang Pang. The convergence behavior of Adam under heavy-tailed noise. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 5244–5265. PMLR, 2026.

Nikita Puchkin, Eduard Gorbunov, Nikolay Kutuzov, and Alexander Gasnikov. Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 856–864. PMLR, 2024.

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

Pedro Savarese, David McAllester, Sudarshan Babu, and Michael Maire. Domain-independent dominance of adaptive methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16286–16295, 2021.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5827–5837. PMLR, 2019.

Tao Sun, Xinwang Liu, and Kun Yuan. Revisiting gradient normalization and clipping for nonconvex SGD under heavy-tailed noise: Necessity, sufficiency, and acceleration. *Journal of Machine Learning Research*, 26(237):1–42, 2025.

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

Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Şimşekli, and Murat A. Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18866–18877, 2021.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6677–6686. PMLR, 2019.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Open problem: Is AdamW effective under heavy-tailed noise? *arXiv preprint arXiv:2606.23676*, 2026a.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. *arXiv preprint arXiv:2602.07425*, 2026b.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020.

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

Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *Transactions on Machine Learning Research*, 2024.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11127–11135, 2019.

A Quantifiers and Notation

We use f_* and Δ_f as defined in Section 2.3. Conditional expectations at clock t are taken with respect to the history sigma-algebra $\mathscr{F}_{t-1}$.

For Theorem 1, use the moment coefficients specified in Section 2.2 and allow every base stepsize $\gamma > 0$. The remaining parameters

$$1 < \mathfrak{p} \le 2, \qquad L > 0, \qquad d \in \mathbb{N},\ d \ge 1, \qquad v > 0, \qquad \epsilon \ge 0, \qquad \sigma \ge 0$$

are fixed independently of T and γ. Upper bounds hold uniformly over admissible instances with each prescribed initial gap $\Delta_f \ge 0$. All constants and horizon thresholds in the energy argument for the upper bound depend only on $(\mathfrak{p}, d, v, \beta_1)$. They never depend on L, γ, ϵ, σ, Δ_f, T, $\mathbf{x}_0$, the admissible objective or oracle, a stopping height, or a confidence parameter. Recall ρ_T from (1):

$$\rho_T = 1 + \sigma^2 T^{(\frac{2}{\mathfrak{p}} - \frac{3}{2})_+}.$$

All vector products, quotients, and square roots are coordinatewise. Theorem 3 also uses Section 2.2, with $L = 1$; its assertion is uniform over all $\gamma > 0$. For $T \ge 2$ and $0 \le t < T$, the positive initial accumulator gives

$$\begin{gathered} v_{t,i} \ge \beta_2^{t+1} v \ge \beta_2^T v \ge \frac{v}{4}, \qquad \beta_2 v_{t-1,i} \ge \frac{v}{4}, \\ \sqrt{v_{t,i}} + \epsilon \ge \epsilon + \frac{\sqrt{v}}{2} > 0. \end{gathered} \tag{8}$$

The explicit threshold T_{u} in (37) satisfies $T_{\mathrm{u}} \ge 10$.

All conditional identities below hold almost surely after conditioning on $\mathscr{F}_{t-1}$. The conditional row at clock t may be any $\mathscr{F}_{t-1}$-measurable function of the completed history, provided that its conditional mean and conditional $\mathfrak{p}$-moment of the full vector satisfy Assumption 3. The rule selecting those rows is fixed before the run. The upper proof applies to these predictable conditional laws, including rows independent of the trajectory when the rule ignores its history.

The effective stepsize is

$$\lambda_{t,i} = \frac{\gamma}{\sqrt{v_{t,i}} + \epsilon}. \tag{9}$$

The accumulator recursion also gives

$$0 \le \frac{g_{t,i}^2}{T v_{t,i}} \le 1.$$

The auxiliary point that removes the direct first moment recursion is

$$\mathbf{z}_t = \frac{\mathbf{x}_t - \beta_1 \mathbf{x}_{t-1}}{1 - \beta_1}. \tag{10}$$

Translation of the initial point. For a prescribed $\mathbf{x}_0 \in \mathbb{R}^d$, translate positions by $\mathbf{y} = \mathbf{x} - \mathbf{x}_0$ and evaluate the objective as $f(\mathbf{x}_0 + \mathbf{y})$. Translation is a bijection of $\mathbb{R}^d$, so

$$\inf_{\mathbf{y}} f(\mathbf{x}_0 + \mathbf{y}) = f_*, \qquad f(\mathbf{x}_0) - \inf_{\mathbf{y}} f(\mathbf{x}_0 + \mathbf{y}) = \Delta_f,$$

and the chain rule gives

$$\nabla_{\mathbf{y}} \big(f(\mathbf{x}_0 + \mathbf{y}) \big) = \nabla f(\mathbf{x}_0 + \mathbf{y}).$$

Thus continuous differentiability and the gradient bound in Assumption 2 are preserved. Transport the oracle policy by adding $\mathbf{x}_0$ to each iterate in its input history and leaving all other history variables unchanged. This deterministic, measurable transformation defines a policy fixed before the run. On the same probability space, use the same noise draws and filtration. Starting the transported instance at zero, induction in the three updates of Algorithm 1 gives

$$\mathbf{y}_t = \mathbf{x}_t - \mathbf{x}_0.$$

The stochastic gradients and both moment states coincide at every clock: the initial moment states agree, the translated objective and oracle give identical stochastic gradients, and the two moment recursions therefore produce identical states. Their normalized increments preserve the iterate difference. This also holds with the convention for zero denominators. Conditional unbiasedness and the moment budget for the full vector are unchanged because the noise and conditioning sigma-algebra are unchanged. The gradient norms agree at every clock. The inverse translation gives a bijection between admissible instances at the two starting points, preserving their gaps and expected performance. Thus uniform upper guarantees and individual witnesses for lower bounds transfer between the two starting points with exactly the same expected performance. The constructions for the lower bounds, with $\Delta_f = 1$, may therefore use $\mathbf{x}_0 = 0$.

The generic objective is denoted by f. Only the constructive lower bound uses f_T, meaning a deterministic objective selected for the announced horizon and fixed before that run.

B Exact Momentum Reduction

Lemma 1 (Pathwise momentum reduction)**.** *Run Algorithm 1 with the moment coefficients in Section 2.2, $T \geq 10$, and $\gamma > 0$. With $\boldsymbol{\lambda}_t$ and $\mathbf{z}_t$ defined in (9) and (10), the following hold pathwise for $0 \leq t < T$, $1 \leq i \leq d$, and every integer endpoint $0 \leq N \leq T$, including stopped endpoints:*

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\boldsymbol{\lambda}_t \odot g_t + \frac{\beta_1}{1-\beta_1}(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t) \odot \mathbf{m}_{t-1}, \tag{11}$$

$$\sum_{t=0}^{N-1} \|\boldsymbol{\lambda}_t \odot \mathbf{m}_t\|^2 + \sum_{t=0}^{N-1} \|\boldsymbol{\lambda}_{t-1} \odot \mathbf{m}_{t-1}\|^2 \leq 8 \sum_{t=0}^{N-1} \|\boldsymbol{\lambda}_t \odot g_t\|^2, \tag{12}$$

$$|(\lambda_{t-1,i} - \lambda_{t,i}) m_{t-1,i}| \leq 2\gamma\sqrt{T}\left(\frac{g_{t,i}^2}{T v_{t,i}} + \frac{2}{T}\right). \tag{13}$$

Proof. Iterating the first moment recursion from $\mathbf{m}_{-1} = 0$ gives

$$\mathbf{m}_t = (1-\beta_1)\sum_{k=0}^{t} \beta_1^{t-k} g_k. \tag{14}$$

Since $\mathbf{x}_{t+1} = \mathbf{x}_t - \boldsymbol{\lambda}_t \odot \mathbf{m}_t$,

$$\begin{aligned} \mathbf{z}_{t+1} &= \mathbf{x}_t - \frac{\boldsymbol{\lambda}_t \odot \mathbf{m}_t}{1-\beta_1} \\ &= \mathbf{x}_t - \frac{\beta_1}{1-\beta_1}\boldsymbol{\lambda}_t \odot \mathbf{m}_{t-1} - \boldsymbol{\lambda}_t \odot g_t, \end{aligned}$$

whereas the definition (10) gives

$$\mathbf{z}_t - \mathbf{x}_t = -\frac{\beta_1}{1-\beta_1}\boldsymbol{\lambda}_{t-1} \odot \mathbf{m}_{t-1} \qquad (0 \leq t \leq T). \tag{15}$$

Subtracting proves (11), including $t = 0$ because $\mathbf{x}_{-1} = \mathbf{x}_0$ and $\mathbf{m}_{-1} = 0$.

The accumulator recursion gives $v_{\ell,i} \geq \beta_2 v_{\ell-1,i}$, so, for $0 \leq k \leq h < T$,

$$v_{h,i} \geq \beta_2^{h-k} v_{k,i}, \qquad \lambda_{h,i} \leq 2\lambda_{k,i}, \tag{16}$$

where the second bound follows from

$$\frac{\lambda_{h,i}}{\lambda_{k,i}} \leq \beta_2^{-\frac{h-k}{2}} \leq \beta_2^{-\frac{T}{2}} < 2.$$

Since $v_{t,i} \geq g_{t,i}^2/T$, equations (14) and (16) give

$$\begin{aligned} |\lambda_{t,i} g_{t,i}| &\leq \gamma\sqrt{T}, \\ |\lambda_{t,i} m_{t,i}| &\leq 2(1-\beta_1)\sum_{k=0}^{t} \beta_1^{t-k} |\lambda_{k,i} g_{k,i}| \leq 2\gamma\sqrt{T}. \end{aligned} \tag{17}$$

Weighted Cauchy–Schwarz similarly gives

$$(\lambda_{t,i} m_{t,i})^2 \le 4(1-\beta_1) \sum_{k=0}^{t} \beta_1^{t-k} (\lambda_{k,i} g_{k,i})^2.$$

Summing first over $t < N$ and then reversing the finite sums proves half of (12); the shifted sum has the same bound and a zero initial term. The proof is pathwise, so a random stopped endpoint is allowed.

If $v_{t,i} \ge v_{t-1,i}$, rationalization gives

$$\frac{\lambda_{t-1,i} - \lambda_{t,i}}{\lambda_{t-1,i}} \le \frac{v_{t,i} - v_{t-1,i}}{v_{t,i}} \le \frac{g_{t,i}^2}{T v_{t,i}}.$$

If $v_{t,i} < v_{t-1,i}$, then $v_{t,i} \ge \beta_2 v_{t-1,i}$ and

$$\frac{\lambda_{t,i} - \lambda_{t-1,i}}{\lambda_{t-1,i}} \le \beta_2^{-\frac{1}{2}} - 1 \le \frac{2}{T}.$$

Taking positive parts in the two cases gives

$$0 \le (\lambda_{t-1,i} - \lambda_{t,i})_+ \le \lambda_{t-1,i} \frac{g_{t,i}^2}{T v_{t,i}}, \tag{18}$$

$$0 \le (\lambda_{t,i} - \lambda_{t-1,i})_+ \le \frac{2}{T} \lambda_{t-1,i}. \tag{19}$$

Multiplying the sum of (18) and (19) by $|m_{t-1,i}|$ and applying (17) proves (13); at $t = 0$, use $\mathbf{m}_{-1} = 0$. □

The next lemma controls first moment momentum for the process stopped at its first crossing of a prescribed function value. Its estimates hold pathwise.

Lemma 2 (Descent inequality for the stopped process)**.** *Under Assumptions 1 and 2, run Algorithm 1 with the moment coefficients specified in Section 2.2, $T \ge 10$, and $\gamma > 0$. For $H \ge \max\{1, L\Delta_f\}$, define the stopping time*

$$\tau_H = \inf\{1 \le t \le T : L(f(\mathbf{x}_t) - f_*) > H\}, \qquad \inf \varnothing = T + 1. \tag{20}$$

Consider the stopped process $(\mathbf{x}_{n\wedge\tau_H}, \mathbf{z}_{n\wedge\tau_H})_{n=0}^{T}$. Pathwise, simultaneously for every $0 \le n \le T$,

$$\begin{aligned} L\big(f(\mathbf{z}_{n\wedge\tau_H}) - f(\mathbf{z}_0)\big) \le -L \sum_{t=0}^{n\wedge\tau_H - 1} \langle \nabla f(\mathbf{x}_t), \boldsymbol{\lambda}_t \odot g_t \rangle + L^2 \sum_{t=0}^{n\wedge\tau_H - 1} \|\boldsymbol{\lambda}_t \odot g_t\|^2 \\ + \frac{128 d^2}{(1-\beta_1)^2} (L\gamma\sqrt{T(H+1)} + L^2\gamma^2 T) \left(1 + \sum_{t=0}^{n\wedge\tau_H - 1} \sum_{i=1}^{d} \frac{g_{t,i}^2}{T v_{t,i}}\right). \end{aligned} \tag{21}$$

Proof. Since $\mathbf{x}_t$ is $\mathscr{F}_{t-1}$-measurable, the event $\{t < \tau_H\}$ is $\mathscr{F}_{t-1}$-measurable. The stopped auxiliary process satisfies, for every $0 \le t < T$,

$$\mathbf{z}_{(t+1)\wedge\tau_H} - \mathbf{z}_{t\wedge\tau_H} = \mathbf{1}_{\{t<\tau_H\}} \Big(-\boldsymbol{\lambda}_t \odot g_t + \frac{\beta_1}{1-\beta_1} (\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t) \odot \mathbf{m}_{t-1} \Big),$$

by (11). Fix $0 \le n \le T$ and put $N = n \wedge \tau_H$. The stopping rule (20), together with $L\Delta_f \le H$, gives $L(f(\mathbf{x}_t) - f_*) \le H$ for every $t < N$. Lemma 1 applies pathwise to this endpoint.

Smoothness and the finite lower bound imply the gradient bound

$$\|\nabla f(\mathbf{x})\|^2 \le 2L(f(\mathbf{x}) - f_*), \tag{22}$$

because the descent inequality gives

$$f_* \le f\left(\mathbf{x} - \frac{\nabla f(\mathbf{x})}{L}\right) \le f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2.$$

Equations (15), (17), and $\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\| \le L\|\mathbf{z}_t - \mathbf{x}_t\|$ give

$$\begin{aligned} \|\mathbf{z}_t - \mathbf{x}_t\| &\le \frac{2\beta_1\gamma\sqrt{dT}}{1-\beta_1}, && 0 \le t \le T, \\ \|\nabla f(\mathbf{z}_t)\| &\le \sqrt{2H} + \frac{2\beta_1 L\gamma\sqrt{dT}}{1-\beta_1}, && 0 \le t < T \wedge \tau_H. \end{aligned} \tag{23}$$

The gradient estimate also uses (22) and the stopping rule.

The individual bounds on the sums of squares in the proof of Lemma 1 and Cauchy–Schwarz over time give

$$\begin{aligned} L\sum_{t=0}^{N-1} |\langle \nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t), \boldsymbol{\lambda}_t \odot g_t\rangle| &\le \frac{\beta_1 L^2}{1-\beta_1}\left(\sum_{t=0}^{N-1}\|\boldsymbol{\lambda}_{t-1}\odot \mathbf{m}_{t-1}\|^2\right)^{\frac{1}{2}}\left(\sum_{t=0}^{N-1}\|\boldsymbol{\lambda}_t \odot g_t\|^2\right)^{\frac{1}{2}} \\ &\le \frac{2\beta_1 L^2}{1-\beta_1}\sum_{t=0}^{N-1}\|\boldsymbol{\lambda}_t \odot g_t\|^2. \end{aligned}$$

The effective stepsize (9) gives

$$L^2\sum_{t=0}^{N-1}\|\boldsymbol{\lambda}_t \odot g_t\|^2 \le L^2\gamma^2 T\sum_{t=0}^{N-1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{Tv_{t,i}}.$$

Summing (13) over coordinates yields

$$\|(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t)\odot \mathbf{m}_{t-1}\|_1 \le 2\gamma\sqrt{T}\left(\sum_{i=1}^{d}\frac{g_{t,i}^2}{Tv_{t,i}} + \frac{2d}{T}\right).$$

Put $S = \sum_{t=0}^{N-1}\sum_{i=1}^{d} g_{t,i}^2/(Tv_{t,i})$. Using (23), $d \ge 1$, $T \ge 1$, and $0 \le \beta_1 < 1$, we obtain

$$\begin{aligned} L\sum_{t=0}^{N-1}|\langle \nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t), \boldsymbol{\lambda}_t \odot g_t\rangle| &\le \frac{2d^2}{(1-\beta_1)^2}L^2\gamma^2 T(1+S), \\ \frac{\beta_1 L}{1-\beta_1}\sum_{t=0}^{N-1}|\langle \nabla f(\mathbf{z}_t), (\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t)\odot \mathbf{m}_{t-1}\rangle| & \\ \le \frac{2\beta_1 L\gamma\sqrt{T}}{1-\beta_1}\left(\sqrt{2H} + \frac{2\beta_1 L\gamma\sqrt{dT}}{1-\beta_1}\right)(S+2d) & \\ \le \frac{d^2}{(1-\beta_1)^2}\left(6L\gamma\sqrt{T(H+1)} + 8L^2\gamma^2 T\right)(1+S). & \end{aligned}$$

Squaring the coordinatewise bound in (13) and using $0 \le g_{t,i}^2/(Tv_{t,i}) \le 1$ gives

$$\begin{aligned} \frac{\beta_1^2 L^2}{(1-\beta_1)^2}\sum_{t=0}^{N-1}\|(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t)\odot \mathbf{m}_{t-1}\|^2 & \\ \le \frac{8\beta_1^2}{(1-\beta_1)^2}L^2\gamma^2 TS + \frac{32\beta_1^2}{(1-\beta_1)^2}L^2\gamma^2 d & \\ \le \frac{32d^2}{(1-\beta_1)^2}L^2\gamma^2 T(1+S). & \end{aligned}$$

Consequently,

$$
\begin{aligned}
& L \sum_{t=0}^{N-1} |\langle \nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t), \boldsymbol{\lambda}_t \odot g_t \rangle| \\
& \quad + \frac{\beta_1 L}{1-\beta_1} \sum_{t=0}^{N-1} |\langle \nabla f(\mathbf{z}_t), (\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t) \odot \mathbf{m}_{t-1} \rangle| \\
& \quad + \frac{\beta_1^2 L^2}{(1-\beta_1)^2} \sum_{t=0}^{N-1} \|(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t) \odot \mathbf{m}_{t-1}\|^2 \\
& \quad \le \frac{d^2}{(1-\beta_1)^2} \big(6L\gamma\sqrt{T(H+1)} + 42L^2\gamma^2 T\big)(1+S) \\
& \quad \le \frac{128 d^2}{(1-\beta_1)^2} \big(L\gamma\sqrt{T(H+1)} + L^2\gamma^2 T\big)(1+S). \qquad (24)
\end{aligned}
$$

