HIGH-CONFIDENCE STATIONARITY OF MOMENTUM MUON: A HEAVY-TAILED SEPARATION FROM ADAM

Anonymous authors
Paper under double-blind review

ABSTRACT

As an emerging optimization method, Muon has been empirically shown to outperform Adam/AdamW on modern machine learning tasks, yet the mechanism underlying this advantage remains unclear. We identify a key mechanism behind Muon's high-probability advantage: its instantaneous singular-value compression provides the scale regularization sought by Adam's second-moment normalization while directly controlling unpreconditioned stationarity. By avoiding the confidence loss incurred when converting Adam's preconditioned adaptive energy into unpreconditioned stationarity, Muon achieves sharper confidence dependence under heavy-tailed noise. Specifically, under conditional p-moment noise for any $p \in (1, 2]$, we prove a finite-time high-probability theorem for exact canonical prepolar momentum Muon, with $\beta \in [0, 1)$ fixed within each run. Choosing momentum and stepsize for the given horizon at fixed batch size yields the average-gradient rate $\mathcal{O}(T^{-a}(1 + \log(1/\delta)))$, where $a = (p - 1)/(3p - 2)$. Even when stochastic gradients have infinite variance, the dependence on the confidence parameter δ remains only logarithmic. For the five-step Newton–Schulz (NS5) variant, we obtain the same rate up to an additive error floor under an explicit approximation-error condition. In contrast, for Adam with fixed $\beta_2 \in (0, 1)$, $\epsilon > 0$ and initial moments, we construct a uniformly p-moment-bounded two-dimensional family and prove a minimax stationarity lower bound $T^{-1} \sum_{t=1}^{T} \|\nabla f(X_t)\|_{\mathrm{F}} \geq c \min\{1, T^{-a}\delta^{-q}\}$ with probability greater than δ.

1 INTRODUCTION

Muon has shown strong empirical performance in modern machine learning (Jordan et al., 2024; Liu et al., 2025). A theoretical comparison with Adam requires specifying both the optimization criterion and the noise model. Here we study high-probability stationarity: expectations can conceal rare, large errors, whereas a $1 - \delta$ guarantee exposes performance at a small failure probability δ.

This perspective is relevant under heavy-tailed noise. Heavy-tailed stochastic gradients or gradient noise have been observed in neural networks, attention models, reinforcement learning, and generative and discriminative tasks (Simsekli et al., 2019; Zhang et al., 2020; Garg et al., 2021; Battash et al., 2024); language tasks also exhibit heavy-tailed class imbalance (Kunstner et al., 2024), and multiplicative noise can produce heavy-tailed trajectories even from light-tailed data (Hodgkinson & Mahoney, 2021; Gurbuzbalaban et al., 2021).

For a rank-s compact SVD $M = U_s \Sigma_s V_s^\top$, Muon's exact normalization is

$$\mathrm{polar}(M) = U_s V_s^\top,$$

with $\mathrm{polar}(0) = 0$. This replaces every nonzero singular value by one and gives

$$\langle M, \mathrm{polar}(M) \rangle = \|M\|_*, \qquad \|\mathrm{polar}(M)\|_{\mathrm{op}} \leq 1.$$

The update is bounded while retaining a nuclear-norm descent certificate. For pre-polar momentum, conditional p-moment estimates over momentum-length blocks, geometric contraction, and a log-shock supermartingale give logarithmic confidence dependence. This argument directly controls the unweighted gradient sum without removing a random historical denominator. For smooth lower-bounded objectives under conditional p-moment noise, $p \in (1, 2]$, choosing the momentum and

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.

stepsize of Choudhury et al. (2026) for the given horizon at fixed batch size yields an average nuclear-gradient bound of $\mathcal{O}\Big(T^{-\frac{p-1}{3p-2}}\,\big(1+\log\frac{1}{\delta}\big)\Big)$ with probability at least $1-\delta$.

Our comparison allows Adam to tune its first-moment memory for each horizon. Fix its second-moment parameter $\beta_2 \in (0,1)$, denominator offset $\epsilon > 0$, and initialization $m_0 = 0$, $v_0 = \nu\mathbf{1}$, $\nu \geq 0$. The first-moment parameter ranges over all of $[0,1)$, and the positive constant learning rate is unrestricted. We study the worst-case high-confidence rate on a common class of smooth objectives with bounded conditional p-moment noise.