Finally, smoothness along the stopped auxiliary process gives

$$
\begin{aligned}
& L\big(f(\mathbf{z}_{(t+1)\wedge\tau_H}) - f(\mathbf{z}_{t\wedge\tau_H})\big) \\
& \quad \le \mathbf{1}_{\{t<\tau_H\}} \Big(-L\langle \nabla f(\mathbf{z}_t), \boldsymbol{\lambda}_t \odot g_t \rangle + L^2 \|\boldsymbol{\lambda}_t \odot g_t\|^2 \\
& \qquad + \frac{\beta_1 L}{1-\beta_1} \langle \nabla f(\mathbf{z}_t), (\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t) \odot \mathbf{m}_{t-1} \rangle \\
& \qquad + \frac{\beta_1^2 L^2}{(1-\beta_1)^2} \|(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t) \odot \mathbf{m}_{t-1}\|^2 \Big).
\end{aligned}
$$

Using $\nabla f(\mathbf{z}_t) = \nabla f(\mathbf{x}_t) + (\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t))$, sum over $0 \le t < n$. The left side telescopes to $L(f(\mathbf{z}_{n\wedge\tau_H}) - f(\mathbf{z}_0))$, while the right side contains precisely the updates $t < N$. Applying (24) proves (21). When $\tau_H \le n$, the sum includes the update at $t = \tau_H - 1$ that produces the first crossing. All estimates hold on every path for each n, hence simultaneously for $0 \le n \le T$. The argument uses realized normalized squares throughout. □

C Formal Statement and Proof of the Expected Upper Bound

Theorem 4 (Formal statement of Theorem 1)**.** *Fix $1 < \mathfrak{p} \le 2$, an integer $d \ge 1$, $L > 0$, $v > 0$, and $\epsilon \ge 0$. Under Assumptions 1–3, run Algorithm 1 from any $\mathbf{x}_0 \in \mathbb{R}^d$. Use the moment coefficients specified in Section 2.2. There are $C_E, T_{\mathrm{u}} > 0$, depending only on $(\mathfrak{p}, d, v, \beta_1)$, such that, for every $\Delta_f, \sigma \ge 0$ and every integer $T \ge T_{\mathrm{u}}$, every $\gamma > 0$ gives the following bound, uniformly over all admissible objectives and oracle policies with initial gap Δ_f and noise budget $\sigma^{\mathfrak{p}}$:*

$$
\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] & \le \frac{2C_E(1+L\Delta_f)}{\sqrt{T}} \left(\epsilon + \sqrt{dv} + \sqrt{2}\sigma T^{\frac{1}{\mathfrak{p}} - \frac{1}{2}}\right) \\
& \quad \times \left(\rho_T + L\gamma\sqrt{T} + \frac{1}{L\gamma\sqrt{T}}\right) \log^2\!\big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2\big) \\
& \quad + \frac{4C_E^2(1+L\Delta_f)^2}{T} \left(\rho_T + L\gamma\sqrt{T} + \frac{1}{L\gamma\sqrt{T}}\right)^2 \\
& \quad \times \log^4\!\big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2\big). \qquad (25)
\end{aligned}
$$

Here ρ_T is defined in (1). The constants T_{u} and C_E are specified in (37) and (38).

The upper proof has one conditional one-dimensional ingredient and one global stopping argument. First moment momentum enters only through Lemma 2.

C.1 CHORD CANCELLATION

Lemma 3 (Chord cancellation and deposit inequality). *For $T \geq 2$, fix a clock $0 \leq t < T$ and a coordinate i, and condition on the predictable history. Expectations in this lemma and its proof are with respect to the resulting fixed one-dimensional probability law. The accumulator satisfies $\beta_2 v_{t-1,i} \geq \frac{v}{4}$. Let $h \geq 0$, $g = h + \xi$, $\mathbb{E}[\xi] = 0$, and $m = \mathbb{E}[|\xi|^{\mathfrak{p}}] < \infty$. Write*

$$D(z) = \sqrt{\beta_2 v_{t-1,i} + \frac{z^2}{T}} + \epsilon, \qquad W(z) = \frac{z}{D(z)}, \qquad Q(z) = \frac{z^2}{T\beta_2 v_{t-1,i} + z^2}. \tag{26}$$

Then

$$[W(h) - \mathbb{E}[W(g)]]_+^2 \leq \frac{Cm}{D(h)} \mathbb{E}\left[|g|^{2-\mathfrak{p}} Q(g)\right]. \tag{27}$$

Proof. Differentiating the response W in (26) gives

$$\begin{aligned} W'(z) &= \frac{\frac{\beta_2 v_{t-1,i}}{\sqrt{\beta_2 v_{t-1,i} + \frac{z^2}{T}}} + \epsilon}{(\sqrt{\beta_2 v_{t-1,i} + \frac{z^2}{T}} + \epsilon)^2}, \\ |W''(z)| &\leq \frac{C}{\sqrt{T}(\sqrt{\beta_2 v_{t-1,i}} + \epsilon + \frac{|z|}{\sqrt{T}})^2}. \end{aligned} \tag{28}$$

Here $\beta_2 v_{t-1,i} \geq \frac{v}{4} > 0$ ensures positive denominators. Thus W' is positive, even, and nonincreasing on $[0, \infty)$. For $\ell < h < u$, let $r = h - \ell$, $s = u - h$, and let

$$S_- = \frac{W(h) - W(\ell)}{r}, \qquad S_+ = \frac{W(u) - W(h)}{s}. \tag{29}$$

We first prove the deterministic secant estimate

$$D(h)(S_- - S_+)_+^2 \leq C(r^{\mathfrak{p}-1} + s^{\mathfrak{p}-1}) \left\{ \frac{|\ell|^{2-\mathfrak{p}} Q(\ell)}{r} + \frac{|u|^{2-\mathfrak{p}} Q(u)}{s} \right\}. \tag{30}$$

Put $Z = \max\{|\ell|, u\}$. The right side of (30) is at least a fixed multiple of $Q(Z)$. Indeed, if $Z = u$, then $u \geq s$ and

$$s^{\mathfrak{p}-1} \frac{u^{2-\mathfrak{p}} Q(u)}{s} = \left(\frac{u}{s}\right)^{2-\mathfrak{p}} Q(u) \geq Q(u).$$

If $Z = |\ell|$, necessarily $\ell < 0$, and $|\ell| \leq r \leq 2|\ell|$. Hence

$$r^{\mathfrak{p}-1} \frac{|\ell|^{2-\mathfrak{p}} Q(\ell)}{r} = \left(\frac{|\ell|}{r}\right)^{2-\mathfrak{p}} Q(\ell) \geq 2^{\mathfrak{p}-2} Q(\ell).$$

Suppose first that $Z \leq \sqrt{T\beta_2 v_{t-1,i}}$. The secant slopes in (29) are averages of W' over subintervals of $[-Z, Z]$. Therefore (28) gives

$$|S_- - S_+| \leq 2Z \sup_{|z| \leq Z} |W''(z)| \leq \frac{CZ}{\sqrt{T}(\sqrt{\beta_2 v_{t-1,i}} + \epsilon)^2}.$$

Since $h \le Z$, $D(h) \le C(\sqrt{\beta_2 v_{t-1,i}} + \epsilon)$, and $\beta_2 v_{t-1,i} \ge \frac{v}{4}$,

$$D(h)(S_- - S_+)^2 \le \frac{CZ^2}{T(\sqrt{\beta_2 v_{t-1,i}} + \epsilon)^3} \le C\frac{Z^2}{T\beta_2 v_{t-1,i} + Z^2} = CQ(Z).$$

Suppose next that $Z > \sqrt{T\beta_2 v_{t-1,i}}$. The definition of Q in (26) gives $Q(Z) > \frac{1}{2}$, while $S_+ \ge 0$. We claim that $D(h)S_- \le 2$. If $\ell \ge 0$, then $D(h) \ge D(\ell)$ and

$$D(h)S_- = \frac{h - \frac{\ell D(h)}{D(\ell)}}{h - \ell} \le 1.$$

If $\ell = -z$ with $z \ge h$, then $D(z) \ge D(h)$ and

$$D(h)S_- = \frac{h + \frac{zD(h)}{D(z)}}{h + z} \le 1.$$

Finally, if $\ell = -z$ with $0 < z < h$, direct squaring gives $zD(h) \le hD(z)$, and hence

$$D(h)S_- = \frac{h + \frac{zD(h)}{D(z)}}{h + z} \le 2.$$

Consequently,

$$D(h)(S_- - S_+)_+^2 \le D(h)S_-^2 \le \frac{4}{D(h)} \le CQ(Z),$$

where the last constant is uniform because $D(h) \ge \sqrt{v/4} + \epsilon$. This proves (30).

Let ν be the conditional law of g at the fixed history. The measures

$$\mathrm{d}\lambda_-(\ell) = (h - \ell)_+ \,\mathrm{d}\nu(\ell), \qquad \mathrm{d}\lambda_+(u) = (u - h)_+ \,\mathrm{d}\nu(u) \tag{31}$$

have the same mass because $\int z \,\mathrm{d}\nu(z) = h$. If that common mass is zero, then $g = h$ almost surely and the result is immediate. Otherwise, pair the measures in (31) by their normalized product. Giving the pair (ℓ, u) total mixture weight

$$\frac{r+s}{rs}\,\frac{\mathrm{d}\lambda_-(\ell)\,\mathrm{d}\lambda_+(u)}{\lambda_-(\mathbb{R})}$$

and then placing weights $s/(r+s)$ and $r/(r+s)$ on ℓ and u, respectively, reproduces ν away from its possible atom at h. That atom is represented by the degenerate law at h. On one nondegenerate chord, the Jensen loss, its $\mathfrak{p}$-moment mass, and its deposit mass are, respectively,

$$\frac{rs}{r+s}(S_- - S_+), \qquad \frac{rs}{r+s}(r^{\mathfrak{p}-1} + s^{\mathfrak{p}-1}),$$

$$\frac{rs}{r+s}\left\{\frac{|\ell|^{2-\mathfrak{p}}Q(\ell)}{r} + \frac{|u|^{2-\mathfrak{p}}Q(u)}{s}\right\}.$$

After integration against the preceding mixture, the second quantity is exactly $\mathbb{E}\left[|g-h|^{\mathfrak{p}}\right] = m$, while the third is at most $\mathbb{E}\left[|g|^{2-\mathfrak{p}}Q(g)\right]$, with the missing nonnegative contribution coming only from a possible atom at h. Multiplying (30) by $(rs/(r+s))^2$, taking square roots, and applying the inequality for positive parts before Cauchy–Schwarz with respect to the mixing measure gives

$$\sqrt{D(h)}[W(h) - \mathbb{E}[W(g)]]_+ \le C\sqrt{m}\left(\mathbb{E}\left[|g|^{2-\mathfrak{p}}Q(g)\right]\right)^{\frac{1}{2}}.$$

Squaring proves (27), with D and Q as defined in (26). The $\mathfrak{p}$-moment is spent exactly once. □

The constant in Lemma 3 can be chosen independently of $\epsilon \ge 0$. To record its numerical value, put $w(z) = \sqrt{\beta_2 v_{t-1,i} + z^2/T}$. Direct differentiation gives

$$|W''(z)| = \frac{|z|\left(3\beta_2 v_{t-1,i} w(z) + \beta_2 v_{t-1,i}\epsilon + 2\epsilon w(z)^2\right)}{T w(z)^3 (w(z)+\epsilon)^3} \le \frac{3}{\sqrt{T}\, D(z)^2}.$$

In the case of small Z in (30), let $b_0 = \sqrt{\beta_2 v_{t-1,i}} + \epsilon$. Since $D(h) \le 2b_0$,

$$D(h)(S_- - S_+)^2 \le \frac{72Z^2}{Tb_0^3} \le \frac{288}{\sqrt{v}} Q(Z).$$

The product on the right of (30), before its constant, is at least $\frac{Q(Z)}{2}$. In the case of large Z, $D(h)(S_- - S_+)_+^2 \le 4/D(h) \le (16/\sqrt{v})Q(Z)$. The mixture argument preserves the resulting secant constant. Thus the coefficient in (27) can be taken as $\frac{2^{11}}{\sqrt{v}}$.

C.2 Deposits, martingale control, and the stopping argument

Replacing the current random gradient by its conditional mean gives the predictable accumulator

$$\bar{v}_{t,i} = \beta_2 v_{t-1,i} + \frac{(\nabla f(\mathbf{x}_t))_i^2}{T}. \tag{32}$$

We use the potential $L(f - f_*)$ and define the adaptive energy of the true gradients

$$\mathcal{A}_N = \sum_{t=0}^{N-1}\sum_{i=1}^{d} L\gamma \frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{\bar{v}_{t,i}} + \epsilon}. \tag{33}$$

For fixed model parameters and initial gap, with $T \ge 1$ and $\gamma > 0$, define the deterministic energy scale

$$\begin{aligned}\mathcal{E}_T(\gamma,\delta) &= (1 + L\Delta_f)(1 + L\gamma\sqrt{T}\rho_T + L^2\gamma^2 T) \\ &\quad \times \log^2\left(\frac{eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2}{\delta}\right), \qquad 0 < \delta \le 1.\end{aligned} \tag{34}$$

Here ρ_T is defined in (1). The scale $\mathcal{E}_T$ controls both the stopping height and the adaptive energy $\mathcal{A}_T$ in (33); its value at $\delta = 1$ controls the energy moments.

Explicit constants and threshold. The stopping argument uses one coefficient, depending only on $(\mathfrak{p}, d, v, \beta_1)$:

$$\begin{aligned} C_{\mathrm{stop}} = 1 + \left[7 + \log\left(6 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}\right)\right] \Bigg\{ & \\ & \frac{2^{25} d}{\sqrt{v}} \left[16d\left(1 + \log\left(1 + \frac{1}{v}\right)\right) + 1 + \frac{8}{1 - 2^{2-2\mathfrak{p}}}\right] \\ & + 32d\left(1 + \log\left(1 + \frac{1}{v}\right)\right) \\ & + 4\sqrt{d}\left(1 + \sqrt{136 + 128\log\left(1 + \frac{1}{v}\right)}\right)\left(1 + \sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}}\right) \\ & + \frac{128 d^2}{(1-\beta_1)^2}\left[1 + 32d\left(1 + \log\left(1 + \frac{1}{v}\right)\right)\right] \\ & \times \left(3 + 2\sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}}\right)\Bigg\}, \end{aligned} \tag{35}$$

The height and energy tail coefficients are

$$K = 16(1 + C_{\mathrm{stop}})^2, \qquad K_E = 16 C_{\mathrm{stop}}(1 + \sqrt{K}). \tag{36}$$

The corresponding horizon threshold and energy moment coefficient are

$$T_{\mathrm{u}} = \left\lceil \max\left\{10,\ 24 \cdot 2^{12} \cdot 16^{\mathfrak{p}},\ K\right\}\right\rceil, \tag{37}$$

$$C_E = K_E \max_{q \in \{1, 2, \frac{\mathfrak{p}}{\mathfrak{p}-1}\}} \left(\int_0^\infty (1+u)^{2q} e^{-u}\, \mathrm{d}u\right)^{\frac{1}{q}}. \tag{38}$$

The integral is finite for every $\mathfrak{p} > 1$. Neither T_{u} nor C_E depends on L, the base stepsize, damping, noise scale, or initial gap.

Proposition 5 (Uniform stopped energy and moment bounds)**.** *Suppose $T \geq T_{\mathrm{u}}$ and $\gamma > 0$. For $H \geq \max\{1, L\Delta_f\}$ and $0 < \delta < 1$, let τ_H be the stopping time in (20). Use the energy $\mathcal{A}_N$ from (33), its deterministic scale $\mathcal{E}_T$ from (34), C_{stop} from (35), (K, K_E) from (36), and ρ_T from (1). With probability at least $1 - \delta$, simultaneously for all $0 \leq n < T$,*

$$\begin{aligned} & L\big(f(\mathbf{z}_{\min\{n+1,\tau_H\}}) - f_*\big) + \frac{1}{16}\mathcal{A}_{\min\{n+1,\tau_H\}} \\ & \quad \leq L\Delta_f + C_{\mathrm{stop}} L\gamma\sqrt{T}(\rho_T + \sqrt{H} + L\gamma\sqrt{T}) \log\frac{eT(1 + H + \sigma^2)}{\delta}. \end{aligned} \tag{39}$$

The objective height and energy satisfy

$$\mathbb{P}\left(\max_{0 \leq t \leq T}\{L(f(\mathbf{x}_t) - f_*)\} > K\mathcal{E}_T(\gamma, \delta) + \frac{8\beta_1^2 d L^2 \gamma^2 T}{(1-\beta_1)^2} + 2\right) \leq \delta, \tag{40}$$

$$\mathbb{P}(\mathcal{A}_T > K_E \mathcal{E}_T(\gamma, \delta)) \leq \delta. \tag{41}$$

With C_E from (38),

$$\|\mathcal{A}_T\|_{L^q} \leq C_E \mathcal{E}_T(\gamma, 1), \qquad q \in \left\{1, 2, \frac{\mathfrak{p}}{\mathfrak{p}-1}\right\}.$$

All statements hold for every $L > 0$, $\Delta_f \geq 0$, and $\gamma > 0$ with the same constants and horizon threshold.

Proof. First suppose $\sigma > 0$. If $\sigma = 0$, the conditional moment assumption implies $\xi_t = 0$ almost surely at every clock. In that case the conditional biases, martingale differences, and every realized noise square vanish. Set the deposit and its conditional mean (U_t, P_t) in (45) to zero, omit the dyadic split, and use the pathwise EMA bound directly. This gives the asserted estimates with $\rho_T = 1$, as defined in (1).

To estimate the conditional normalization bias, fix $t < T$ and a coordinate i. The accumulator satisfies

$$\beta_2 v_{t-1,i} \geq \beta_2^{t+1} v \geq \beta_2^T v \geq \frac{v}{4},$$

so Lemma 3 applies conditionally with predictable accumulator argument $\beta_2 v_{t-1,i}$. Apply it directly when $(\nabla f(\mathbf{x}_t))_i \geq 0$, and reflect the gradient and noise when $(\nabla f(\mathbf{x}_t))_i < 0$. Using W from (26) and $\bar{v}_{t,i}$ from (32), oddness of W and evenness of the fraction for the current squared gradient give

$$\begin{aligned}
&\left[\frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{\bar{v}_{t,i}}+\epsilon} - \mathbb{E}\left[\frac{(\nabla f(\mathbf{x}_t))_i\, g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right]_+^2 \\
&\quad\leq \frac{2^{11}\,(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{v}\left(\sqrt{\bar{v}_{t,i}}+\epsilon\right)}\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\mid\mathscr{F}_{t-1}\right]\mathbb{E}\left[|g_{t,i}|^{2-\mathfrak{p}}\frac{g_{t,i}^2}{T v_{t,i}}\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned}$$

When $(\nabla f(\mathbf{x}_t))_i \neq 0$, Young's inequality yields

$$\begin{aligned}
&\left[(1-2^{-12})\frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{\bar{v}_{t,i}}+\epsilon} - \mathbb{E}\left[\frac{(\nabla f(\mathbf{x}_t))_i\, g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right]_+ \\
&\quad\leq \frac{2^{21}}{\sqrt{v}}\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\mid\mathscr{F}_{t-1}\right]\mathbb{E}\left[|g_{t,i}|^{2-\mathfrak{p}}\frac{g_{t,i}^2}{T v_{t,i}}\,\middle|\,\mathscr{F}_{t-1}\right].
\end{aligned} \tag{42}$$

If $(\nabla f(\mathbf{x}_t))_i = 0$, the left side is zero, so the same bound holds.

To absorb the bias when the predictable gradient is large, reflection reduces to $h > 0$. Write $m = \mathbb{E}\left[|\xi|^{\mathfrak{p}}\mid\mathscr{F}_{t-1}\right]$. On $|\xi| \leq \frac{h}{2}$, every point between h and $h+\xi$ has denominator at least $\frac{D(h)}{2}$. The numerical derivative bound above, Taylor's formula, and

$$\mathbb{E}\left[|\xi|^2\mathbf{1}_{\{|\xi|\leq\frac{h}{2}\}}\,\middle|\,\mathscr{F}_{t-1}\right] \leq h^{2-\mathfrak{p}}m, \qquad D(h) \geq \frac{h}{\sqrt{T}}$$

give

$$\mathbb{E}\left[|W(h+\xi)-W(h)-W'(h)\xi|\mathbf{1}_{\{|\xi|\leq\frac{h}{2}\}}\,\middle|\,\mathscr{F}_{t-1}\right] \leq 6mh^{-\mathfrak{p}}W(h).$$

On the complementary event, oddness and monotonicity imply

$$|W(h+\xi)| \leq W(h)\left(1+\frac{|\xi|}{h}\right), \qquad hW'(h) \leq W(h).$$

Conditional Markov and the bound

$$\mathbb{E}\left[|\xi|\mathbf{1}_{\{|\xi|>\frac{h}{2}\}}\,\middle|\,\mathscr{F}_{t-1}\right] \leq 2^{\mathfrak{p}-1}h^{1-\mathfrak{p}}m$$

show that the complementary remainder is at most $3\cdot 2^{\mathfrak{p}}mh^{-\mathfrak{p}}W(h) \leq 12mh^{-\mathfrak{p}}W(h)$. Subtracting the centered linear term and combining the two pieces yields

$$|W(h) - \mathbb{E}\left[W(h+\xi)\mid\mathscr{F}_{t-1}\right]| \leq 24mh^{-\mathfrak{p}}W(h).$$

If $|(\nabla f(\mathbf{x}_t))_i| > \sigma T^{1/\mathfrak{p}}/16$, the preceding estimate gives

$$\begin{aligned}
&\left|\frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{\bar{v}_{t,i}}+\epsilon} - \mathbb{E}\left[\frac{(\nabla f(\mathbf{x}_t))_i\, g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right| \\
&\quad\leq \frac{24\cdot 16^{\mathfrak{p}}}{T}\frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{\bar{v}_{t,i}}+\epsilon} \leq 2^{-12}\frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{\bar{v}_{t,i}}+\epsilon},
\end{aligned}$$

provided $T \geq 24\cdot 2^{12}\cdot 16^{\mathfrak{p}}$, as required in (37). Thus the left side of (42) is zero in this range. If $|(\nabla f(\mathbf{x}_t))_i| \leq \sigma T^{1/\mathfrak{p}}/16$, decompose its right side into dyadic bins $2^k < |g_{t,i}| \leq 2^{k+1}$, with

$k \in \mathbb{Z}$. On each bin, $|g_{t,i}|^{2-\mathfrak{p}} \le 2^{1+k(2-\mathfrak{p})}$. When $2^k > \sigma T^{1/\mathfrak{p}}$, the bin also implies $|\xi_{t,i}| > 2^{k-3}$. Using the conditional moment bound gives

$$\begin{aligned}
&\left[(1-2^{-12})\frac{(\nabla f(\mathbf{x}_t))_i^2}{\sqrt{v_{t,i}}+\epsilon} - \mathbb{E}\left[\frac{(\nabla f(\mathbf{x}_t))_i\, g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]\right]_+ \\
&\le \frac{2^{22}}{\sqrt{v}}\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right] \sum_{k=-\infty}^{\lfloor \log_2(\sigma T^{\frac{1}{\mathfrak{p}}})\rfloor} 2^{k(2-\mathfrak{p})} \\
&\quad \times \mathbb{E}\left[\frac{g_{t,i}^2}{T v_{t,i}}\mathbf{1}_{\{2^k<|g_{t,i}|\le 2^{k+1}\}}\,\middle|\,\mathscr{F}_{t-1}\right] \\
&\quad + \frac{2^{22}\sigma^{\mathfrak{p}}}{\sqrt{v}} \sum_{k=\lfloor \log_2(\sigma T^{\frac{1}{\mathfrak{p}}})\rfloor+1}^{\infty} 2^{k(2-\mathfrak{p})}\mathbb{P}\left(|\xi_{t,i}| > 2^{k-3}\,\middle|\,\mathscr{F}_{t-1}\right).
\end{aligned} \tag{43}$$

This holds for every predictable $(\nabla f(\mathbf{x}_t))_i$, including the range in which its left side is zero.

To aggregate (43), we use the moment bound for the full vector

$$\sum_{i=1}^{d}\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right] \le d^{1-\frac{\mathfrak{p}}{2}}\mathbb{E}\left[\|\xi_t\|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right] \le d^{1-\frac{\mathfrak{p}}{2}}\sigma^{\mathfrak{p}}. \tag{44}$$

Collect the realized contributions at low amplitudes and their conditional mean in

$$\begin{aligned}
U_t &= \frac{1}{d^{1-\frac{\mathfrak{p}}{2}}\sigma^2 T^{\frac{2}{\mathfrak{p}}-1}}\sum_{i=1}^{d}\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right] \\
&\quad \times \sum_{k=-\infty}^{\lfloor \log_2(\sigma T^{\frac{1}{\mathfrak{p}}})\rfloor} 2^{k(2-\mathfrak{p})}\frac{g_{t,i}^2}{T v_{t,i}}\mathbf{1}_{\{2^k<|g_{t,i}|\le 2^{k+1}\}}, \\
P_t &= \mathbb{E}\left[U_t\,|\,\mathscr{F}_{t-1}\right].
\end{aligned} \tag{45}$$

For each coordinate at most one dyadic bin is active. Moreover, for every i,

$$\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right] \le \mathbb{E}\left[\|\xi_t\|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right] \le \sigma^{\mathfrak{p}}.$$

Hence (44), $g_{t,i}^2/(Tv_{t,i}) \le 1$, and $2^{k(2-\mathfrak{p})} \le \sigma^{2-\mathfrak{p}}T^{\frac{2}{\mathfrak{p}}-1}$, applied to (45), show that $0 \le U_t \le 1$. They also show pathwise that, for every $N \le T$,

$$\sum_{t=0}^{N-1} U_t \le \sum_{t=0}^{N-1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{T v_{t,i}}.$$

For any adapted $Y_t \in [0,1]$, with $R_t = \mathbb{E}\left[Y_t\,|\,\mathscr{F}_{t-1}\right]$, convexity of $y \mapsto e^{-y}$ on $[0,1]$ gives

$$\mathbb{E}\left[e^{-Y_t}\,\middle|\,\mathscr{F}_{t-1}\right] \le 1-(1-e^{-1})R_t \le e^{-(1-e^{-1})R_t}.$$

$$\exp\left\{(1-e^{-1})\sum_{t=0}^{N-1}R_t - \sum_{t=0}^{N-1}Y_t\right\}, \qquad 0 \le N \le T,$$

is a nonnegative supermartingale starting from one. Stopping it at its first crossing and applying Markov's inequality, then taking $(Y_t, R_t) = (U_t, P_t)$ from (45), gives

$$\mathbb{P}\left(\sum_{t=0}^{N-1} P_t > 4\sum_{t=0}^{N-1} U_t + 8u \text{ for some } N \le T\right) \le e^{-u}. \tag{46}$$

The realized deposits from the current squared gradients satisfy the pathwise EMA potential

$$\sum_{t=0}^{N-1}\sum_{i=1}^{d} \frac{g_{t,i}^2}{T v_{t,i}} \le 2d + \sum_{i=1}^{d} \log\left(1 + \frac{\sum_{t=0}^{N-1} g_{t,i}^2}{Tv}\right). \tag{47}$$

Indeed, the accumulator update gives

$$\frac{g_{t,i}^2}{T v_{t,i}} \le \log \frac{v_{t,i}}{\beta_2 v_{t-1,i}}.$$

Summation telescopes the accumulator logarithms, $-N\log\beta_2 \le 2$, and $v_{N-1,i} \le v + T^{-1}\sum_{t=0}^{N-1} g_{t,i}^2$, proving (47). If every update $t < N$ starts below height H, (22) gives $\|\nabla f(\mathbf{x}_t)\|^2 \le 2H$. Together with $(h+\xi)^2 \le 2h^2 + 2\xi^2$, this yields

$$\sum_{t=0}^{N-1}\sum_{i=1}^{d} \frac{g_{t,i}^2}{T v_{t,i}} \le 2d + d\log\left(1 + \frac{4H}{v} + \frac{2}{Tv}\sum_{s=0}^{T-1}\|\xi_s\|^2\right). \tag{48}$$

The expected average squared noise need not be finite when $\mathfrak{p} < 2$, but subadditivity at exponent $\frac{\mathfrak{p}}{2} \le 1$, including equality when $\mathfrak{p} = 2$, followed by the tower property, gives

$$\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=0}^{T-1}\|\xi_t\|^2\right)^{\frac{\mathfrak{p}}{2}}\right] \le T^{-\frac{\mathfrak{p}}{2}}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\xi_t\|^{\mathfrak{p}}\right] \le \sigma^{\mathfrak{p}} T^{1-\frac{\mathfrak{p}}{2}}. \tag{49}$$

Fix $H \ge \max\{1, L\Delta_f\}$. By Markov's inequality, outside an event of probability at most $\frac{\delta}{16}$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\xi_t\|^2 \le \sigma^2 T^{\frac{2}{\mathfrak{p}}-1}\left(\frac{16}{\delta}\right)^{\frac{2}{\mathfrak{p}}}.$$

Since $T \ge 10$,

$$\log\frac{16}{\delta} \le \log\frac{eT(1+H+\sigma^2)}{\delta},$$
$$\log\left(1 + \frac{1}{T}\sum_{t=0}^{T-1}\|\xi_t\|^2\right) \le 3\log\frac{eT(1+H+\sigma^2)}{\delta}.$$

Using

$$1 + \frac{4H}{v} + \frac{2}{Tv}\sum_{t=0}^{T-1}\|\xi_t\|^2 \le \left(1+\frac{4}{v}\right)(1+H)\left(1 + \frac{1}{T}\sum_{t=0}^{T-1}\|\xi_t\|^2\right)$$

in (48) gives, simultaneously for all prefixes whose updates start below height H,

$$\begin{aligned}\sum_{t=0}^{N-1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{Tv_{t,i}} &\le d\left(6 + \log\left(1+\frac{4}{v}\right)\right)\log\frac{eT(1+H+\sigma^2)}{\delta} \\ &\le 32d\left(1+\log\left(1+\frac{1}{v}\right)\right)\log\frac{eT(1+H+\sigma^2)}{\delta}.\end{aligned} \tag{50}$$

Apply (46) to the deposit U_t in (45) and to $d^{-1}\sum_{i=1}^{d} g_{t,i}^2/(Tv_{t,i})$, each with $u = \log(16/\delta)$. The three events jointly fail with probability at most $\frac{3\delta}{16}$. On their intersection, simultaneously for all

such stopped prefixes,

$$\sum_{t=0}^{N-1} P_t \le 8\left[16d\left(1+\log\left(1+\frac{1}{v}\right)\right)+1\right]\log\frac{eT(1+H+\sigma^2)}{\delta},$$
$$\sum_{t=0}^{N-1}\sum_{i=1}^{d}\mathbb{E}\left[\frac{g_{t,i}^2}{Tv_{t,i}}\,\middle|\,\mathscr{F}_{t-1}\right] \le 8d\left[17+16\log\left(1+\frac{1}{v}\right)\right]\log\frac{eT(1+H+\sigma^2)}{\delta}.$$

By the definition of P_t in (45), the sum over the low bins in (43), before its outer coefficient, is exactly

$$d^{1-\frac{\mathfrak{p}}{2}}\sigma^2 T^{\frac{2}{\mathfrak{p}}-1}\sum_{t=0}^{N-1} P_t$$
$$\le 8d\sigma^2 T^{\frac{2}{\mathfrak{p}}-1}\left[16d\left(1+\log\left(1+\frac{1}{v}\right)\right)+1\right]\log\frac{eT(1+H+\sigma^2)}{\delta}.$$

For the high bins, conditional Markov and the budget for the full vector give

$$\sum_{i=1}^{d} 2^{k(2-\mathfrak{p})}\mathbb{P}\big(|\xi_{t,i}|>2^{k-3}\,\big|\,\mathscr{F}_{t-1}\big) \le 2^{3\mathfrak{p}+k(2-2\mathfrak{p})}\sum_{i=1}^{d}\mathbb{E}\left[|\xi_{t,i}|^{\mathfrak{p}}\,|\,\mathscr{F}_{t-1}\right]$$
$$\le d\sigma^{\mathfrak{p}}2^{6+k(2-2\mathfrak{p})}.$$

The geometric sum converges because $\mathfrak{p}>1$. For $\sigma>0$,

$$\sum_{k=\lfloor\log_2(\sigma T^{\frac{1}{\mathfrak{p}}})\rfloor+1}^{\infty} 2^{k(2-2\mathfrak{p})} \le \frac{T^{\frac{2}{\mathfrak{p}}-2}}{\sigma^{2\mathfrak{p}-2}(1-2^{2-2\mathfrak{p}})}.$$

After summing over at most T clocks, the scale factors for the low and high bins are controlled by ρ_T from (1):

$$\frac{\sigma^2 T^{\frac{2}{\mathfrak{p}}-1}}{\sqrt{T}} = \sigma^{2\mathfrak{p}}\sqrt{T}\frac{T^{\frac{2}{\mathfrak{p}}-2}}{\sigma^{2\mathfrak{p}-2}} = \sigma^2 T^{\frac{2}{\mathfrak{p}}-\frac{3}{2}} \le \rho_T.$$

Since $\log(eT(1+H+\sigma^2)/\delta)\ge 1$, summing (43) and multiplying by $L\gamma$ gives the accumulated bound on the conditional bias, in terms of $\mathcal{A}_N$ from (33):

$$(1-2^{-12})\mathcal{A}_N - L\gamma\sum_{t=0}^{N-1}\mathbb{E}\left[\sum_{i=1}^{d}\frac{(\nabla f(\mathbf{x}_t))_i\, g_{t,i}}{\sqrt{v_{t,i}}+\epsilon}\,\middle|\,\mathscr{F}_{t-1}\right]$$
$$\le \frac{2^{25}dL\gamma\sqrt{T}\rho_T}{\sqrt{v}}\log\frac{eT(1+H+\sigma^2)}{\delta} \tag{51}$$
$$\times\left[16d\left(1+\log\left(1+\frac{1}{v}\right)\right)+1+\frac{8}{1-2^{2-2\mathfrak{p}}}\right].$$

On the same event, this holds for every stopped prefix, including large predictable gradients, for which the positive remainder is zero. The dyadic bins indexed by $k\in\mathbb{Z}$ cover all positive sample magnitudes for every $\sigma>0$.

Let Z_t be the normalized gradient alignment, including the potential scale L, and let Δ_t be its centered fluctuation:

$$Z_t = L\gamma\sum_{i=1}^{d}\frac{(\nabla f(\mathbf{x}_t))_i\, g_{t,i}}{\sqrt{\beta_2 v_{t-1,i}+\frac{g_{t,i}^2}{T}}+\epsilon}, \tag{52}$$
$$\Delta_t = \mathbb{E}\left[Z_t\,|\,\mathscr{F}_{t-1}\right]-Z_t.$$

The stopping rule (20) makes $\{t<\tau_H\}$ $\mathscr{F}_{t-1}$-measurable, so the centered increment Δ_t in (52), multiplied by $\mathbf{1}_{\{t<\tau_H\}}$, is a martingale difference. On this stopped region, current normalization, Cauchy–Schwarz over coordinates, and (22) give

$$|Z_t| \le L\gamma\sqrt{2dHT}, \qquad |\Delta_t| \le 2L\gamma\sqrt{2dHT},$$
$$\mathbb{E}\left[\Delta_t^2 \,\middle|\, \mathscr{F}_{t-1}\right] \le 2L^2\gamma^2 TH \sum_{i=1}^{d} \mathbb{E}\left[\frac{g_{t,i}^2}{T v_{t,i}} \,\middle|\, \mathscr{F}_{t-1}\right]. \tag{53}$$

If a conditionally centered variable X satisfies $|X| \le R$ and $V_X = \mathbb{E}\left[X^2 \,\middle|\, \mathscr{F}_{t-1}\right]$, expansion of the exponential gives, for $0 \le \lambda < \frac{3}{R}$,

$$\mathbb{E}\left[e^{\lambda X} \,\middle|\, \mathscr{F}_{t-1}\right] \le \exp\left\{\frac{\lambda^2 V_X}{2(1-\frac{\lambda R}{3})}\right\}.$$

Iteration, Markov's inequality, optimization in λ, and application to both signs yield

$$\mathbb{P}\left(\left|\sum_{t=0}^{n-1} X_t\right| > b,\ \sum_{t=0}^{n-1} \mathbb{E}\left[X_t^2 \,\middle|\, \mathscr{F}_{t-1}\right] \le F\right) \le 2\exp\left(-\frac{b^2}{2F + \frac{2Rb}{3}}\right). \tag{54}$$

On the event where the predictable deposits are bounded for all prefixes, the variance and increment bounds in (53) are

$$F = 16dL^2\gamma^2 TH\left[17 + 16\log\left(1+\frac{1}{v}\right)\right]\log\frac{eT(1+H+\sigma^2)}{\delta},$$
$$R = 2L\gamma\sqrt{2dHT}.$$

Apply (54) to $X_t = \mathbf{1}_{\{t<\tau_H\}}\Delta_t$. For $N = \min\{n+1, \tau_H\}$, its prefix sum at the deterministic endpoint $n+1$ is exactly $\sum_{t=0}^{N-1}\Delta_t$. Take

$$w = \log\frac{8T}{\delta} \le 2\log\frac{eT(1+H+\sigma^2)}{\delta}.$$

The deviation

$$\sqrt{2Fw} + \frac{2Rw}{3} \le 4L\gamma\sqrt{dTH}\left(1+\sqrt{136+128\log\left(1+\frac{1}{v}\right)}\right) \times \log\frac{eT(1+H+\sigma^2)}{\delta}$$

makes the right side of (54) at most $2e^{-w} = \frac{\delta}{4T}$. A union bound over the T endpoints therefore adds failure probability at most $\frac{\delta}{4}$. Thus, with total failure probability at most $\frac{7\delta}{16}$,

$$\left|\sum_{t=0}^{N-1}\Delta_t\right| \le 4L\gamma\sqrt{dTH}\left(1+\sqrt{136+128\log\left(1+\frac{1}{v}\right)}\right) \times \log\frac{eT(1+H+\sigma^2)}{\delta} \tag{55}$$

simultaneously for all stopped prefixes. The conditional variance is controlled through the bounded normalized scalar.