For every $p \in (1,2]$, write $a = (p-1)/(3p-2)$ and fix $0 < q < a/(2p)$. There are constants $c, \bar{\delta} > 0$ such that, for every horizon T, first-moment parameter β_1, learning rate $\gamma > 0$ and confidence level $0 < \delta < \bar{\delta}$, Adam satisfies on a preselected instance

$$\mathbb{P}\left\{\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} \geq c\min\{1, T^{-a}\delta^{-q}\}\right\} > \delta. \tag{1}$$

Every instance is a 1×2 row-matrix problem with smoothness bound one, initial objective gap one, and centered primitive p-moment at most one. Both c and $\bar{\delta}$ are independent of $T, \delta, \beta_1, \gamma$ and the selected instance. The result therefore covers every confidence-independent rule $\beta_1(T) \in [0,1)$, including memories approaching one as the horizon grows. At $p = 2$, it gives, for example, $c\min\{1, T^{-1/4}\delta^{-1/32}\}$.

The construction combines complementary effects of Adam's two memories. Rare impulses send the iterates onto a long plateau, while deterministic oscillating gradients sustain motion under long first-moment memory. An interpolation that preserves the queried gradients controls the objective along the entire path, including arbitrarily large learning rates. These mechanisms yield one uniform rate across the full tuning domain.

On row matrices, nuclear and Frobenius norms coincide. Muon's upper bound on the same selected objective and oracle has a constant depending only on p and the batch size. For $T^{-a/q} \leq \delta < \bar{\delta}$, the two bounds have the common factor T^{-a}, and their ratio is a positive constant times $\delta^{-q}/(1+\log(1/\delta))$. The comparison isolates a confidence advantage that persists under first-moment tuning. The matrix upper theorem describes the role of nuclear–operator duality, while the row-matrix lower bound identifies the effects of temporal normalization and momentum. The controlled UCI Iris experiment uses 5,000 coupled trajectories per tail index with mean-one Pareto radial noise to examine confidence profiles, and directional noise to compare momentum lengths. At 20.48M input tokens per method, Qwen3-0.6B yields endpoint development NLLs of 5.0501 for Muon, 5.2440 for Adam, and 5.2230 for AdamW. Native gradient-noise diagnostics at fixed checkpoints yield fitted GPD tail indices below two at all four thresholds for four of 24 state–matrix pairs.

Contributions.

1. For every $p \in (1,2]$, we prove a finite-time high-probability guarantee for exact-polar momentum Muon. Conditional moment estimates and geometric tracking yield logarithmic confidence dependence with fixed batch size. A companion NS5 theorem specifies its approximation-error condition.
2. We prove Adam's polynomial-confidence lower bound uniformly over all first-moment parameters and positive constant learning rates, with a fixed second-moment parameter and positive offset. A two-dimensional instance family has common smoothness, initial gap, noise budget and rate constants.
3. On the same selected instances, Adam's lower bound and Muon's upper guarantee have the same time exponent and different confidence dependence. Controlled real-data experiments examine normalization and momentum, while language-model training provides a practical comparison with native gradient-noise tail diagnostics.

2 Problem Setup

Fix dimensions $m, n \in \mathbb{N}_+$ and consider the matrix optimization problem $\min_{X\in\mathbb{R}^{m\times n}} f(X)$, where $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is differentiable. For this matrix block, fix a batch size $B \in \mathbb{N}_+$, a horizon

$T \in \mathbb{N}_+$, and a deterministic step $\eta > 0$, with iterates $X_t \in \mathbb{R}^{m\times n}$. Set $r := \min\{m, n\}$ and $\langle A, B\rangle := \mathrm{tr}(A^\top B)$; write the nuclear, Frobenius, and operator norms as $\|\cdot\|_*$, $\|\cdot\|_{\mathrm{F}}$, and $\|\cdot\|_{\mathrm{op}}$, with $\|Z\|_* \le \sqrt{r}\,\|Z\|_{\mathrm{F}}$.

Let $(\mathscr{F}_t)_{t=0}^T$ be the filtration of the algorithm and oracle, including any initial oracle state. Take X_1 deterministic and X_t to be $\mathscr{F}_{t-1}$-measurable. Given $\mathscr{F}_{t-1}$, the oracle draws an independent fresh batch $G_{t,1}, \dots, G_{t,B} \in \mathbb{R}^{m\times n}$, and we set $M_t := \frac{1}{B}\sum_{j=1}^B G_{t,j}$. Both M_t and $M_t - \nabla f(X_t)$ are $\mathscr{F}_t$-measurable, with dependence across times allowed. Each iteration averages B primitive stochastic gradients, and the algorithm runs for T iterations.

2.1 Muon

Muon turns the stochastic matrix estimate into a spectrally normalized update. The polar-and-momentum rule and five-step Newton–Schulz candidate used here follow the core Muon construction introduced by Jordan et al. (2024) and subsequently developed at scale by Liu et al. (2025).

We isolate the spectral map used by the update. Unlike coordinatewise normalization, it acts on the singular geometry of the entire matrix. For a rank-s compact SVD $M = U_s\Sigma_s V_s^\top$, with $U_s \in \mathbb{R}^{m\times s}$, $\Sigma_s \in \mathbb{R}^{s\times s}$, and $V_s \in \mathbb{R}^{n\times s}$, define the canonical partial polar map $\mathrm{polar}(M) := M\big((M^\top M)^\dagger\big)^{\frac{1}{2}} = U_s V_s^\top$, $\qquad \mathrm{polar}(0) := 0$. Here $\dagger$ denotes the Moore–Penrose pseudoinverse, which keeps the definition valid when M is rank deficient. For a nonzero matrix, the map preserves its left and right singular subspaces while replacing every nonzero singular value by one.

Momentum is accumulated before this spectral normalization. The exact polar map is the reference direction. In the NS5 mode, $\mathsf{NS}_5(N_t; \varepsilon_{\mathrm{NS}})$ denotes the standard five-step Newton–Schulz map used by the cited Muon implementations.

Algorithm 1: Muon

1 **Input:** X_1, $U_0 = 0$, horizon T, batch size $B \in \mathbb{N}_+$, step $\eta > 0$, momentum $\beta \in [0, 1)$, and mode exact or NS5.
  In NS5 mode, also input $\varepsilon_{\mathrm{NS}} > 0$.
2 **for** $t = 1, \dots, T$ **do**
3 Query the fresh oracle batch and form $M_t = B^{-1}\sum_{j=1}^B G_{t,j}$.
4 Set $U_t = \beta U_{t-1} + (1-\beta)M_t$ and $N_t = (1-\beta)M_t + \beta U_t = \beta^2 U_{t-1} + (1-\beta^2)M_t$.
5 Set $P_t = \mathrm{polar}(N_t)$.
6 In NS5 mode, replace $P_t \leftarrow \mathsf{NS}_5(N_t; \varepsilon_{\mathrm{NS}})$.
7 Update $X_{t+1} = X_t - \eta P_t$.
8 **end for**; **return** $X_{1:T+1}$.

Here U_t carries the exponentially weighted history, N_t combines that history with the current batch estimate, and P_t is the normalized update direction. The matrix N_t is the Nesterov numerator. The recursion is adapted to the same filtration: U_{t-1} is known before call t, while U_t, N_t, P_t, and X_{t+1} are $\mathscr{F}_t$-measurable after the call.

Line 5 supplies the exact polar direction, while Line 6 replaces it by the five-step Newton–Schulz direction in NS5 mode. The batch size, momentum, and stepsize remain constant throughout a run, and $U_0 = 0$.

2.2 Technical Assumptions

Assumption 1 (Finite lower bound)**.** *There exists a deterministic $f_\star > -\infty$ such that*

$$f(X) \ge f_\star \qquad \text{for every } X \in \mathbb{R}^{m\times n}.$$

The initial gap $\Delta_1 := f(X_1) - f_\star$ bounds accumulated objective decrease. This matches the lower-boundedness used for Adam and normalized SGD (Jin et al., 2026; Hübler et al., 2025), including an unattained lower bound $f_\star$.

Assumption 2 (Global Frobenius gradient Lipschitzness)**.** *For a deterministic $L > 0$,*

$$\|\nabla f(X) - \nabla f(Y)\|_{\mathrm{F}} \le L\|X - Y\|_{\mathrm{F}} \quad \text{for all } X, Y \in \mathbb{R}^{m\times n}. \tag{2}$$

This standard smoothness condition controls the quadratic remainder of a stochastic first-order step. The same global smoothness model appears in prior Adam and normalized-SGD analyses (Jin et al., 2026; Hübler et al., 2025).

Assumption 3 (Conditional p-moment oracle)**.** *Fix $p \in (1, 2]$ and deterministic $\sigma_0, \sigma_1 \geq 0$. For every $1 \leq t \leq T$ and $1 \leq j \leq B$, the primitive stochastic gradient is conditionally unbiased and satisfies $\mathbb{E}[G_{t,j} \mid \mathscr{F}_{t-1}] = \nabla f(X_t)$, and*

$$\mathbb{E}[\|G_{t,j} - \nabla f(X_t)\|_{\mathrm{F}}^p \mid \mathscr{F}_{t-1}] \leq \sigma_0^p + \sigma_1^p \|\nabla f(X_t)\|_{\mathrm{F}}^p.$$

Assumption 3 allows additive noise, measured by σ_0, and gradient-dependent noise, measured by σ_1. The order-p moment model permits heavy tails and unbounded gradients or iterates; $\sigma_1 > 0$ extends the p-BCM model of Hübler et al. (2025).