For the stopped endpoint $N = \min\{n+1, \tau_H\}$, every update $t < N$ starts below height H, including the update that may produce the first crossing. The centering identity (52) gives

$$Z_t = \mathbb{E}\left[Z_t \,|\, \mathscr{F}_{t-1}\right] - \Delta_t.$$

Using (21) at time $n+1$, the bound on normalized squares, and $L(f(\mathbf{z}_0)-f_*) = L(f(\mathbf{x}_0)-f_*) = L\Delta_f$, we obtain

$$\begin{aligned} &L\big(f(\mathbf{z}_N)-f_*\big) \\ &\quad\le L\Delta_f - \sum_{t=0}^{N-1}\mathbb{E}\left[Z_t \,|\, \mathscr{F}_{t-1}\right] + \sum_{t=0}^{N-1}\Delta_t + L^2\gamma^2\sum_{t=0}^{N-1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{v_{t,i}} \\ &\qquad + \frac{128d^2}{(1-\beta_1)^2}(L\gamma\sqrt{T(H+1)} + L^2\gamma^2 T) \\ &\qquad \times\left(1+\sum_{t=0}^{N-1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{Tv_{t,i}}\right). \end{aligned}$$

On the joint event above, (50) and $\log(eT(1+H+\sigma^2)/\delta)\ge 1$ give

$$1+\sum_{t=0}^{N-1}\sum_{i=1}^{d}\frac{g_{t,i}^2}{Tv_{t,i}} \le \left[1+32d\left(1+\log\left(1+\frac{1}{v}\right)\right)\right]\log\frac{eT(1+H+\sigma^2)}{\delta}.$$

Combining this with $\sqrt{H+1}\le 2\sqrt{H}$, (51), and (55), using ρ_T from (1), gives

$$\begin{aligned} L\big(f(\mathbf{z}_N)-f_*\big) + (1-2^{-12})\mathcal{A}_N \le L\Delta_f + L\gamma\sqrt{T}\log\frac{eT(1+H+\sigma^2)}{\delta}\Bigg\{ \\ \frac{2^{25}d\rho_T}{\sqrt{v}}\left[16d\left(1+\log\left(1+\frac{1}{v}\right)\right)+1+\frac{8}{1-2^{2-2\mathfrak{p}}}\right] \\ + 4\sqrt{dH}\left(1+\sqrt{136+128\log\left(1+\frac{1}{v}\right)}\right) \\ + 32dL\gamma\sqrt{T}\left(1+\log\left(1+\frac{1}{v}\right)\right) \\ + \frac{128d^2}{(1-\beta_1)^2}\left[1+32d\left(1+\log\left(1+\frac{1}{v}\right)\right)\right]\left(2\sqrt{H}+L\gamma\sqrt{T}\right)\Bigg\}. \end{aligned} \tag{56}$$

The definition of C_{stop} in (35) dominates each coefficient of ρ_T, $\sqrt{H}$, and $L\gamma\sqrt{T}$ in (56). Since $1-2^{-12}\ge 1/16$ and the joint failure probability is at most $7\delta/16<\delta$, this proves (39). At $\sigma=0$, the pathwise deposit and momentum bounds give (56) directly.

The displacement identity and normalized caps give $\|\mathbf{z}_t-\mathbf{x}_t\|\le \frac{2\beta_1\gamma\sqrt{dT}}{1-\beta_1}$. Smoothness and the gradient bound imply that, whenever $\|\mathbf{u}-\mathbf{w}\|\le\frac{2\beta_1\gamma\sqrt{dT}}{1-\beta_1}$,

$$L(f(\mathbf{u})-f_*)\le 2L(f(\mathbf{w})-f_*)+\frac{4\beta_1^2dL^2\gamma^2T}{(1-\beta_1)^2}. \tag{57}$$

Indeed, the descent inequality bounds the left side by

$$L(f(\mathbf{w})-f_*)+\frac{2\beta_1L\gamma\sqrt{2dTL(f(\mathbf{w})-f_*)}}{1-\beta_1}+\frac{2\beta_1^2dL^2\gamma^2T}{(1-\beta_1)^2},$$

and Young's inequality gives the display.

Using the energy scale $\mathcal{E}_T$ in (34), set

$$H_\delta = K\mathcal{E}_T(\gamma,\delta)+\frac{8\beta_1^2dL^2\gamma^2T}{(1-\beta_1)^2}+2. \tag{58}$$

By (36), $K\ge 1$. Since the logarithm is at least one, the height (58) is at least $\max\{1, L\Delta_f\}$. To compare its logarithm uniformly in the initial gap and base stepsize, the profile ρ_T in (1) gives, for $T\ge 1$,

$$\begin{aligned} \rho_T &\le (1+\sigma^2)\sqrt{T}, \\ (1+L\Delta_f)(1+L\gamma\sqrt{T}\rho_T+L^2\gamma^2T) &\le (1+L\Delta_f)(1+\sigma^2)(1+L\gamma\sqrt{T})^2\sqrt{T} \\ &\le (2+L\Delta_f+\sigma^2)^2(1+L\gamma\sqrt{T})^2\sqrt{T}. \end{aligned}$$

Since $T \geq K$, $L^2\gamma^2 T \leq (1 + L\Delta_f)(1 + L\gamma\sqrt{T}\rho_T + L^2\gamma^2 T)$, and $\log\big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2/\delta\big) \geq 1$,

$$1 + H_\delta + \sigma^2 \leq \left(6 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}\right)(2 + L\Delta_f + \sigma^2)^2(1 + L\gamma\sqrt{T})^2 T^{\frac{3}{2}} \times \left[\log \frac{eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2}{\delta}\right]^2.$$

Taking logarithms and using $\log u \leq u$ for $u > 0$ gives

$$\log \frac{eT(1 + H_\delta + \sigma^2)}{\delta} \leq \left[7 + \log\left(6 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}\right)\right] \times \log \frac{eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2}{\delta}.$$

Indeed, apart from the logarithm of the square and $\log\left(6 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}\right)$, the terms are at most $\frac{5}{2}\log\big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2/\delta\big)$. The logarithm of the square is at most twice the same logarithm, which gives the stated coefficient. In particular, the comparison constant does not depend on K, Δ_f, γ, or σ.

The height definition (58) and the scale (34) give

$$\sqrt{H_\delta} \leq \left(\sqrt{K} + \sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}}\right)\sqrt{\mathcal{E}_T(\gamma, \delta)}.$$

Indeed, $\mathcal{E}_T(\gamma, \delta) \geq 1 + L^2\gamma^2 T$. The same definition gives the bounds

$$(1 + L\Delta_f + L\gamma\sqrt{T}\rho_T + L^2\gamma^2 T)\log^2\left(\frac{eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2}{\delta}\right) \leq \mathcal{E}_T(\gamma, \delta),$$

$$L\gamma\sqrt{T}\sqrt{\mathcal{E}_T(\gamma, \delta)}\log\frac{eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2}{\delta} \leq \mathcal{E}_T(\gamma, \delta).$$

For the second inequality, square both sides and use $L^2\gamma^2 T \leq (1 + L\Delta_f)(1 + L\gamma\sqrt{T}\rho_T + L^2\gamma^2 T)$. Together with the logarithm comparison above, these inequalities bound the right side of (56) at $H = H_\delta$ by $C_{\mathrm{stop}}(1 + \sqrt{K})\mathcal{E}_T(\gamma, \delta)$. The coefficient follows directly from (35) and

$$\sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}} + \sqrt{K} \leq \left(1 + \sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}}\right)(1 + \sqrt{K}),$$

$$1 + 2\sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}} + 2\sqrt{K} \leq \left(3 + 2\sqrt{2 + \frac{8\beta_1^2 d}{(1-\beta_1)^2}}\right)(1 + \sqrt{K}).$$

Thus, on the same event as (56),

$$L\big(f(\mathbf{z}_N) - f_*\big) + \frac{1}{16}\mathcal{A}_N \leq C_{\mathrm{stop}}(1 + \sqrt{K})\mathcal{E}_T(\gamma, \delta), \qquad N = \min\{n + 1, \tau_{H_\delta}\}.$$

If the stopping time (20) at height (58) satisfies $\tau_{H_\delta} \leq T$, then (57) at the crossing implies

$$L\big(f(\mathbf{z}_{\tau_{H_\delta}}) - f_*\big) > \frac{H_\delta}{2} - \frac{2\beta_1^2 d L^2\gamma^2 T}{(1-\beta_1)^2} > \frac{K}{2}\mathcal{E}_T(\gamma, \delta).$$

Apply the stopped estimate, which holds for all prefixes, at $n = T - 1$. Its left side is at least the preceding quantity, whereas its right side has the upper bound just proved. The choice of K in (36) makes these bounds incompatible because

$$\frac{K}{2} - C_{\mathrm{stop}}(1 + \sqrt{K}) = 8 + 11C_{\mathrm{stop}} + 4C_{\mathrm{stop}}^2 > 0.$$

There is consequently no crossing on the event where the stopped estimate holds, which proves (40). Taking the endpoint $N = T$ on that same event and dropping $L(f(\mathbf{z}_T) - f_*) \geq 0$ proves (41), with exactly $K_E = 16C_{\mathrm{stop}}(1 + \sqrt{K})$, as in (36).

For the energy $\mathcal{A}_T$ in (33), we pass from the tail to moments using the deterministic scale (34). Its baseline logarithm is at least one, so

$$\mathcal{E}_T(\gamma, e^{-u}) \leq \mathcal{E}_T(\gamma, 1)(1 + u)^2, \qquad u \geq 0.$$

Setting $\delta = e^{-u}$ in (41) therefore gives

$$\mathbb{P}\big(\mathcal{A}_T > K_E \mathcal{E}_T(\gamma, 1)(1 + u)^2\big) \leq e^{-u}, \qquad u \geq 0,$$

where the bound at $u = 0$ is automatic. Equivalently, $\mathcal{A}_T$ is stochastically dominated by $K_E \mathcal{E}_T(\gamma, 1)(1 + E)^2$, where E has density e^{-u} on $[0, \infty)$. Layer-cake integration gives

$$\begin{aligned} \|\mathcal{A}_T\|_{L^q} &\leq K_E \mathcal{E}_T(\gamma, 1) \left(\int_0^\infty (1 + u)^{2q} e^{-u} \, \mathrm{d}u \right)^{\frac{1}{q}} \\ &\leq C_E \mathcal{E}_T(\gamma, 1) \end{aligned}$$

for each of the three exponents specified in (38). □

C.3 Removing preconditioning

Lemma 4 (Conversion to ordinary squared gradient)**.** *For the gradient energy $\mathcal{A}_T$ in (33), pathwise,*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\mathcal{A}_T}{L\gamma T} \left(\epsilon + \sqrt{dv + \frac{2}{T} \sum_{s=0}^{T-1} \|\xi_s\|^2} \right) + \frac{4\mathcal{A}_T^2}{L^2 \gamma^2 T^2}. \tag{59}$$

Proof. Unrolling the second moment recursion, dropping its weights, and using $(h+\xi)^2 \leq 2h^2 + 2\xi^2$ in the predictable accumulator (32) gives

$$\sum_{i=1}^{d} \bar{v}_{t,i} \leq dv + \frac{1}{T} \sum_{s=0}^{T-1} \left(3\|\nabla f(\mathbf{x}_s)\|^2 + 2\|\xi_s\|^2 \right).$$

The energy definition (33) therefore yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{\mathcal{A}_T}{L\gamma T} \left(\epsilon + \sqrt{dv + \frac{1}{T} \sum_{s=0}^{T-1} \left(3\|\nabla f(\mathbf{x}_s)\|^2 + 2\|\xi_s\|^2 \right)} \right).$$

Separating the gradient contribution from the square root and applying Young's inequality gives

$$\frac{\sqrt{3}\mathcal{A}_T}{L\gamma T} \sqrt{\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2} \leq \frac{1}{2T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{3\mathcal{A}_T^2}{2L^2\gamma^2 T^2}.$$

Combining these inequalities yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2\mathcal{A}_T}{L\gamma T} \left(\epsilon + \sqrt{dv + \frac{2}{T} \sum_{s=0}^{T-1} \|\xi_s\|^2} \right) + \frac{3\mathcal{A}_T^2}{L^2 \gamma^2 T^2}.$$

Enlarging the quadratic coefficient from three to four proves (59). □

Proof of Theorem 4. For every $L > 0$, $\Delta_f \geq 0$, $T \geq T_{\mathrm{u}}$, and $\gamma > 0$, Proposition 5, with $\mathcal{A}_T$ from (33), $\mathcal{E}_T$ from (34), and C_E from (38), gives

$$\|\mathcal{A}_T\|_{L^q} \leq C_E \mathcal{E}_T(\gamma, 1), \qquad q \in \left\{ 1, 2, \frac{\mathfrak{p}}{\mathfrak{p} - 1} \right\}.$$

The noise moment bound (49) implies

$$\left\| \left(\frac{1}{T} \sum_{t=0}^{T-1} \|\xi_t\|^2 \right)^{\frac{1}{2}} \right\|_{L^{\mathfrak{p}}} = \left(\mathbb{E}\left[\left(\frac{1}{T} \sum_{t=0}^{T-1} \|\xi_t\|^2 \right)^{\frac{\mathfrak{p}}{2}} \right] \right)^{\frac{1}{\mathfrak{p}}} \le \sigma T^{\frac{1}{\mathfrak{p}} - \frac{1}{2}}.$$

Take expectations in (59), and use

$$\sqrt{dv + \frac{2}{T} \sum_{t=0}^{T-1} \|\xi_t\|^2} \le \sqrt{dv} + \sqrt{2} \left(\frac{1}{T} \sum_{t=0}^{T-1} \|\xi_t\|^2 \right)^{\frac{1}{2}}.$$

Hölder's inequality with conjugate exponents $\mathfrak{p}/(\mathfrak{p}-1)$ and $\mathfrak{p}$ gives

$$\mathbb{E}\left[\mathcal{A}_T \left(\frac{1}{T} \sum_{t=0}^{T-1} \|\xi_t\|^2 \right)^{\frac{1}{2}} \right] \le C_E \mathcal{E}_T(\gamma, 1) \sigma T^{\frac{1}{\mathfrak{p}} - \frac{1}{2}}.$$

The first and second energy moments, after expanding the scale (34) with ρ_T from (1), yield

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] &\le \frac{2C_E(1 + L\Delta_f)}{\sqrt{T}} \left(\epsilon + \sqrt{dv} + \sqrt{2}\sigma T^{\frac{1}{\mathfrak{p}} - \frac{1}{2}} \right) \left(\rho_T + L\gamma\sqrt{T} + \frac{1}{L\gamma\sqrt{T}} \right) \\ &\quad \times \left[\log\Big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2 \Big) \right]^2 \\ &\quad + \frac{4C_E^2(1 + L\Delta_f)^2}{T} \left(\rho_T + L\gamma\sqrt{T} + \frac{1}{L\gamma\sqrt{T}} \right)^2 \left[\log\Big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2 \Big) \right]^4. \end{aligned}$$

This proves (25), with T_{u} and C_E specified in (37) and (38). For fixed parameters, the inequality

$$\begin{aligned} &\log\big(eT(2 + L\Delta_f + \sigma^2)(1 + L\gamma\sqrt{T})^2 \big) \\ &\qquad \le [2 + \log(2 + L\Delta_f + \sigma^2)] \log\big(eT(1 + L\gamma\sqrt{T}) \big) \end{aligned}$$

holds uniformly over $\gamma > 0$. Absorbing the fixed coefficients into $\mathcal{O}$ yields (4) and proves Theorem 1. □

D Formal Statements and Proofs of Proposition 2 and Corollary 1

This section proves Proposition 2 and Corollary 1. Theorem 4 supplies (25) for every $\gamma > 0$ with the threshold T_{u} in (37). Throughout this section, $T \ge T_{\mathrm{u}}$ and the common setup of Section 3.2, with the moment coefficients from Section 2.2, applies. The parameters $L > 0$ and $\Delta_f \ge 0$ are held fixed when optimizing over γ.

Proposition 6 (Formal statement of Proposition 2)**.** *For every $\Delta_f, \sigma \ge 0$ and integer $T \ge T_{\mathrm{u}}$, there exists $\eta^* = \eta^*(T, \mathfrak{p}, L, \sigma, \Delta_f) \in (0, \frac{1}{L})$ such that $\gamma_T^* = \eta^*/\sqrt{T}$ minimizes the right side of (25) over $\gamma > 0$. With this stepsize, uniformly over the instances in Theorem 4,*

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] &\le C_*(1 + L\Delta_f)^2(1 + \sigma)(2 + \sigma^2)^2 \\ &\quad \times \log^2\big(eT(2 + L\Delta_f + \sigma^2) \big) \left(1 + \frac{\log^2\big(eT(2 + L\Delta_f + \sigma^2) \big)}{\sqrt{T}} \right) T^{r_*(\mathfrak{p})}, \end{aligned} \tag{60}$$

where $C_* > 0$ is given in (67) and depends only on $(\mathfrak{p}, d, v, \epsilon, \beta_1)$, and the exponent $r_*(\mathfrak{p})$ is defined in (1). For every such T, a constant base stepsize depending only on the model parameters, T, and Δ_f can be chosen so that, uniformly over the same instances, the expected average squared gradient norm is at most the smaller of $2L\Delta_f + 1$ and the right side of (60). For fixed model parameters and initial gap, this gives (5).

For this proof, let $U_T(\gamma; \Delta_f)$ denote the right side of (25), with C_E from (38) and ρ_T from (1). At each base stepsize $\gamma > 0$, Theorem 4 bounds the expected average squared gradient norm by $U_T(\gamma; \Delta_f)$ for every admissible objective and oracle, with constants and a threshold independent of the instance and of γ.