Immediate batch consequence. Under Assumption 3, conditional symmetrization and nuclear–Frobenius comparison give

$$\mathbb{E}[\|M_t - \nabla f(X_t)\|_* \mid \mathscr{F}_{t-1}] \leq 2\sqrt{r}\, B^{-\frac{p-1}{p}} \big(\sigma_0 + \sigma_1 \|\nabla f(X_t)\|_*\big).$$

Averaging B primitive stochastic gradients thus reduces the centered noise.

Parameter choice. Given T and $p \in (1, 2]$, choose the Nesterov-Muon parameters (Choudhury et al., 2026), $\beta = 1 - T^{-\frac{p}{3p-2}}$, $\eta = (1-\beta)^{\frac{2p-1}{p}}$. Momentum averages noise across iterates, while the stepsize controls gradient drift over that memory.

2.3 Stationarity and comparison protocol

We compare the methods using the average Frobenius-gradient norm $T^{-1} \sum_{t=1}^{T} \|\nabla f(X_t)\|_{\mathrm{F}}$. The average includes the initial point X_1 and uses the true objective gradient at each iterate. On single-row instances, it is the average Euclidean-gradient norm.

A high-confidence comparison uses the same selected objective, initialization, oracle law, batch size B, horizon T, and the gradient-norm average above. Fix p, B and Adam's $\beta_2 \in (0, 1)$, $\epsilon > 0$ and $v_0 = \nu \mathbf{1}$, $\nu \geq 0$, with $m_0 = 0$. Its first-moment parameter $\beta_1 \in [0, 1)$ and constant learning rate $\gamma > 0$ may be chosen for each horizon before sampling. Our lower bound assigns a preselected hard instance to each $T, \delta, \beta_1, \gamma$, with common dimension, smoothness, initial gap and conditional noise-moment bounds. Both the Adam lower constant and the Muon upper constant are uniform over this family. In particular, the comparison covers $\beta_1(T)$ rules chosen independently of δ.

3 Main Results

Muon and Adam rescale noisy gradients in different spaces and at different times. Muon applies a spectral map to the current momentum numerator, while Adam divides each coordinate by a random history-dependent scale. These operations lead to different confidence estimates and different responses to temporally structured noise.

3.1 Differences between the two normalizations

Polar normalization gives a direct support inequality for descent. Adam combines a signed first-moment history with a coordinatewise second-moment scale. The interaction between these two histories determines its response to rare impulses and persistent noise.

Muon: confidence preservation. Muon normalizes the singular values of its momentum numerator. For the canonical partial polar, $\langle M, \mathrm{polar}(M) \rangle = \|M\|_*$, $\|\mathrm{polar}(M)\|_{\mathrm{op}} \leq 1$, $\|\mathrm{polar}(M)\|_{\mathrm{F}}^2 = \mathrm{rank}(M) \leq r$. The pairing links descent to the nuclear norm; the bounds control update size. Combined with momentum tracking and the conditional p-moment analysis, this gives a nuclear-gradient bound with logarithmic dependence on $1/\delta$. For every trajectory,

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(X_t)\|_{\mathrm{F}} \leq \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(X_t)\|_*.$$

Thus the average Frobenius-gradient norm inherits the bound on the same $1-\delta$ event, preserving the logarithmic confidence dependence.

Adam: historical normalization. Adam instead divides each coordinate by its historical second-moment scale. Write $x_t = \mathrm{vec}(X_t)$ and $g_t = \mathrm{vec}(M_t)$. The Adam recursion uses first- and second-moment estimates, with all scalar parameters shared across coordinates and fixed during a run.

Algorithm 2: Adam

Input: x_1, $m_0 = 0$, $v_0 = \nu\mathbf{1}$ with $\nu \geq 0$, fixed $B, T \in \mathbb{N}_+$, $\gamma > 0$, $\epsilon > 0$, $\beta_1 \in [0, 1)$ and $\beta_2 \in (0, 1)$.
for $t = 1, \ldots, T$ **do**
  Form g_t by averaging B fresh primitive gradients.
  Set $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_t \odot g_t)$.
  Update $x_{t+1} = x_t - \gamma \dfrac{m_t}{\sqrt{v_t} + \epsilon}$ coordinatewise.
end for; **return** $x_{1:T+1}$.

Under bounded conditional variance, Jin et al. (2026) use fixed $v_0 = v\mathbf{1} > 0$, $\epsilon > 0$, and sufficiently small fixed $\eta_{\mathrm{ref}} > 0$. For $\beta_2 = 1 - 1/T$ and $\gamma = \eta_{\mathrm{ref}}/\sqrt{T}$, with probability at least $1-\delta$,

$$\mathcal{E}_T := \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{\eta_{\mathrm{ref}} (\nabla_i f(x_t))^2}{\sqrt{T}(\sqrt{v_{t,i}} + \epsilon)} \lesssim \log^6 \frac{1}{\delta}.$$

Writing $\Sigma_T := \sum_i v_{T,i}$, their terminal comparison bounds the average squared gradient norm by $\frac{2(\sqrt{\Sigma_T}+\epsilon)}{\eta_{\mathrm{ref}}\sqrt{T}} \mathcal{E}_T$. The tail estimate $\mathbb{P}\{\sqrt{\Sigma_T} + \epsilon > H\} \lesssim H^{-2}$ then gives the confidence cost: taking $H \asymp \delta^{-1/2}$ and allocating failure probability $\delta/2$ to each event yields

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \lesssim \frac{\log^6(1/\delta)}{\sqrt{\delta T}} \qquad \text{with probability at least } 1-\delta.$$

For this schedule, the $\delta^{-1/2}$ loss enters when the random denominator is removed.

The two moment recursions respond differently to an isolated large gradient. The second moment records its squared magnitude, while the first moment retains its signed direction. The lower-bound constructions in Section C exploit this interaction through rare impulses, persistent asymmetric noise, and deterministic oscillating gradients. They cover the entire first-moment domain with fixed β_2 and $\epsilon > 0$. Their uniform normalization permits a direct comparison with Muon's high-probability guarantee on every selected instance.

3.2 High-probability guarantees for Muon

Large-batch NSGD reduces heavy-tailed noise by averaging more samples at each iterate (Hübler et al., 2025). Following Choudhury et al. (2026) and Hübler et al. (2026), we choose a large momentum β for the given budget T while keeping B fixed. Both average gradients: batching across samples at one iterate, and momentum across time over a window of order $1/(1-\beta)$. The stepsize controls gradient drift over this window. Parameters are deterministic and constant within each run.

Theorem 1 (Conditional p-moment momentum Muon)**.** *For every $p \in (1, 2]$, there is a constant $C_p \geq 1$, depending only on p, with the following property. Suppose Assumptions 1 to 3 hold. Given $T \in \mathbb{N}_+$, choose $\eta > 0$ and β satisfying $0 \leq \beta \leq 1 - \frac{1}{T}$, $\quad C_p \sigma_1 \sqrt{r} \left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}} \leq \frac{1}{2}$. Then Algorithm 1 in* exact *mode satisfies, for every $\delta \in (0, 1)$, with probability at least $1-\delta$,*

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla f(X_t)\|_* \leq C_p \Big[\frac{\Delta_1}{\eta T} + \sigma_0 \sqrt{r} \big(\frac{1-\beta}{B}\big)^{\frac{p-1}{p}} + \big(\frac{\sqrt{2Lr\Delta_1}}{(1-\beta)T} + \frac{Lr\eta}{1-\beta}\big) \log \frac{e}{\delta} \Big]. \tag{3}$$

The upper bound on β keeps its memory window within the given budget T. The noise condition in Theorem 1 requires enough averaging to absorb gradient-dependent noise.

Corollary 1 (Fixed-batch iteration rate)**.** *Under the assumptions of Theorem 1, for the given T, choose $\beta = 1 - T^{-\frac{p}{3p-2}}$, $\eta = (1-\beta)^{\frac{2p-1}{p}}$, with β satisfying $C_p\sigma_1\sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}} \le \frac{1}{2}$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* \le C_p T^{-\frac{p-1}{3p-2}}\Big[\Delta_1 + \sigma_0\sqrt{r}\,B^{-\frac{p-1}{p}} + \big(\sqrt{2Lr\Delta_1}T^{-\frac{p-1}{3p-2}} + Lr\big)\log\frac{e}{\delta}\Big]. \quad (4)$$

In particular,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} = \mathcal{O}\left(T^{-\frac{p-1}{3p-2}}\left(1+\log\frac{1}{\delta}\right)\right),$$

where the hidden constant depends only on $p, B, L, r, \Delta_1, \sigma_0$.

Practical Muon with NS5. Practical Muon replaces the exact polar factor by the five-step Newton–Schulz direction in Algorithm 1. Its finite-time guarantee uses a condition on the resulting normalization error.