D.1 UNIQUE GLOBAL MINIMUM OF THE EXPLICIT UPPER BOUND

The map

$$u \longmapsto \frac{2C_E(1+L\Delta_f)}{\sqrt{T}}\left(\epsilon+\sqrt{dv}+\sqrt{2}\sigma T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}\right)u+\frac{4C_E^2(1+L\Delta_f)^2}{T}u^2$$

is strictly increasing on $[0,\infty)$, because both coefficients are positive. With ρ_T defined in (1), it suffices to minimize

$$\left(\rho_T+L\gamma\sqrt{T}+\frac{1}{L\gamma\sqrt{T}}\right)\left[\log\left(eT(2+L\Delta_f+\sigma^2)(1+L\gamma\sqrt{T})^2\right)\right]^2$$

over $\gamma>0$. Differentiation isolates the factor P that determines the derivative's sign:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\gamma}\left\{\left(\rho_T+L\gamma\sqrt{T}+\frac{1}{L\gamma\sqrt{T}}\right)\left[\log\left(eT(2+L\Delta_f+\sigma^2)(1+L\gamma\sqrt{T})^2\right)\right]^2\right\}\\
&\qquad=\frac{\log\left(eT(2+L\Delta_f+\sigma^2)(1+L\gamma\sqrt{T})^2\right)}{L\gamma^2\sqrt{T}}\,P(\gamma),\\
&P(\gamma)=(L^2\gamma^2T-1)\log\left(eT(2+L\Delta_f+\sigma^2)(1+L\gamma\sqrt{T})^2\right)\\
&\qquad+\frac{4L\gamma\sqrt{T}(L^2\gamma^2T+\rho_TL\gamma\sqrt{T}+1)}{1+L\gamma\sqrt{T}}.
\end{aligned}\tag{61}$$

The function P extends continuously to $\left[0,1/(L\sqrt{T})\right]$, with

$$P(0)=-\log(eT(2+L\Delta_f+\sigma^2))<0,\qquad P\left(\frac{1}{L\sqrt{T}}\right)=2(\rho_T+2)>0.$$

For $0<\gamma<1/(L\sqrt{T})$, the quotient rule and $\rho_T\geq 1$ give

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\gamma}\frac{L\gamma\sqrt{T}(L^2\gamma^2T+\rho_TL\gamma\sqrt{T}+1)}{1+L\gamma\sqrt{T}}\\
&\qquad=L\sqrt{T}\,\frac{1+2\rho_TL\gamma\sqrt{T}+(3+\rho_T)L^2\gamma^2T+2L^3\gamma^3T^{\frac{3}{2}}}{(1+L\gamma\sqrt{T})^2}\geq L\sqrt{T}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
P'(\gamma)&\geq 2L^2\gamma T\log\left(eT(2+L\Delta_f+\sigma^2)(1+L\gamma\sqrt{T})^2\right)\\
&\quad+2L\sqrt{T}(L\gamma\sqrt{T}-1)+4L\sqrt{T}\\
&=2L^2\gamma T\left[\log\left(eT(2+L\Delta_f+\sigma^2)(1+L\gamma\sqrt{T})^2\right)+1\right]+2L\sqrt{T}>0.
\end{aligned}$$

The intermediate value theorem and strict monotonicity give exactly one root $\gamma_T^*\in\left(0,1/(L\sqrt{T})\right)$ of P in (61). On $(0,\gamma_T^*)$, $P<0$, and on $\left(\gamma_T^*,1/(L\sqrt{T})\right)$, $P>0$. For every $\gamma\geq 1/(L\sqrt{T})$, the first term in $P(\gamma)$ is nonnegative and the second is strictly positive. The prefactor of $P(\gamma)$ in (61) is positive. The displayed factor, and therefore $U_T(\cdot;\Delta_f)$, decreases strictly up to γ_T^* and increases strictly thereafter. This proves a unique global minimum over all $\gamma>0$. The factor also tends to infinity as $\gamma\downarrow 0$ and as $\gamma\to\infty$. In particular, since $\gamma_T^*<1/(L\sqrt{T})$, strict increase to the right of the minimizer gives

$$U_T(\gamma_T^*;\Delta_f)<U_T\left(\frac{1}{L\sqrt{T}};\Delta_f\right).$$

Define the rescaled minimizing stepsize

$$\eta^*=\eta^*(T,\mathfrak{p},L,\sigma,\Delta_f):=\gamma_T^*\sqrt{T}\in\left(0,\frac{1}{L}\right).\tag{62}$$

Dividing $P(\gamma_T^*) = 0$ in (61) by $(L\eta^*)^2$ gives its characterization by a unique root, with ρ_T from (1):

$$\left(1 - \frac{1}{(L\eta^*)^2}\right) \log\bigl(eT(2 + L\Delta_f + \sigma^2)(1 + L\eta^*)^2\bigr) + \frac{4\left(\rho_T + L\eta^* + \frac{1}{L\eta^*}\right)}{1 + L\eta^*} = 0. \tag{63}$$

Thus the bound U_T defined by (25) satisfies

$$\inf_{\gamma>0} U_T(\gamma; \Delta_f) = U_T\left(\frac{\eta^*}{\sqrt{T}}; \Delta_f\right).$$

The root characterization (63) shows that η^* depends only on $(T, \mathfrak{p}, L, \sigma, \Delta_f)$. The positive coefficients in $U_T(\cdot; \Delta_f)$ do not change the minimizer because their combination is strictly increasing in the displayed factor.

For fixed L, Δ_f, and σ, if $\mathfrak{p} \geq \frac{4}{3}$ or $\sigma = 0$, then ρ_T in (1) equals the constant $1 + \sigma^2$. For every fixed $\theta \in (0, 1)$, (61) gives

$$P\left(\frac{\theta}{L\sqrt{T}}\right) = (\theta^2 - 1)\log\bigl(eT(2 + L\Delta_f + \sigma^2)(1 + \theta)^2\bigr) + \frac{4\theta(\theta^2 + \rho_T\theta + 1)}{1 + \theta} \longrightarrow -\infty.$$

Strict monotonicity gives $\gamma_T^* > \theta/(L\sqrt{T})$, and hence $L\eta^* > \theta$, for all sufficiently large T. Together with $L\eta^* < 1$, this proves $\eta^* \to \frac{1}{L}$. If $1 < \mathfrak{p} < \frac{4}{3}$ and L, Δ_f, and $\sigma > 0$ are fixed, then $\rho_T = 1 + \sigma^2 T^{2/\mathfrak{p}-3/2}$ grows faster than $\log(eT(2 + L\Delta_f + \sigma^2))$. For every fixed $\theta \in (0, 1)$, the second term in the displayed expression for $P\bigl(\theta/(L\sqrt{T})\bigr)$ dominates its negative first term. Thus $L\eta^* < \theta$ eventually and $L\eta^* \to 0$. The root equation (63) can be written as

$$\rho_T(L\eta^*)^2 + L\eta^* + (L\eta^*)^3 = \frac{(1 - L\eta^*)(1 + L\eta^*)^2}{4} \log\bigl(eT(2 + L\Delta_f + \sigma^2)(1 + L\eta^*)^2\bigr).$$

Since

$$\log\bigl(eT(2 + L\Delta_f + \sigma^2)(1 + L\eta^*)^2\bigr) = \log(eT(2 + L\Delta_f + \sigma^2)) + 2\log(1 + L\eta^*),$$

dividing the preceding root identity by $\log(eT(2 + L\Delta_f + \sigma^2))$, using $L\eta^* \to 0$, gives

$$(L\eta^*)^2 \sim \frac{\log(eT(2 + L\Delta_f + \sigma^2))}{4\rho_T}.$$

These asymptotics characterize the stepsize that minimizes the explicit upper bound $U_T(\cdot; \Delta_f)$.

D.2 Stationarity rates

Use ρ_T and $r_*(\mathfrak{p})$ from (1). The proof in Section D.1 minimizes the profile itself, with $\gamma_T^* = \eta^*/\sqrt{T}$ as in (62), so its value at γ_T^* is at most its value at $\gamma = 1/(L\sqrt{T})$. Since $\rho_T + 2 \leq 3\rho_T$ and $\log(eT(2 + L\Delta_f + \sigma^2)) \geq 1$,

$$\begin{aligned}
&\left(\rho_T + L\eta^* + \frac{1}{L\eta^*}\right)\left[\log\bigl(eT(2 + L\Delta_f + \sigma^2)(1 + L\eta^*)^2\bigr)\right]^2 \\
&\quad \leq (\rho_T + 2)\log^2\bigl(4eT(2 + L\Delta_f + \sigma^2)\bigr) \\
&\quad \leq 3(1 + \log 4)^2\rho_T \log^2\bigl(eT(2 + L\Delta_f + \sigma^2)\bigr).
\end{aligned} \tag{64}$$

Since $\rho_T \leq (1 + \sigma^2)T^{(2/\mathfrak{p}-3/2)_+}$ by (1), substituting (64) into the bound $U_T(\cdot; \Delta_f)$ defined by (25) yields, for every admissible objective and oracle at $\gamma = \gamma_T^*$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\bigl[\|\nabla f(\mathbf{x}_t)\|^2\bigr] \leq U_T(\gamma_T^*; \Delta_f) \\
&\leq 6C_E(1 + \log 4)^2(1 + L\Delta_f)(1 + \sigma^2)\log^2\bigl(eT(2 + L\Delta_f + \sigma^2)\bigr) \\
&\quad \times \left[(\epsilon + \sqrt{dv})T^{\left(\frac{2}{\mathfrak{p}} - \frac{3}{2}\right)_+ - \frac{1}{2}} + \sqrt{2}\sigma T^{\left(\frac{2}{\mathfrak{p}} - \frac{3}{2}\right)_+ + \frac{1}{\mathfrak{p}} - 1}\right] \\
&\quad + 36C_E^2(1 + \log 4)^4(1 + L\Delta_f)^2(1 + \sigma^2)^2 \log^4\bigl(eT(2 + L\Delta_f + \sigma^2)\bigr) T^{\left(\frac{4}{\mathfrak{p}} - 3\right)_+ - 1}.
\end{aligned} \tag{65}$$

For $1 < \mathfrak{p} < \frac{4}{3}$, the three powers of T in this bound have exponents

$$\frac{2}{\mathfrak{p}} - 2, \qquad \frac{3}{\mathfrak{p}} - \frac{5}{2}, \qquad \frac{4}{\mathfrak{p}} - 4.$$

The second minus the first equals $\frac{1}{\mathfrak{p}} - \frac{1}{2} \geq 0$, and the second minus the third equals $\frac{3}{2} - \frac{1}{\mathfrak{p}} > 0$. For $\frac{4}{3} \leq \mathfrak{p} \leq 2$, the exponents are $-\frac{1}{2}, \frac{1}{\mathfrak{p}} - 1, -1$; their maximum is $\frac{1}{\mathfrak{p}} - 1$. At $\mathfrak{p} = \frac{4}{3}$, both formulas give $-\frac{1}{4}$, and at $\mathfrak{p} = 2$ the first two exponents coincide at $-\frac{1}{2}$. In particular, $(2/\mathfrak{p} - 3/2)_+ + 1/\mathfrak{p} - 1 = r_*(\mathfrak{p})$ throughout $1 < \mathfrak{p} \leq 2$, with r_* from (1). For the quadratic term, the separation of exponents is

$$r_*(\mathfrak{p}) + 1 - \left(\frac{4}{\mathfrak{p}} - 3\right)_+ = \begin{cases} \dfrac{3}{2} - \dfrac{1}{\mathfrak{p}}, & 1 < \mathfrak{p} \leq \frac{4}{3}, \\ \dfrac{1}{\mathfrak{p}}, & \frac{4}{3} \leq \mathfrak{p} \leq 2, \end{cases} \geq \frac{1}{2}. \tag{66}$$

Thus the first two powers are bounded by $T^{r_*(\mathfrak{p})}$, and the third by $T^{r_*(\mathfrak{p})-1/2}$. Take

$$C_* = 6C_E(1 + \log 4)^2\big(\epsilon + \sqrt{dv} + \sqrt{2}\big) + 36C_E^2(1 + \log 4)^4. \tag{67}$$

Using $1 + \sigma^2 \leq 2 + \sigma^2$, $1 + L\Delta_f \leq (1 + L\Delta_f)^2$, and $\sigma \leq 1 + \sigma$, gives

$$\begin{aligned} U_T(\gamma_T^*; \Delta_f) &\leq C_*(1 + L\Delta_f)^2(1 + \sigma)(2 + \sigma^2)^2 \\ &\quad \times \log^2\big(eT(2 + L\Delta_f + \sigma^2)\big)\left(1 + \frac{\log^2\big(eT(2 + L\Delta_f + \sigma^2)\big)}{\sqrt{T}}\right) T^{r_*(\mathfrak{p})}, \end{aligned}$$

which proves (60). The constant C_* depends only on $(\mathfrak{p}, d, v, \epsilon, \beta_1)$, and the estimate holds for every $T \geq T_{\mathrm{u}}$.

To determine the horizon dependence with fixed parameters, for $T \geq 1$ we have

$$\log T = \int_1^T \frac{1}{u}\, \mathrm{d}u \leq \int_1^T u^{-\frac{3}{4}}\, \mathrm{d}u = 4\left(T^{\frac{1}{4}} - 1\right).$$

Hence $\log(eT) \leq 4T^{1/4}$, which gives

$$\frac{[\log(eT)]^2}{\sqrt{T}} \leq 16.$$

Furthermore,

$$\begin{aligned} \log\big(eT(2 + L\Delta_f + \sigma^2)\big) &= \log(eT) + \log(2 + L\Delta_f + \sigma^2) \\ &\leq [1 + \log(2 + L\Delta_f + \sigma^2)]\log(eT). \end{aligned}$$

Combining the two inequalities gives, for every $T \geq 1$,

$$\begin{aligned} &\log^2\big(eT(2 + L\Delta_f + \sigma^2)\big)\left(1 + \frac{\log^2\big(eT(2 + L\Delta_f + \sigma^2)\big)}{\sqrt{T}}\right) \\ &\leq [1 + \log(2 + L\Delta_f + \sigma^2)]^2\big(1 + 16[1 + \log(2 + L\Delta_f + \sigma^2)]^2\big)[\log(eT)]^2. \end{aligned} \tag{68}$$

For fixed model parameters and initial gap, this gives the following rate, with $r_*(\mathfrak{p})$ defined in (1):

$$U_T(\gamma_T^*; \Delta_f) = \mathcal{O}\Big([\log(eT)]^2 T^{r_*(\mathfrak{p})}\Big).$$

For fixed model parameters and initial gap with $\sigma > 0$, the squared logarithm also gives the exact order of the explicit bound U_T from (25). With ρ_T and $r_*(\mathfrak{p})$ defined in (1), for every $\gamma > 0$ its profile is at least $\rho_T \log^2(eT(2 + L\Delta_f + \sigma^2))$; its linear noise term and $\rho_T \geq \sigma^2 T^{(2/\mathfrak{p}-3/2)_+}$ give

$$\begin{aligned} U_T(\gamma; \Delta_f) &\geq 2\sqrt{2}C_E(1 + L\Delta_f)\sigma T^{\frac{1}{\mathfrak{p}}-1}\rho_T \log^2\big(eT(2 + L\Delta_f + \sigma^2)\big) \\ &\geq 2\sqrt{2}C_E(1 + L\Delta_f)\sigma^3 T^{r_*(\mathfrak{p})}[\log(eT)]^2. \end{aligned}$$

Together with the upper estimate, this proves

$$\min_{\gamma > 0} U_T(\gamma; \Delta_f) = \Theta\Big([\log(eT)]^2 T^{r_*(\mathfrak{p})}\Big) \qquad (\sigma > 0 \text{ fixed}).$$

This establishes the exact logarithmic order of the optimized expression U_T for the upper bound.

D.3 Deterministic update control and the minimax upper bound

For every $T \geq 2$ and every $\gamma > 0$, the accumulator recursion gives $v_{t,i} \geq g_{t,i}^2/T$ and $v_{t,i} \geq \beta_2^{t-k} v_{k,i}$ for $k \leq t < T$. The effective stepsizes $\lambda_{t,i}$ from (9) therefore satisfy

$$|\lambda_{t,i} g_{t,i}| \leq \gamma\sqrt{T}, \qquad \lambda_{t,i} \leq 2\lambda_{k,i}.$$

Expanding $m_{t,i} = (1-\beta_1)\sum_{k=0}^{t} \beta_1^{t-k} g_{k,i}$ shows

$$|\lambda_{t,i} m_{t,i}| \leq 2\gamma\sqrt{T}(1-\beta_1)\sum_{k=0}^{t} \beta_1^{t-k} \leq 2\gamma\sqrt{T}.$$

Summation over coordinates and over updates yields, on every history,

$$\|\mathbf{x}_t - \mathbf{x}_0\| \leq 2\gamma t\sqrt{dT}, \qquad 0 \leq t \leq T.$$

L-smoothness gives $f(\mathbf{x} - \nabla f(\mathbf{x})/L) \leq f(\mathbf{x}) - \|\nabla f(\mathbf{x})\|^2/(2L)$, so $\|\nabla f(\mathbf{x}_0)\|^2 \leq 2L\Delta_f$. The gradient Lipschitz bound implies, simultaneously for all histories,

$$\begin{aligned}\|\nabla f(\mathbf{x}_t)\| &\leq \|\nabla f(\mathbf{x}_0)\| + L\|\mathbf{x}_t - \mathbf{x}_0\| \\ &\leq \sqrt{2L\Delta_f} + 2L\gamma t\sqrt{dT} \leq \sqrt{2L\Delta_f} + 2L\gamma(T-1)\sqrt{dT}, \qquad t < T.\end{aligned}$$

It follows, for every admissible objective and oracle at the stated base stepsize γ, that

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] \leq \big(\sqrt{2L\Delta_f} + 2L\gamma(T-1)\sqrt{dT}\big)^2. \tag{69}$$

Indeed, the same deterministic bound holds for the average squared gradient norm on every history, and hence also bounds its expectation. For every tolerance $\delta > 0$ and $T \geq 2$, the positive stepsize

$$\gamma_\delta = \frac{\sqrt{2L\Delta_f + \delta} - \sqrt{2L\Delta_f}}{2L(T-1)\sqrt{dT}} \tag{70}$$

depends only on $(L, d, T, \Delta_f, \delta)$ and makes the right side of (69) equal to $2L\Delta_f + \delta$. When $T = 1$, every positive stepsize gives $\|\nabla f(\mathbf{x}_0)\|^2 \leq 2L\Delta_f$. Thus the bound $2L\Delta_f$ can be approached arbitrarily closely by a positive stepsize chosen uniformly for all admissible instances, for every $T \geq 1$.

For $T \geq T_{\mathrm{u}}$ and any $\delta > 0$, use $\gamma_T^* = \eta^*/\sqrt{T}$ from (62), γ_δ from (70), and the bound U_T defined by (25) to choose

$$\gamma = \begin{cases} \gamma_T^*, & U_T(\gamma_T^*; \Delta_f) \leq 2L\Delta_f, \\ \gamma_\delta, & U_T(\gamma_T^*; \Delta_f) > 2L\Delta_f. \end{cases}$$

Both choices depend only on the fixed parameters, T, Δ_f, and δ. In the first case, Theorem 4 bounds the expected average squared gradient norm by $U_T(\gamma_T^*; \Delta_f)$; in the second case, (69) bounds it by $2L\Delta_f + \delta$. Consequently, at this single stepsize, every admissible objective and oracle satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] \leq \min\{2L\Delta_f, U_T(\gamma_T^*; \Delta_f)\} + \delta.$$

Using the bound on $U_T(\gamma_T^*; \Delta_f)$ established in the proof of (60), together with (68) and $\delta = T^{-1}$, gives, for fixed model parameters and initial gap,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] = \mathcal{O}\Big(\min\Big\{1, [\log(eT)]^2 T^{r_*(\mathfrak{p})}\Big\}\Big),$$

because the exponent in (1) satisfies $r_*(\mathfrak{p}) \geq -(\mathfrak{p}-1)/\mathfrak{p} \geq -\frac{1}{2}$, so the term T^{-1} is absorbed by the displayed bound. The stepsize and the implied constant are independent of the particular objective and oracle policy.