Assumption 4 (NS5 approximation error for practical Muon)**.** *Fix deterministic $1 \le K < 2$ and $\tau \ge 0$. For every t, let $P_t := \mathsf{NS}_5(N_t;\varepsilon_{\mathrm{NS}})$. Almost surely, $\|P_t\|_{\mathrm{op}} \le K$, and either $\|N_t\|_* \le \tau$ or $\langle N_t, P_t\rangle \ge (2-K)\|N_t\|_*$. Any randomness used by the normalization routine is included in the filtration, and K and τ do not depend on the confidence level.*

The operator-norm condition controls the update size, while the alignment condition retains a $(2-K)$-fraction of exact polar support outside the threshold τ. This pathwise model is adapted from Qian & Richtárik (2026) and quantifies normalization of the momentum numerator N_t into P_t. The recursion and Assumption 3 control the momentum error.

Theorem 2 (Conditional p-moment NS5 momentum Muon)**.** *The constant C_p in Theorem 1 can be chosen so that the following also holds. Suppose Assumptions 1 to 4 hold. Given $T \in \mathbb{N}_+$, choose $\eta > 0$ and β satisfying $0 \le \beta \le 1 - \frac{1}{T}$, $\quad C_p\sigma_1\sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}} \le \frac{2-K}{2}$. Then Algorithm 1 in $\mathsf{NS5}$ mode satisfies, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* \le \frac{C_p}{2-K}\Big[\frac{\Delta_1}{\eta T} + \sigma_0\sqrt{r}\big(\frac{1-\beta}{B}\big)^{\frac{p-1}{p}} + \frac{\sqrt{2Lr\Delta_1} + LrK\eta T}{(1-\beta)T}\log\frac{e}{\delta} + \tau\Big]. \quad (5)$$

Corollary 2 (Fixed-batch NS5 iteration rate)**.** *Under the assumptions of Theorem 2, for the given T, choose $\beta = 1 - T^{-\frac{p}{3p-2}}$, $\eta = (1-\beta)^{\frac{2p-1}{p}}$, with β satisfying $C_p\sigma_1\sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}} \le \frac{2-K}{2}$. Assume that Assumption 4 holds with the same $1 \le K < 2$ and $\tau \ge 0$ across these choices. Then, with probability at least $1-\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} = \mathcal{O}\left(T^{-\frac{p-1}{3p-2}}\left(1+\log\frac{1}{\delta}\right) + \tau\right),$$

where the hidden constant depends only on $p, B, K, L, r, \Delta_1, \sigma_0$.

The threshold contributes an additive accuracy floor. A fixed positive τ therefore gives a neighborhood guarantee, whereas $\tau = 0$ retains convergence to stationarity at each fixed confidence level.

3.3 Adam's confidence lower bound under first-moment tuning

We allow Adam's first-moment parameter to range over $[0,1)$ and its constant learning rate to range over $(0,\infty)$. Both are selected before sampling and held constant within a run. The second-moment parameter, positive denominator offset, and initial moments are fixed: $\beta_2 \in (0,1)$, $\epsilon > 0$, $m_0 = 0$, $v_0 = \nu\mathbf{1}$, $\nu \ge 0$. The lower bound is uniform over the first-moment and learning-rate choices on a common class of smooth objectives and conditional p-moment oracles.

Theorem 3 (Uniform Adam confidence lower bound). *Fix $p \in (1,2]$, $B \in \mathbb{N}_+$ and β_2, ϵ, ν as above. Let $a = \frac{p-1}{3p-2}$, $0 < q < \frac{p-1}{2p(3p-2)}$. There exist $c > 0$ and $\bar{\delta} \in (0, 1/16]$, depending only on $p, B, \beta_2, \epsilon, \nu, q$, with the following property. For every $T \in \mathbb{N}_+$, $\beta_1 \in [0,1)$, $\gamma > 0$, and $0 < \delta < \bar{\delta}$, there is a preselected instance $\mathcal{I}_{T,\delta,\beta_1,\gamma}$ on $\mathbb{R}^{1\times 2}$ on which Adam satisfies*

$$\mathbb{P}\left\{ \frac{1}{T}\sum_{t=1}^{T} \|\nabla f(X_t)\|_{\mathrm{F}} \geq c \min\{1, T^{-a}\delta^{-q}\} \right\} > \delta. \tag{6}$$

Every instance satisfies Assumptions 1 to 3 with the common bounds $L = 1$, $f_\star = 0$, $\Delta_1 = 1$, $(\sigma_0, \sigma_1) = (1, 0)$, $r = 1$. The constants $c, \bar{\delta}$ are independent of $T, \delta, \beta_1, \gamma$ and of the selected instance.

The theorem covers every horizon-dependent rule $\beta_1(T) \in [0,1)$ chosen independently of confidence. Its pointwise quantifiers also permit β_1 and γ to be selected using δ, with β_2, ϵ, ν fixed. Each instance is chosen before oracle sampling; its dimension, smoothness, initial gap, and noise budget remain uniform. For $p = 2$, one may take $q = 1/32$, obtaining $c\min\{1, T^{-1/4}\delta^{-1/32}\}$.

Comparison with Muon. For the same budget T, use the momentum and stepsize in Corollary 1 and run Muon on the same selected objective, initialization, oracle law, and batch size. Frobenius and nuclear norms coincide on row matrices. Corollary 1 therefore gives $T^{-1}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} \leq C_{\mathrm{M}}(1+\log(1/\delta))T^{-a}$ with probability at least $1-\delta$, where $C_{\mathrm{M}} = C_p\big(2+\sqrt{2}+B^{-(p-1)/p}\big)$ depends only on p, B. In particular, changing the hard instance does not change this comparison constant. On the unsaturated range $T^{-a/q} \leq \delta < \bar{\delta}$, the ratio of the two rate thresholds is $\frac{c}{C_{\mathrm{M}}}\frac{\delta^{-q}}{1+\log(1/\delta)}$. The bounds have the same horizon exponent and polynomial versus logarithmic confidence dependence. For example, taking $\delta = T^{-u}$ with any fixed $0 < u < a/q$ makes the ratio grow as $T^{uq}/(1+u\log T)$. The full construction and proof appear in Section C.

4 Proof Sketch

We sketch the proof mechanisms behind the three main theorems; complete arguments appear in Sections B.2 and C.

4.1 Muon upper bounds

Step 1 (Descent and alignment). Write $g_t = \nabla f(X_t)$ and define the alignment deficit $\varepsilon_t = \|g_t\|_* - \langle g_t, \mathrm{polar}(N_t)\rangle$. Smoothness and $\|\mathrm{polar}(N_t)\|_{\mathrm{F}}^2 \leq r$ give the descent inequality; telescoping yields

$$f(X_{t+1}) \leq f(X_t) - \eta\|g_t\|_* + \eta\varepsilon_t + \frac{Lr\eta^2}{2}, \quad \frac{1}{T}\sum_{t=1}^{T}\|g_t\|_* \leq \frac{\Delta_1}{\eta T} + \frac{Lr\eta}{2} + \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t.$$

Polar duality bounds the deficit by a capped tracking error:

$$0 \leq \varepsilon_t \leq 2\min\{\|g_t\|_*, \|N_t - g_t\|_*\}.$$

Step 2 (Momentum tracking). The tracking recurrence (for $t \geq 2$) and Nesterov identity are

$$U_t - g_t = \beta(U_{t-1} - g_{t-1}) + \beta(g_{t-1} - g_t) + (1-\beta)(M_t - g_t),$$
$$N_t - g_t = \beta(U_t - g_t) + (1-\beta)(M_t - g_t).$$

They separate contraction, gradient drift, and centered noise. The choice $\beta \leq 1 - 1/T$ permits blocks of length $\ell_k \asymp \ell \asymp (1-\beta)^{-1}$, merging the final remainder so that $\beta^{\ell_k} \leq 1/2$. For a block starting at s_k, conditional symmetrization and a martingale maximal inequality yield

$$Z_k := \max_{1\leq l\leq \ell_k}\left\|(1-\beta)\sum_{j=1}^{l}\beta^{l-j}(M_{s_k+j-1} - g_{s_k+j-1})\right\|_*,$$

$$\mathbb{E}[Z_k \mid \mathscr{F}_{s_k-1}] \leq u_k \lesssim_p \sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}}\left(\sigma_0 + \sigma_1\|g_{s_k}\|_{\mathrm{F}} + \sigma_1 L\sqrt{r}\,\eta\ell\right),$$

where u_k is the predictable bound in Lemma B.5. This combines batch and momentum averaging.