Taking $\delta = 1$ and choosing γ_T^* when $U_T(\gamma_T^*; \Delta_f) \leq 2L\Delta_f + 1$, and γ_1 from (70) otherwise, gives a single stepsize that is independent of the instance and satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] \leq \min\{2L\Delta_f + 1, U_T(\gamma_T^*; \Delta_f)\},$$

which proves the alternative for finite horizons in Proposition 6 by (60).

This proves (5) and completes Proposition 6 and the order statement in Proposition 2.

D.4 Proof of Corollary 1

Corollary 2 (Formal statement of Corollary 1)**.** *Under the conditions of Theorem 4, taking* $\gamma = 1/(L\sqrt{T})$ *gives* (60) *for every* $T \geq T_{\mathrm{u}}$*, with the same constant* C_* *from* (67)*. For fixed model parameters and initial gap, this stepsize also gives* (6)*, with* $r_*(\mathfrak{p})$ *defined in* (1)*, uniformly over the admissible instances.*

Proof. For $\gamma = 1/(L\sqrt{T})$, one has $L\gamma\sqrt{T} = 1$. Substitution into (25) gives

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|^2\big] &\leq \frac{2C_E(1+L\Delta_f)}{\sqrt{T}}\left(\epsilon + \sqrt{dv} + \sqrt{2}\sigma T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}\right)(\rho_T+2)\\ &\quad\times \log^2\big(4eT(2+L\Delta_f+\sigma^2)\big)\\ &\quad + \frac{4C_E^2(1+L\Delta_f)^2}{T}(\rho_T+2)^2\log^4\big(4eT(2+L\Delta_f+\sigma^2)\big).\end{aligned}$$

Using ρ_T from (1) and $T \geq T_{\mathrm{u}} \geq 1$,

$$\rho_T + 2 \leq 3(1+\sigma^2)T^{\left(\frac{2}{\mathfrak{p}}-\frac{3}{2}\right)_+}, \qquad \log\big(eT(2+L\Delta_f+\sigma^2)\big) \geq 1.$$

Therefore

$$\begin{aligned}\log\big(4eT(2+L\Delta_f+\sigma^2)\big) &= \log\big(eT(2+L\Delta_f+\sigma^2)\big) + \log 4\\ &\leq (1+\log 4)\log\big(eT(2+L\Delta_f+\sigma^2)\big).\end{aligned}$$

Substituting these two bounds gives the right side of (65), whose three powers of T are bounded in Section D.2. In particular, (66) gives an extra factor $T^{-1/2}$ for the term containing $\log^4(eT(2+L\Delta_f+\sigma^2))$. The same bounds on the noise and gap factors therefore give (60) with C_* from (67). Applying (68) proves (6) with the exponent $r_*(\mathfrak{p})$ from (1). This exponent is negative exactly when $\mathfrak{p} > \frac{6}{5}$; for $1 < \mathfrak{p} \leq \frac{6}{5}$, the guarantee of constant order instead follows from the stepsize selection in Section D.3. This proves Corollary 2 and Corollary 1. □

E Formal Statement and Proof of the Constructive Lower Bound

Take $L = 1$. Fix $1 < \mathfrak{p} \leq 2$, $v > 0$, and $\epsilon \geq 0$, and use the moment coefficients specified in Section 2.2. The base stepsize $\gamma > 0$ is arbitrary. Run Algorithm 1 from $x_0 = 0$, $m_{-1} = 0$, and $v_{-1} = v$. The stochastic constructions use a fixed $\sigma > 0$; their constants and horizon thresholds depend only on $(\mathfrak{p}, v, \epsilon, \beta_1, \sigma)$, unless another dependence is stated. These parameters are independent of (T, γ). The notation $\asymp$ denotes upper and lower bounds up to positive constants with this dependence. The deterministic plateau and gap completion arguments also hold when $\sigma = 0$. Parameters introduced within a construction are local to that lemma.

We first record the scalar path bounds, then give explicit objectives and noise distributions. Their formulas verify nonnegativity and continuous differentiability, derivatives with Lipschitz constant one, and conditionally centered noise with $\mathfrak{p}$-moment at most $\sigma^{\mathfrak{p}}$, as required by Assumptions 1–3. The final step completes the initial gap to one and embeds the scalar construction into the first coordinate of any fixed dimension.

E.1 Scalar path bounds and geometric flushing

For $T \geq 2$, put

$$\lambda_t = \frac{\gamma}{\sqrt{v_t}+\epsilon} \qquad (-1 \leq t < T), \qquad v_{-1} = v. \tag{71}$$

Unrolling the two moment recursions gives

$$m_t = (1-\beta_1)\sum_{k=0}^{t}\beta_1^{t-k}g_k, \qquad v_t = \beta_2^{t+1}v + \frac{1}{T}\sum_{k=0}^{t}\beta_2^{t-k}g_k^2.$$

Since $\beta_2^T \geq \frac{1}{4}$,

$$v_t \geq \frac{v}{4}, \qquad \sqrt{v_t} + \epsilon \geq \epsilon + \frac{\sqrt{v}}{2} > 0. \tag{72}$$

For $-1 \leq k \leq h < T$, the same expansion gives $v_h \geq \beta_2^{h-k} v_k$, hence $\lambda_h \leq 2\lambda_k$. Also $v_t \geq g_t^2/T$. Combining these facts with the nonnegative geometric weights in m_t gives

$$\lambda_h \leq 2\lambda_k \quad (-1 \leq k \leq h < T), \qquad |\lambda_t g_t| \leq \gamma\sqrt{T}, \qquad |\lambda_t m_t| \leq 2\gamma\sqrt{T}. \tag{73}$$

In particular, every history satisfies

$$|x_t| \leq 2\gamma T\sqrt{T} \qquad (0 \leq t \leq T). \tag{74}$$

When $T = 1$, this last bound follows directly from $v_0 = g_0^2$, $m_0 = (1-\beta_1)g_0$, and the convention for zero denominators in Algorithm 1. All stochastic horizon thresholds below are chosen to be at least two.

Define

$$J_{\beta_1}(T) = \left\lceil \frac{12\log(eT)}{1-\beta_1} \right\rceil. \tag{75}$$

Then $\beta_1^{J_{\beta_1}(T)} \leq (eT)^{-12}$, including $\beta_1 = 0$. If $g_s = \cdots = g_{s+J-1} = 0$, then $m_{s+j} = \beta_1^{j+1} m_{s-1}$, and (73) gives, for every such prefix before clock T,

$$\sum_{j=0}^{J-1} |\lambda_{s+j} m_{s+j}| \leq \frac{2}{1-\beta_1} |\lambda_{s-1} m_{s-1}|, \qquad |m_{s+J-1}| = \beta_1^J |m_{s-1}|.$$

The same geometric sum bounds the remaining movement under continued zero gradients by a constant depending only on β_1 times $\beta_1^J |\lambda_{s-1} m_{s-1}|$. Thus logarithmically many clocks with zero gradients remove the inherited momentum while preserving any fixed positive spatial margin, or any margin larger than this bound.

E.2 Deterministic gap contribution and exact gap completion

Lemma 5 (Deterministic plateau lower bound)**.** *For every integer $T \geq 1$ and every $\gamma > 0$, there is a scalar, nonnegative, continuously differentiable objective with a derivative of Lipschitz constant one and initial gap $\Delta_f = 1$, together with the noiseless oracle, such that Adam in Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[|f'(x_t)|^2\right] = \min\left\{1, \frac{\epsilon + \frac{\sqrt{v}}{2}}{2\gamma T}\right\}. \tag{76}$$

The same conclusion holds in every fixed dimension $d \geq 1$.

Proof. Define

$$a^2 = \min\left\{1, \frac{\epsilon + \frac{\sqrt{v}}{2}}{2\gamma T}\right\}, \qquad a > 0, \qquad \ell = \frac{1 - \frac{a^2}{2}}{a}. \tag{77}$$

In particular, $a^2 \leq 1$ and $\ell \geq 1/(2a) > 0$. Set

$$f(x) = \begin{cases} 0, & x \leq -\ell - a, \\ \frac{1}{2}(x + \ell + a)^2, & -\ell - a \leq x \leq -\ell, \\ a(x+\ell) + \frac{1}{2}a^2, & x \geq -\ell. \end{cases} \tag{78}$$

Figure 4 illustrates the construction and the interval containing the observed iterates.

The values and first derivatives agree at both junctions. Its derivative is the clipped affine function $f'(x) = \min\{a, \max\{0, x + \ell + a\}\}$, and therefore has Lipschitz constant one. Moreover, $f \geq 0$, $\inf f = 0$, and $f(0) = a\ell + a^2/2 = 1$. These properties verify Assumptions 1 and 2. Assumption 3 follows by taking $\xi_t \equiv 0$, since

$$\mathbb{E}\left[\xi_t \,|\, \mathscr{F}_{t-1}\right] = 0, \qquad \mathbb{E}\left[|\xi_t|^{\mathfrak{p}} \,|\, \mathscr{F}_{t-1}\right] = 0 \leq \sigma^{\mathfrak{p}}.$$

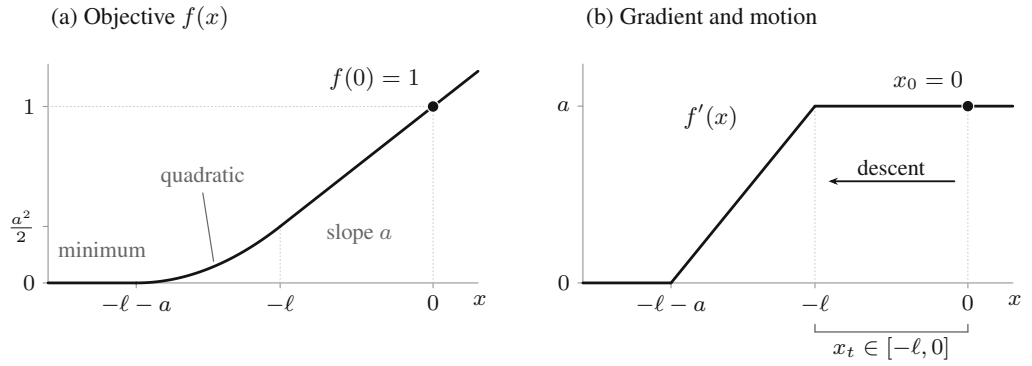


Figure 4: The scalar hard instance in Lemma 5, shown for $a = \frac{4}{5}$ and $\ell = \frac{17}{20}$. Left: a quadratic transition joins the flat minimum to an affine region. Right: the derivative equals a throughout $[-\ell, 0]$. The arrow indicates the direction of noiseless Adam updates; all observed squared derivatives equal a^2, while $f(0) - \inf f = 1$.

The parameter choice in (77) also gives

$$\frac{\gamma a T}{\epsilon + \frac{\sqrt{v}}{2}} \le \frac{1}{2a} \le \ell. \tag{79}$$

Indeed, the first inequality is equivalent to $a^2 \le (\epsilon + \sqrt{v}/2)/(2\gamma T)$, which follows directly from the definition of a, in both branches of the minimum.

If $T = 1$, the only observed derivative is $f'(0) = a$, so the claimed equality holds. For the objective in (78) and $T \ge 2$, we claim that every update start x_t, $0 \le t < T$, lies in $[-\ell, 0]$, where $f'(x_t) = a$. This follows by induction. As long as all earlier update starts lie in this interval, their gradients equal a, so

$$m_t = a(1 - \beta_1^{t+1}) \in [0, a].$$

By (72), the denominator is at least $\epsilon + \sqrt{v}/2$, and hence

$$0 \le x_t - x_{t+1} = \frac{\gamma m_t}{\sqrt{v_t} + \epsilon} \le \frac{\gamma a}{\epsilon + \frac{\sqrt{v}}{2}}.$$

Starting from $x_0 = 0$, summation and (79) give

$$-\ell \le -\frac{\gamma a(t+1)}{\epsilon + \frac{\sqrt{v}}{2}} \le x_{t+1} \le 0 \qquad (0 \le t < T).$$

This closes the induction, including the last updated iterate. Every observed derivative consequently equals a, proving (76) exactly.

For $d \ge 1$, define $F(x_1, \dots, x_d) = f(x_1)$ and use $\xi_t \equiv 0$. Then

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| = |f'(x_1) - f'(y_1)| \le |x_1 - y_1| \le \|\mathbf{x} - \mathbf{y}\|.$$

The initial gap remains 1. All coordinates other than the first remain zero, and the first coordinate follows the scalar recursion exactly. Thus

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla F(\mathbf{x}_t)\|^2 = a^2,$$

proving (76) in every fixed dimension. □

Lemma 6 (Prescribing the initial gap outside a scalar trajectory region)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable and bounded below, with a derivative of Lipschitz constant one and initial gap $\Delta_f \le 1$ at $x_0 = 0$, as defined in Section 2.3. For every finite $A < 0$, there is a nonnegative, continuously differentiable objective $\widetilde{f}$ with a derivative of Lipschitz constant one such that*

$$\widetilde{f}'(x) = f'(x) \quad \textit{for all } x \ge A, \qquad \widetilde{f}(0) - \inf \widetilde{f} = 1.$$

Proof. Put $m = \inf f$, $r = f'(A)$, $B = A - |r|$, and

$$q = f(0) - 1 \le m, \qquad V = f(A) - \frac{1}{2} r|r|.$$

Smoothness and boundedness below give $f(A) - r^2/2 \ge m$, by applying the descent inequality at $A - r$. Therefore $V \ge m$: for $r \ge 0$ this is the preceding inequality, and for $r < 0$ it follows from $V = f(A) + r^2/2$.

Keep f unchanged on $[A, \infty)$. If $r \neq 0$, extend its derivative to $[B, A]$ by

$$u(x) = \frac{r}{|r|}(x - B).$$

This derivative starts at zero, ends at r, and has Lipschitz constant one. The corresponding function value at B is V, and all values on $[B, A]$ are at least m. If $r = 0$, this extension interval is empty and $B = A$, $V = f(A)$.

Since $V - q \ge 0$, put $b = \sqrt{V - q}$. To the left of B, join the constant value q to V using a nonnegative triangular derivative on $[B-2b, B]$: it increases linearly from zero to b on $[B-2b, B-b]$ and decreases linearly to zero on $[B - b, B]$. Its area is $b^2 = V - q$, and both slopes have absolute value one. When $b = 0$, the triangle is empty. Set the function equal to q on the remaining left half-line.

The resulting function F equals f on $[A, \infty)$. Its derivative is continuous at every junction and globally Lipschitz with constant one. Its global minimum is q, attained on its constant left tail. Thus $\widetilde{f} = F - q$ is nonnegative and has

$$\widetilde{f}(0) - \inf \widetilde{f} = f(0) - q = 1.$$

On $[A, \infty)$, the two objectives have identical derivatives. Using the same noise policy, induction over the updates therefore gives identical gradients, first and second moment states, and iterates on every history that remained in this region for the original instance. This also preserves the centered conditional noise law and its moment budget. □

Remark 1. By (74), a deterministic choice $A < -2\gamma T\sqrt{T}$ lies strictly to the left of every scalar trajectory. Lemma 6 therefore completes any auxiliary objective with initial gap at most one to initial gap exactly one, without changing the complete finite horizon process or its conditional noise budget.

E.3 Two piecewise linear derivative profiles

For $h_0 > 0$, $\ell \le r$, $\mu > 0$, and $b \in \mathbb{R}$, define the following profiles. These definitions are independent of $\mathfrak{p}$.

$$\mathrm{Trap}_{h_0;\ell,r}(y) = \begin{cases} 0, & y \le \ell - h_0, \\ y - \ell + h_0, & \ell - h_0 \le y \le \ell, \\ h_0, & \ell \le y \le r, \\ r + h_0 - y, & r \le y \le r + h_0, \\ 0, & y \ge r + h_0, \end{cases} \tag{80}$$

$$\mathrm{Tri}_{\mu;b}(y) = \begin{cases} 0, & y \le b, \\ y - b, & b \le y \le b + \mu, \\ 2\mu - (y - b), & b + \mu \le y \le b + 2\mu, \\ 0, & y \ge b + 2\mu. \end{cases} \tag{81}$$

E.4 Occupation after rare shocks

Lemma 7 (Uniform occupation after rare shocks). *Fix $1 < \mathfrak{p} \le 2$ and $\sigma > 0$. There are $c, T_0 > 0$, depending only on $(\mathfrak{p}, v, \epsilon, \beta_1, \sigma)$, such that for every $T \ge T_0$ and every $\gamma > 0$, a scalar admissible*

instance satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[|f'(x_t)|^2\right] \geq c\left[\gamma^2 + \min\left\{\gamma^2 T,\, \gamma T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}},\, T^{\frac{2}{\mathfrak{p}}-1}\right\}\right]. \tag{82}$$

Proof. Take $f(x) = x^2/2$, put

$$a_\sigma = \min\{1, \sigma^{\mathfrak{p}}\}, \qquad q = \frac{a_\sigma}{16T}, \qquad M = (16T)^{\frac{1}{\mathfrak{p}}}, \tag{83}$$

and use independent rows

$$\mathbb{P}(\xi_t = M) = \mathbb{P}(\xi_t = -M) = \frac{q}{2}, \qquad \mathbb{P}(\xi_t = 0) = 1 - q. \tag{84}$$

The objective is nonnegative with derivative $f'(x) = x$, so Assumptions 1 and 2 hold. Its initial gap is zero. Independence of the rows gives

$$\mathbb{E}\left[\xi_t \,|\, \mathscr{F}_{t-1}\right] = 0, \qquad \mathbb{E}\left[|\xi_t|^{\mathfrak{p}} \,|\, \mathscr{F}_{t-1}\right] = qM^{\mathfrak{p}} = a_\sigma \leq \sigma^{\mathfrak{p}},$$

verifying Assumption 3. Let E_τ be the event that $+M$ is the unique nonzero atom and occurs at $\tau < \lfloor T/4 \rfloor$. The events are disjoint and

$$\sum_{\tau=0}^{\lfloor \frac{T}{4} \rfloor - 1} \mathbb{P}(E_\tau) = \lfloor \frac{T}{4} \rfloor \frac{q}{2}(1-q)^{T-1} \geq 2^{-10} a_\sigma. \tag{85}$$

On the event E_τ in (85), which contains a single shock, the iterate and first moment state are zero before the shock and

$$v_\tau = \beta_2^{\tau+1} v + \frac{M^2}{T}, \qquad m_\tau = (1-\beta_1)M.$$

Since $M^2/T \geq 16$,

$$c\gamma\sqrt{T} \leq |x_{\tau+1}| = \frac{(1-\beta_1)\gamma M}{\sqrt{\beta_2^{\tau+1} v + \frac{M^2}{T}} + \epsilon} \leq C\gamma\sqrt{T}.$$

The clock $\tau + 1 < T$ therefore contributes at least $c\gamma^2$ to the average squared derivative on the event.

It remains to prove the occupation term. Suppose first that $\gamma\sqrt{T} \leq c_0 M$. Fix an integer J_0, depending only on β_1, with $J_0 \geq 1$ and $\beta_1^{J_0+1} \leq 1/4$, and fix a sufficiently large constant B, depending only on the fixed problem parameters. We then choose $c_0, c_1 > 0$ sufficiently small, in that order after B, and put

$$n = \left\lfloor c_1 \min\left\{T, \frac{M}{\gamma\sqrt{T}}\right\}\right\rfloor. \tag{86}$$

We also take $c_1 \leq 1/4$, so that all clocks below remain before T whenever $\tau < T/4$. If $n < 2J_0$, the bound from one clock dominates the desired occupation term after reducing c. We may therefore assume $n \geq 2J_0$. We claim that, for all $1 \leq r \leq n+1$,

$$|x_{\tau+r}| \leq B\gamma\sqrt{T}, \tag{87}$$

where B is the constant fixed above.

Indeed, on a prefix satisfying (87), the exact accumulator recursion gives, for $0 \leq j \leq n$,

$$v_{\tau+j} = \beta_2^j v_\tau + \frac{1}{T}\sum_{r=1}^{j} \beta_2^{j-r} x_{\tau+r}^2.$$

Here $j \leq c_1 T$, and

$$\frac{1}{T}\sum_{r=1}^{j} x_{\tau+r}^2 \leq B^2\gamma^2 j \leq CB^2 c_1 \frac{M^2}{T}.$$

Using the occupation length in (86), the last inequality follows from either $j \le c_1 T$ and $\gamma \le M/T^{3/2}$, or $j \le c_1 M/(\gamma\sqrt{T})$ and $\gamma\sqrt{T} \le c_0 M$. With B fixed first, the stipulated small choices of c_0, c_1, together with (71), give

$$c\frac{M^2}{T} \le v_{\tau+j} \le C\frac{M^2}{T}, \qquad c\frac{\gamma\sqrt{T}}{M} \le \lambda_{\tau+j} \le C\frac{\gamma\sqrt{T}}{M}. \tag{88}$$

For $j \ge 0$, the exact momentum expansion is

$$m_{\tau+j} = (1-\beta_1)\beta_1^j M + (1-\beta_1)\sum_{r=1}^{j} \beta_1^{j-r} x_{\tau+r}.$$

Summing the updates and substituting this identity yields

$$\begin{aligned} x_{\tau+j+1} = &-(1-\beta_1)M\sum_{k=0}^{j}\beta_1^k\lambda_{\tau+k} \\ &-(1-\beta_1)\sum_{k=0}^{j}\lambda_{\tau+k}\sum_{r=1}^{k}\beta_1^{k-r}x_{\tau+r}. \end{aligned}$$

By (88), the first line has magnitude between $c\gamma\sqrt{T}(1-\beta_1^{j+1})$ and $C\gamma\sqrt{T}$. Reversing the two finite sums bounds the second line by

$$C\frac{\gamma\sqrt{T}}{M}\sum_{r=1}^{j}|x_{\tau+r}| \le CBc_1\gamma\sqrt{T}.$$

The prior choice of B, followed by a further reduction of c_0, c_1 if necessary, makes the last constant smaller than one quarter of the lower constant for the main term and makes B larger than twice the upper constant for that term. Considering the first possible failure closes (87); the same calculation also gives

$$|x_{\tau+j+1}| \ge c\gamma\sqrt{T} \qquad (J_0 \le j \le n).$$

Summing over the occupation length in (86) gives, on every E_τ,

$$\frac{1}{T}\sum_{t=0}^{T-1}|f'(x_t)|^2 \ge c\gamma^2 + c\gamma^2\min\left\{T, \frac{M}{\gamma\sqrt{T}}\right\}.$$

If $\gamma\sqrt{T} > c_0 M$, the contribution from one clock, $c\gamma^2$, dominates cM^2/T. In this regime both $\gamma^2 T$ and $\gamma M/\sqrt{T}$ are at least constant multiples of M^2/T, so the minimum of three terms in (82) is itself bounded by a constant multiple of M^2/T. Thus the contribution from one clock proves the required pathwise bound. For the row law (84), the parameter choice (83) gives $0 < a_\sigma \le 1$, hence $q \le 1/(16T)$ and $(1-q)^{T-1} \ge 1-(T-1)q \ge 15/16$. Thus (85) bounds the total success probability below by a fixed positive multiple of a_σ. Multiplying the pathwise bounds by that probability and using $M = (16T)^{1/\mathfrak{p}}$ from (83) gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[|f'(x_t)|^2\right] \ge ca_\sigma\left[\gamma^2 + \min\left\{\gamma^2 T, \gamma T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}, T^{\frac{2}{\mathfrak{p}}-1}\right\}\right].$$

The trajectory constants are independent of σ, and absorbing its fixed positive factor a_σ proves (82). Finally, the normalized update cap makes the average squared derivative finite on every complementary history. □

E.5 A TUNABLE RAMP WITH MEMORY FROM RARE SHOCKS

Lemma 8 (Ramp with memory from rare shocks). *Fix $1 < \mathfrak{p} \le 2$ and $\sigma > 0$. There are $c, C, T_0 > 0$, depending only on $(\mathfrak{p}, v, \epsilon, \beta_1, \sigma)$, such that for every $T \ge T_0$, whenever*

$$\frac{C\log(eT)}{T} \le \gamma \le \frac{c}{\sqrt{T}},$$

there is a scalar admissible instance satisfying

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[|f'(x_t)|^2\right] \ge c\min\left\{\frac{1}{\log(eT)}, \frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma T}\right\}. \tag{89}$$

Proof. At $\mathfrak{p} = 2$, Lemma 5 gives the claim after reducing its constant. Indeed, $\epsilon + \sqrt{v}/2 > 0$, and $\min\{1, ab\} \geq \min\{1, a\} \min\{1, b\}$ for $a, b > 0$; also

$$\min\left\{1, \frac{1}{\gamma T}\right\} \geq \min\left\{\frac{1}{\log(eT)}, \frac{1}{\gamma T}\right\}.$$

We therefore prove the ramp construction for $1 < \mathfrak{p} < 2$. Use the triangular profile in (81) to set

$$u(x) = -\operatorname{Tri}_{\frac{1}{8};0}(x), \qquad f(x) = \frac{1}{64} + \int_0^x u(s)\, \mathrm{d}s. \tag{90}$$

Thus $f = 1/64$ on the left tail, $f = 0$ on the right tail, $f \geq 0$, $f(0) - \inf f = 1/64$, and $u = f'$ has Lipschitz constant one. Put

$$H = c_H \sqrt{T} \min\left\{T^{\frac{1}{\mathfrak{p}} - \frac{1}{2}}, \frac{\gamma T}{\log(eT)}\right\}. \tag{91}$$

We will choose c_H first and then choose the lower stepsize coefficient $C \geq 1/c_H$. Since $\mathfrak{p} < 2$, we may enlarge T_0 so that $T^{1/\mathfrak{p}-1/2} \geq 1/c_H$. The lower bound on γ then ensures $H/\sqrt{T} \geq 1$; also $H \leq c_H T^{1/\mathfrak{p}}$. All constants in the stopped recurrence estimates below are obtained under $H/\sqrt{T} \geq 1$ and are independent of c_H. Choose c_H small enough for those estimates and also require

$$0 < c_H \leq \frac{\sigma}{(1 + 2^{\mathfrak{p}-1})^{\frac{1}{\mathfrak{p}}}}.$$

Then the following centered row obeys the conditional $\mathfrak{p}$-moment budget:

$$\mathbb{P}(\xi = -H) = \frac{1}{T}, \qquad \mathbb{P}(\xi = 2H) = \frac{1}{2T}, \qquad \mathbb{P}(\xi = 0) = 1 - \frac{3}{2T}. \tag{92}$$

Indeed its $\mathfrak{p}$-moment is $(1 + 2^{\mathfrak{p}-1})H^{\mathfrak{p}}/T \leq \sigma^{\mathfrak{p}}$.

While the iterate remains at zero, use (92) until the first nonzero child, or for at most $\lfloor T/8 \rfloor$ clocks. Immediately after either outcome use the noiseless row forever. This policy is predictable, so the objective and oracle satisfy Assumptions 1–3. Let E_{ret} be the event that the first nonzero child within the waiting block is $-H$. Then

$$\mathbb{P}(E_{\mathrm{ret}}) = \frac{2}{3}\left[1 - \left(1 - \frac{3}{2T}\right)^{\lfloor \frac{T}{8} \rfloor}\right] \geq c_{\mathrm{ret}} > 0. \tag{93}$$

Before this child, $x = m = 0$, $u(x) = 0$, and the accumulator only decays. If s is its clock, then

$$v_s = \beta_2^{s+1} v + \frac{H^2}{T}, \qquad m_s = -(1 - \beta_1)H,$$

and hence

$$c\gamma\sqrt{T} \leq x_{s+1} = \frac{(1 - \beta_1)\gamma H}{\sqrt{v_s} + \epsilon} \leq C\gamma\sqrt{T}.$$

By reducing the upper constant in the assumed range for γ, this interval lies inside $(0, 1/2^{10})$.

For $k \geq 0$, as long as $0 < x_{s+1+k} \leq 1/16$, write

$$y_k = x_{s+1+k}, \qquad n_k = -m_{s+k}. \tag{94}$$

All later noises are zero and $u(y_k) = -y_k$ by (90), so

$$n_{k+1} = \beta_1 n_k + (1 - \beta_1) y_k, \qquad y_{k+1} = y_k + \lambda_{s+1+k} n_{k+1}. \tag{95}$$

On every stopped prefix of length at most $T/2$,

$$v_{s+k} = \beta_2^k v_s + \frac{1}{T} \sum_{j=0}^{k-1} \beta_2^{k-1-j} y_j^2.$$

Since $H^2/T \geq 1$, $\beta_2^T \geq 1/4$, and $(\sum_{j=0}^{k-1} y_j^2)/T \leq 1$, the explicit bounds are

$$\frac{1}{4}\frac{H^2}{T} \leq v_{s+k} \leq (v + 2)\frac{H^2}{T}, \qquad c\frac{\gamma\sqrt{T}}{H} \leq \lambda_{s+k} \leq C\frac{\gamma\sqrt{T}}{H}.$$

Here λ_{s+k} is the effective stepsize from (71); the comparison constants depend only on v, ϵ, independently of c_H, because $\epsilon\sqrt{T}/H \le \epsilon$.

The position and the negated momentum in (94) are both positive and y_k is nondecreasing. Therefore

$$y_{k+1} \ge y_k \left(1 + c\frac{\gamma\sqrt{T}}{H}\right).$$

Conversely, solving the first recursion and using monotonicity gives

$$n_{k+1} \le H\beta_1^{k+1} + y_k,$$

and hence

$$y_{k+1} \le y_k \left(1 + C\frac{\gamma\sqrt{T}}{H}\right) + C\gamma\sqrt{T}\,\beta_1^{k+1}. \tag{96}$$

The last term is precisely the decaying momentum packet from the shock. Let k be the first index with $y_k \ge 1/32$. Until that clock the stopped memory bounds apply and the lower recursion shows that k is at most

$$K \le C\frac{H}{\gamma\sqrt{T}} \log\left(\frac{e}{\gamma\sqrt{T}}\right) \le \frac{T}{8} \tag{97}$$

clocks. The final inequality follows from H in (91): both alternatives in its minimum give

$$\frac{H}{\gamma\sqrt{T}} \le c_H \frac{T}{\log(eT)},$$

and $\log(e/(\gamma\sqrt{T})) \le C\log(eT)$. Thus choose c_H after the fixed recurrence constants so that the resulting coefficient of T is at most $1/8$, then choose the lower stepsize coefficient at least $1/c_H$, and finally enlarge T_0 as specified above. There is no dependence of the recurrence constants on c_H in this order.

This also proves that the first index exists before the stopped horizon. At its preceding clock, $y_{k-1} < 1/32$, and

$$n_k \le H\beta_1^k + \frac{1}{32}.$$

Consequently

$$y_k - \frac{1}{32} \le C\gamma\sqrt{T}\,\beta_1^k + C\frac{\gamma\sqrt{T}}{H}.$$

If k is below a fixed threshold, the first term is bounded by the sufficiently small upper constant in $\gamma\sqrt{T} \le c$; beyond this threshold, geometric decay makes it arbitrarily small. The second term tends to zero uniformly because $H/\sqrt{T} \ge 1$ and $\gamma \le c/\sqrt{T}$. Thus $y_k \le 1/24$.

For subsequent clocks, iterate (96). Over $j \le cH/(\gamma\sqrt{T})$, its multiplicative factor is at most $\exp(Cc)$, and the remaining momentum tail from the shock contributes at most $C_{\beta_1}\gamma\sqrt{T}$. Choose first the residence constant c and then the upper stepsize constant small enough. It follows that for the next $\lfloor cH/(\gamma\sqrt{T})\rfloor$ clocks,

$$\frac{1}{32} \le y_{k+j} \le \frac{1}{16}.$$

This also closes the stopped accumulator bootstrap. The waiting, amplification, and residence blocks together use less than T clocks; more explicitly, after reducing the fixed residence constant if necessary,

$$\lfloor\frac{T}{8}\rfloor + 1 + K + \left\lfloor\frac{cH}{\gamma\sqrt{T}}\right\rfloor < T,$$

by the growth bound (97) and the height choice (91), which make both K and the final term at most fixed small multiples of T. For the derivative in (90), $|u(y_{k+j})|^2 \ge 2^{-10}$. Substituting the height (91) therefore gives, on the event E_{ret} in (93),

$$\frac{1}{T}\sum_{t=0}^{T-1} |f'(x_t)|^2 \ge c\frac{H}{\gamma\sqrt{T}\,T} = c\min\left\{\frac{1}{\log(eT)}, \frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma T}\right\}.$$

Multiplication by (93) proves the claim. On every other history the objective gradient is bounded by $1/8$, so the expectation is finite. □

E.6 A RESCALED SATURATED PUMP

Lemma 9 (Pump rescaled by the stepsize)**.** *Fix $1 < \mathfrak{p} \le 4/3$ and $\sigma > 0$. There are $c, C, T_0 > 0$, depending only on $(\mathfrak{p}, v, \epsilon, \beta_1, \sigma)$, such that for every $T \ge T_0$, whenever*

$$CT^{\frac{1}{\mathfrak{p}}-\frac{3}{2}}\log(eT) \le \gamma \le cT^{\frac{2}{\mathfrak{p}}-2}, \tag{98}$$

there is a scalar admissible instance satisfying

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[|f'(x_t)|^2\right] \ge c\min\left\{1, T^{\frac{3}{\mathfrak{p}}-\frac{5}{2}}\right\}. \tag{99}$$

Proof. Use the scales

$$\begin{gathered} R = \gamma T^{\frac{1}{\mathfrak{p}}}, \qquad M = c_M T^{\frac{1}{\mathfrak{p}}-1}, \qquad h = c_h T^{\frac{1}{\mathfrak{p}}-1}, \\ \mu^2 = c_\mu h R, \qquad H = c_H T^{\frac{1}{\mathfrak{p}}}, \\ J = \left\lceil C_J \frac{T(1+\gamma\sqrt{T})}{R} \right\rceil, \qquad K = \left\lceil C_K \frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma}\log(eT) \right\rceil. \end{gathered} \tag{100}$$

Here R sets the charging distance, M and H are the charging and return noise amplitudes, h is the plateau derivative, and μ is the maximum magnitude of the negative ramp derivative. The integers J and K bound the bridge and growth phases. Choose

$$0 < c_M \le \min\left\{\frac{1}{8}, \frac{\sigma}{2^{\frac{1}{\mathfrak{p}}}}\right\}, \qquad 0 < c_h < \frac{c_M}{4}, \qquad 0 < c_H \le \frac{\sigma}{(1+2^{\mathfrak{p}-1})^{\frac{1}{\mathfrak{p}}}}.$$

Then choose c_μ, c_H sufficiently small for the trajectory estimates, and finally choose the fixed large constants. The upper coefficient c in (98) is chosen afterward, sufficiently small relative to $c_\mu c_h$ and the fixed landing and momentum constants. For the scales in (100), the range (98) implies

$$\begin{gathered} R \to \infty, \quad R = o(T), \quad J = o(T), \quad K \le \frac{T}{8}, \\ \frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma} \le \frac{T}{C\log(eT)}, \qquad \frac{\gamma^2 T}{\mu^2} = \frac{\gamma T}{c_\mu c_h T^{\frac{2}{\mathfrak{p}}-1}} \le \frac{c}{c_\mu c_h}. \end{gathered} \tag{101}$$

At $\mathfrak{p} = 4/3$, the first assertion follows from the logarithm in the lower endpoint; all the other endpoint inequalities remain valid.

With the charging amplitude from (100), use the centered row

$$\mathbb{P}(\xi = -M) = 1 - \frac{1}{T}, \qquad \mathbb{P}\big(\xi = (T-1)M\big) = \frac{1}{T}. \tag{102}$$

Its conditional moment is at most $\sigma^{\mathfrak{p}}$ because

$$\mathbb{E}\left[|\xi|^{\mathfrak{p}}\right] = c_M^{\mathfrak{p}}\left[(1-\frac{1}{T})T^{1-\mathfrak{p}} + (1-\frac{1}{T})^{\mathfrak{p}}\right] \le 2c_M^{\mathfrak{p}} \le \sigma^{\mathfrak{p}}.$$

On the event of no positive compensator, every charging noise equals $-M$.

Define a deterministic reference trajectory for $\lfloor T/8 \rfloor$ clocks by running the scalar recursion with constant objective derivative h and noise $-M$, starting from the prescribed initial states, and let x_{r} be its endpoint. Along that trajectory

$$v_t = \beta_2^{t+1}v + \frac{(M-h)^2}{T}\sum_{j=0}^{t}\beta_2^j, \qquad m_t = (h-M)(1-\beta_1^{t+1}).$$

In particular $m_t < 0$, so the reference trajectory is increasing. Moreover, $\beta_2^{T/4}$ is bounded below, while $M = o(1)$; hence v_t stays between two fixed positive constants and, after a fixed number of clocks, every displacement is between $c\gamma M$ and $C\gamma M$. Consequently

$$cR \le x_{\mathrm{r}} \le CR. \tag{103}$$

Choose a fixed $B > 1$, put $r = x_{\mathrm{r}} + B$, and choose a holding interval to the right of $r + h$ whose width is $B + C_0\gamma\sqrt{T}$, with C_0 fixed large enough for the return estimate below. Denote its right endpoint by b, and take the launch point to be

$$b - \frac{1-\beta_1}{2}\gamma\sqrt{T}. \tag{104}$$

The choice of C_0 puts this point strictly inside the holding interval. Using the profiles in (80) and (81), define

$$u_{T,\gamma}(y) = \mathrm{Trap}_{h;-B,r}(y) - \mathrm{Tri}_{\mu;b}(y), \qquad f_{T,\gamma}(y) = C_{T,\gamma} + \int_0^y u_{T,\gamma}(s)\,\mathrm{d}s, \tag{105}$$

where $C_{T,\gamma}$ makes the global minimum zero. The positive signed area is $h(x_{\mathrm{r}}+2B)+h^2$, while the negative signed area is $\mu^2 = c_\mu hR$. By (103), choosing c_μ small makes the negative area at most half the positive area. Hence $f_{T,\gamma} \geq 0$, $u_{T,\gamma}$ has Lipschitz constant one, and

$$f_{T,\gamma}(0) - \inf f_{T,\gamma} = hB + \frac{h^2}{2} \leq 1.$$

Conditional on no earlier compensator, use (102) during the first $\lfloor T/8 \rfloor$ clocks and then for at most J further clocks, with the bridge length from (100). Stop this row as soon as the iterate reaches the launch point (104). If a compensator appears or this bridge times out before reaching that point, use the noiseless row forever; reaching it instead initiates the flush described below. Thus the row choice at every clock is a complete predictable function of the observed history. On the event without a compensator, the charging trajectory is exactly the reference trajectory, because every update starts on the plateau with gradient h; an argument at the first exit using (103) verifies this premise. Past the reference endpoint one has $0 \leq u_{T,\gamma} \leq h < M/2$, so the stochastic gradient remains at most $-M/2$. The accumulator remains between fixed positive constants: its contribution from the initial state is bounded below because this phase ends before $T/4$, and its additional bridge contribution is at most $CJM^2/T = o(1)$, using $J = o(T)$ and $M = o(1)$ from (100). After a fixed number of updates, the rightward speed is at least $c\gamma M$. The distance through the end of the trapezoid and across the holding interval is $O(1+\gamma\sqrt{T})$. Thus the launch point is reached within J further clocks. Since $T/8 + J < T/4$, the event of no compensator through this bridge has probability at least

$$(1 - \frac{1}{T})^{\lfloor \frac{T}{8} \rfloor + J} \geq c_{\mathrm{ch}} > 0.$$

At the crossing, the overshoot in one step is $O(\gamma M)$. During the next $J_{\beta_1}(T)$ clocks with zero gradients, with this length defined in (75), momentum decays geometrically; since the denominator remains bounded below and above, their total movement is $O_{\beta_1}(\gamma M)$, and the residual momentum is $O_{\beta_1}(M(eT)^{-12})$. Because $\gamma M/(\gamma\sqrt{T}) = M/\sqrt{T} = o(1)$, the holding margins absorb both quantities.

After the flush, use the return amplitude from (100) in the centered row

$$\mathbb{P}(\xi = -H) = \frac{1}{T}, \qquad \mathbb{P}(\xi = 2H) = \frac{1}{2T}, \qquad \mathbb{P}(\xi = 0) = 1 - \frac{3}{2T}$$

until its first nonzero child or for $\lfloor T/8 \rfloor$ clocks, and then use zero noise. Its $\mathfrak{p}$-moment is at most $\sigma^{\mathfrak{p}}$ by the choice of c_H. Together with (105) and the charging row, this verifies Assumptions 1–3. The probability that its first nonzero child is $-H$ is bounded below exactly as in (93). On this desired event the accumulator before the return is bounded, the flushed momentum and its remaining displacement are negligible, and more precisely

$$x_s = b - \frac{1-\beta_1}{2}\gamma\sqrt{T} + O_{\beta_1}(\gamma M), \qquad m_{s-1} = O_{\beta_1}(M(eT)^{-12}).$$

Furthermore,

$$v_s = \beta_2 v_{s-1} + \frac{H^2}{T} \asymp T^{\frac{2}{\mathfrak{p}}-1}, \qquad \lambda_s H = \gamma\sqrt{T}\,(1 + o(1)).$$

Because $H^2/T \to \infty$, the exact return update and the launch point (104) give

$$x_{s+1} - b = \frac{1-\beta_1}{2}\gamma\sqrt{T} + o_{\beta_1}(\gamma\sqrt{T}).$$

In particular the landing distance is between $(1-\beta_1)\gamma\sqrt{T}/4$ and $3(1-\beta_1)\gamma\sqrt{T}/4$. The last inequality in (101) and a sufficiently small upper constant in (98) ensure that this landing lies in $(0, \mu/8)$ relative to b.

Write

$$y_k = x_{s+1+k} - b, \qquad n_k = -m_{s+k}. \tag{106}$$

Until y_k reaches $\mu/2$, the same exact recursion as (95) holds. On every stopped prefix of length at most $K + cT^{1/\mathfrak{p}-1/2}/\gamma$,

$$v_{s+k} = \beta_2^k v_s + \frac{1}{T}\sum_{j=0}^{k-1} \beta_2^{k-1-j} y_j^2.$$

The growth length and ramp height in (100) bound the feedback relative to the deposited return memory by

$$\frac{(K + \frac{cT^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma})\mu^2}{T\,T^{\frac{2}{\mathfrak{p}}-1}} \le CT^{\frac{1}{\mathfrak{p}}-\frac{3}{2}}\log(eT) = o(1).$$

Therefore, uniformly on every such prefix,

$$v_{s+k} \asymp T^{\frac{2}{\mathfrak{p}}-1}, \qquad \lambda_{s+k} \asymp \gamma T^{\frac{1}{2}-\frac{1}{\mathfrak{p}}}.$$

For the displacement and the negated momentum in (106), positivity gives, as in (95),

$$y_{k+1} \ge y_k\left(1 + c\gamma T^{\frac{1}{2}-\frac{1}{\mathfrak{p}}}\right),$$

whereas the exact momentum convolution gives

$$n_{k+1} \le H\beta_1^{k+1} + y_k.$$

The landing is at least $c\gamma\sqrt{T}$, and $\log(\mu/(\gamma\sqrt{T})) \le C\log(eT)$. The lower estimate therefore reaches $\mu/4$ within K clocks, with K defined in (100). At the first crossing, the upper estimate bounds the overshoot by

$$C\gamma\sqrt{T}\,\beta_1^k + C\gamma T^{\frac{1}{2}-\frac{1}{\mathfrak{p}}}\mu.$$

By the last inequality in (101), the first term is at most a chosen small multiple of μ, uniformly in k, after taking the upper constant in (98) sufficiently small; this small multiple is absorbed by the crossing margin. The second term is $o(\mu)$ because $\gamma T^{1/2-1/\mathfrak{p}} \le cT^{1/\mathfrak{p}-3/2}$. Thus the first crossing lies below $\mu/3$. Iterating the same upper estimate over a sufficiently small fixed multiple of $T^{1/\mathfrak{p}-1/2}/\gamma$ keeps the multiplicative factor bounded by a fixed constant and keeps the shock tail below the chosen small multiple of μ. Taking that multiple and the fixed constant controlling residence time sufficiently small, y_k remains in $[\mu/4, \mu/2]$ for at least

$$c\frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma}$$

additional clocks. Equations (101) and $J = o(T)$ place the charging, bridge, flush, return wait, growth, and residence phases strictly before the horizon. Using the phase lengths in (75) and (100), for all sufficiently large T,

$$2\lfloor\frac{T}{8}\rfloor + J + J_{\beta_1}(T) + 1 + K + \left\lfloor c\frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma}\right\rfloor < T.$$

The charging and return events use fresh randomness on predictable successive phases, so their intersection has probability at least a fixed positive constant. The residence bound and $\mu^2 = c_\mu hR$ from (100) give, on that event,

$$\frac{1}{T}\sum_{t=0}^{T-1} |f'_{T,\gamma}(x_t)|^2 \ge c\min\left\{1, T^{\frac{3}{\mathfrak{p}}-\frac{5}{2}}\right\}.$$

On a compensator or timeout, the controller switches permanently to the noiseless row. Every row has finite support and $u_{T,\gamma}$ is bounded for fixed T, so the average squared derivative is finite on every history. This proves (99). □

E.7 SPLICING THE FOUR MECHANISMS

Proposition 7 (Normalized scalar construction uniform over stepsizes). *Fix $(\mathfrak{p}, v, \epsilon, \beta_1, \sigma)$ with $1 < \mathfrak{p} \le 2$, $v, \sigma > 0$, $\epsilon \ge 0$. Use the moment coefficients specified in Section 2.2. There are $c_0, T_0 > 0$, depending only on this tuple, such that for every $T \ge T_0$ and every $\gamma > 0$, there is a scalar auxiliary objective and a predictable oracle policy satisfying Assumptions 1–3, both fixed before the run, with $0 \le f(0) - \inf f \le 1$ and, for the exponent $r_*(\mathfrak{p})$ defined in (1),*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[|f'(x_t)|^2\right] \ge \begin{cases} \dfrac{c_0}{\log(eT)}, & 1 < \mathfrak{p} \le \frac{6}{5}, \\ c_0 T^{r_*(\mathfrak{p})}, & \frac{6}{5} < \mathfrak{p} \le 2. \end{cases} \tag{107}$$

Proof. The explicit objectives and noise distributions in Lemma 5 and Lemmas 7–9 satisfy Assumptions 1–3, as verified in their proofs. We combine their lower bounds over the stepsize ranges. The exact plateau estimate also implies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[|f'(x_t)|^2\right] \ge \min\left\{1, \frac{\epsilon + \frac{\sqrt{v}}{2}}{2}\right\}\min\left\{1, \frac{1}{\gamma T}\right\},$$

so it supplies the comparison for small stepsizes used below. All comparisons below are up to fixed positive constants, which may depend on $\sigma > 0$; adjacent ranges can therefore be made to overlap by reducing the final payoff constant and enlarging the fixed horizon threshold. The required overlaps are nonempty for all sufficiently large T: throughout $\mathfrak{p} > 1$,

$$T^{\frac{1}{\mathfrak{p}} - \frac{3}{2}}\log(eT) = o(T^{-\frac{1}{2}}),$$

while for $\mathfrak{p} > 6/5$,

$$T^{\frac{3}{\mathfrak{p}} - \frac{5}{2}}\log(eT) = o(1),$$

and for $\mathfrak{p} < 4/3$,

$$T^{1-\frac{2}{\mathfrak{p}}} \le T^{\frac{2}{\mathfrak{p}} - 2}.$$

These are exactly the comparisons between the ramp lemma's admissible range, the pump lemma's admissible range, and the joining scales below.

Suppose $1 < \mathfrak{p} \le 6/5$. If $\gamma \lesssim \log(eT)/T$, Lemma 5 gives an expected average squared derivative of at least a fixed positive multiple of $1/\log(eT)$. From this scale through

$$\gamma \lesssim T^{\frac{1}{\mathfrak{p}} - \frac{3}{2}}\log(eT),$$

Lemma 8 gives the same lower bound. From there through

$$\gamma \lesssim T^{\frac{2}{\mathfrak{p}} - 2},$$

Lemma 9 gives a fixed positive lower bound. For every larger γ, Lemma 7 also gives a fixed positive lower bound: all three entries in its minimum are bounded below by positive constants at the joining point and are nondecreasing in γ. This proves the first branch of (107).

Now let $6/5 < \mathfrak{p} < 4/3$. In this case $T^{3/\mathfrak{p} - 5/2} = o(1)$. Lemma 5 covers

$$\gamma \lesssim T^{\frac{3}{2} - \frac{3}{\mathfrak{p}}}.$$

Lemma 8 then covers the interval through

$$\gamma \lesssim T^{1-\frac{2}{\mathfrak{p}}},$$

because throughout this interval both entries on the right of (89) are at least $T^{3/\mathfrak{p} - 5/2}$. Lemma 9 covers from there through $\gamma \lesssim T^{2/\mathfrak{p} - 2}$. Finally, Lemma 7 covers every larger γ, since at the join

$$\gamma T^{\frac{1}{\mathfrak{p}} - \frac{1}{2}} = T^{\frac{3}{\mathfrak{p}} - \frac{5}{2}}.$$

This proves the $T^{r_*(\mathfrak{p})}$ lower bound in the middle range, with the exponent defined in (1).

Finally, let $4/3 \le \mathfrak{p} \le 2$. The target is $T^{1/\mathfrak{p}-1} = T^{-(\mathfrak{p}-1)/\mathfrak{p}}$. Lemma 5 covers $\gamma \lesssim T^{-1/\mathfrak{p}}$, and Lemma 8 covers from there through $\gamma \lesssim T^{-1/2}$, because

$$\frac{T^{\frac{1}{\mathfrak{p}}-\frac{1}{2}}}{\gamma T} \ge T^{\frac{1}{\mathfrak{p}}-1}.$$

Lemma 7 covers the rest, with equality of the relevant scales at $\gamma = T^{-1/2}$. For $\mathfrak{p} = 2$, the ranges covered by the plateau and rare shocks can be chosen to overlap directly.

Each of the four objectives has initial gap at most one, as checked in its construction. The choice of mechanism depends only on the announced (T, γ), and every oracle policy is fixed before sampling starts. Taking the smallest of the finitely many positive comparison constants and the largest of the fixed horizon thresholds gives c_0, T_0 independent of T and γ. This proves the normalized scalar statement. Section E.8 completes each selected objective to initial gap exactly one and embeds it into any fixed dimension. □

E.8 Formal statement and completion of the lower bound

Theorem 8 (Formal statement of Theorem 3)**.** *Set $L = 1$. Fix $1 < \mathfrak{p} \le 2$, an integer $d \ge 1$, $v > 0$, $\epsilon \ge 0$, and $\sigma > 0$, independently of (T, γ). Run Algorithm 1 from $\mathbf{x}_0 = 0$ with the moment coefficients specified in Section 2.2, and use $r_*(\mathfrak{p})$ from (1). There are $c, T_1 > 0$, depending only on $(\mathfrak{p}, d, v, \epsilon, \beta_1, \sigma)$, such that for every integer $T \ge T_1$ and every $\gamma > 0$, there exist an objective f with $\Delta_f = 1$ and an oracle policy satisfying Assumptions 1–3 for which*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \ge \begin{cases} \max\left\{\min\left\{1, \dfrac{\epsilon + \frac{\sqrt{v}}{2}}{2\gamma T}\right\}, \dfrac{2}{T}, \dfrac{c}{\log(eT)}\right\}, & 1 < \mathfrak{p} \le \frac{6}{5}, \\ \max\left\{\min\left\{1, \dfrac{\epsilon + \frac{\sqrt{v}}{2}}{2\gamma T}\right\}, \dfrac{2}{T}, cT^{r_*(\mathfrak{p})}\right\}, & \frac{6}{5} < \mathfrak{p} \le 2. \end{cases}$$

In particular, any choice of $\gamma = \gamma(T) > 0$ independent of the instance admits, at each $T \ge T_1$, an instance satisfying

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \ge \begin{cases} \max\left\{\dfrac{2}{T}, \dfrac{c}{\log(eT)}\right\}, & 1 < \mathfrak{p} \le \frac{6}{5}, \\ \max\left\{\dfrac{2}{T}, cT^{r_*(\mathfrak{p})}\right\}, & \frac{6}{5} < \mathfrak{p} \le 2. \end{cases} \tag{108}$$

The objective and complete predictable oracle policy may depend on (T, γ), but are fixed before the run. All witnesses have the prescribed dimension and moment budget. Neither c nor T_1 depends on γ. The plateau and $\frac{2}{T}$ terms also hold when $\sigma = 0$.

Proof of Theorem 8. Fix the parameter tuple $(\mathfrak{p}, d, v, \epsilon, \beta_1, \sigma)$, with $1 < \mathfrak{p} \le 2$, $d \ge 1$, $v, \sigma > 0$, and $\epsilon \ge 0$, independently of (T, γ), and use the moment coefficients in Section 2.2. Proposition 7 provides a constant $c_0 > 0$ and threshold T_0 depending only on the fixed scalar parameters. For every announced $T \ge T_0$ and $\gamma > 0$, it gives a scalar objective and predictable oracle policy, both fixed before the run, with initial gap at most one and the lower bound (107).

Apply Lemma 6 at a deterministic $A < -2\gamma T\sqrt{T}$ to give the objective initial gap exactly one. Write $f^{(1)}$ and $\xi_t^{(1)}$ for this scalar construction and set

$$F(x_1, \dots, x_d) = f^{(1)}(x_1), \qquad \xi_t = (\xi_t^{(1)}, 0, \dots, 0).$$

Nonnegativity, continuous differentiability, and initial gap one are preserved. Moreover,

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| = |(f^{(1)})'(x_1) - (f^{(1)})'(y_1)| \le |x_1 - y_1| \le \|\mathbf{x} - \mathbf{y}\|,$$

and

$$\mathbb{E}\left[\xi_t \mid \mathscr{F}_{t-1}\right] = 0, \qquad \mathbb{E}\left[\|\xi_t\|^{\mathfrak{p}} \mid \mathscr{F}_{t-1}\right] = \mathbb{E}\left[|\xi_t^{(1)}|^{\mathfrak{p}} \,\middle|\, \mathscr{F}_{t-1}\right] \le \sigma^{\mathfrak{p}}.$$

Every inactive coordinate has $x_{t,i} = m_{t,i} = 0$ and $v_{t,i} = \beta_2^{t+1} v$, while coordinate one follows the scalar recursion exactly. Thus the average squared gradient is unchanged, and the lower bound (107)

is preserved for every $T \geq T_0$ and $\gamma > 0$. Lemma 5 supplies an instance with the exact plateau contribution. For the remaining deterministic term, take zero noise and

$$f(\mathbf{x}) = \frac{1}{2}(x_1 - \sqrt{2})^2.$$

This objective has initial gap one and $\|\nabla f(0)\|^2 = 2$, so its first summand alone gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{2}{T} \qquad (\gamma > 0).$$

For the announced (T, γ), select the construction corresponding to the largest of these three bounds before the run. Denote its objective by f. Its initial gap is exactly one, its dimension is d, and its oracle policy satisfies the same conditional noise budget. Combining (107) with the two deterministic bounds proves the first inequality in Theorem 8, with $c = c_0$ and $T_1 = T_0$. The constant and threshold are uniform over $\gamma > 0$. Choosing between the two noiseless objectives also proves their lower bounds when $\sigma = 0$.

The translation argument in Appendix A transfers these witnesses from the origin to any prescribed $\mathbf{x}_0 \in \mathbb{R}^d$, preserving their gaps, noise budgets, dimension, and payoffs.

Finally, let $\gamma = \gamma(T) > 0$ be any stepsize rule chosen independently of the instance. For every $T \geq T_0$, the preceding construction at that stepsize gives an admissible objective and oracle policy satisfying (108), with $c = c_0$. The objective and policy may depend on the announced horizon and stepsize, and are fixed before sampling starts. This proves (108) and every assertion in Theorem 8. With the fixed parameters absorbed into Ω, the first displayed bound in that theorem gives (7), proving Theorem 3. □