Step 3 (Logarithmic concentration). Smoothness and bounded updates give $\|g_t\|_* \le H := \sqrt{2Lr\Delta_1} + Lr\eta T$. Applying a logarithmic potential to the contracting block errors yields

$$\sum_{t=1}^{T} \varepsilon_t \lesssim \ell\|g_1\|_* + \frac{TLr\eta}{1-\beta} + \ell H \sum_k \log\left(1 + \frac{Z_k}{H}\right).$$

The logarithm measures how long a shock persists in the decaying memory (Lemma B.7). Since $\prod_{k=1}^{j}(H + Z_k)/(H + u_k)$ is a nonnegative supermartingale, Markov's inequality gives, with probability at least $1 - \delta$,

$$H \sum_k \log\left(1 + \frac{Z_k}{H}\right) \le \sum_k u_k + H \log\frac{1}{\delta}.$$

This is the proof's only confidence event (Lemma B.8).

Step 4 (Stationarity bounds). Combining the deficit estimate with the descent bound gives a relative-noise multiple of the same nuclear-gradient average (Lemma B.9); the choice of β absorbs it into the left side, proving Theorem 1. The parameter choice in Section 2.2 balances $1/(\eta T)$, $\eta/(1-\beta)$, and $(1-\beta)^{(p-1)/p}$. Since $\|g_t\|_{\mathrm{F}} \le \|g_t\|_*$, this gives Corollary 1 on the same event. For NS5, the support inequality becomes

$$\langle g_t, P_t\rangle \ge (2 - K)\|g_t\|_* - 2\|N_t - g_t\|_* - 2\tau.$$

The bound $\|P_t\|_{\mathrm{op}} \le K$ adjusts the radius, drift, and smoothness term. Repeating the argument with descent coefficient $2 - K$ gives Theorem 2, with an accuracy floor proportional to $\tau/(2-K)$.

4.2 Adam lower bound

Hard-instance assembly. Put $h = 1 - \beta_1$ and $s = 1/16$. A favorable early impulse of batch amplitude $M = B^{\frac{1}{p}-1}\left(\frac{T}{256\delta}\right)^{\frac{1}{p}}$ occurs with probability greater than δ. The second moment forgets it after

$$k(M) = \left\lceil \frac{\log_+((\nu + M^2)/s^2)}{-\log\beta_2} \right\rceil = \mathcal{O}(\log M)$$

steps as $M \to \infty$, where the hidden constant depends only on β_2, ν. The signed first moment persists for order $1/h$ steps; their overlap drives the iterate across an uphill bank and onto a nonstationary plateau. Deterministic travel and oscillatory gradients cover the complementary step-size and long-memory regimes, persistent asymmetric noise covers shorter memory, and a fixed heavy-tail construction covers the saturated confidence range. These witnesses give the target rate uniformly on every bounded learning-rate interval. For larger γ, deterministic interpolation covers γh small and stationary noise covers γh large. Exactly one normalized instance is selected before sampling for each tuning tuple, which preserves the common class constants and proves Theorem 3.

5 Related Work

Adam and adaptive high-probability analysis. Adam combines temporal first- and coordinatewise second-moment estimates (Kingma & Ba, 2014). Its theory includes counterexamples, long-memory corrections, and nonconvex guarantees for adaptive methods (Reddi et al., 2018; Chen et al., 2019; Zhou et al., 2024; Zhang et al., 2022; Défossez et al., 2022; Ward et al., 2019). High-probability analyses study different oracle models and stationarity criteria (Hong & Lin, 2023; Kavis et al., 2022; Jin et al., 2026). Our lower bound controls the first-power gradient score uniformly over all first-moment parameters and positive constant learning rates, with a fixed second-moment parameter and positive offset. A normalized two-dimensional family gives polynomial confidence dependence at the time exponent of the Muon upper guarantee.

Polar descent and Muon. The polar factor is classical (Higham, 1986), and recent work connects matrix orthogonalization to duality maps and non-Euclidean trust regions (Bernstein & Newhouse, 2024; Kovalev, 2025). Direct nuclear descent, stochastic Muon, momentum, and approximate polar computation have important precedents (Chang et al., 2025; Zhang et al., 2025; Kim & Oh, 2026;

Shulgin et al., 2025; Choudhury et al., 2026; Do et al., 2026). The present contribution gives a unified finite-time high-probability closure for the exact canonical recursion and a companion guarantee for the practical NS5 branch under Assumption 4. The NS5 guarantee applies to trajectories satisfying this pathwise normalization-error condition.

Heavy-tailed normalized methods. Clipping, sign, and normalization support high-probability guarantees under finite moments (Cutkosky & Mehta, 2021; Nguyen et al., 2023; Li & Liu, 2023; Hübler et al., 2025; Kornilov et al., 2025), while general nonlinear mappings and heavy-tailed Muon are studied under other stationarity and convergence notions (Jakovetić et al., 2023; Armacki et al., 2025; Choudhury et al., 2026; Iiduka, 2026; Hübler et al., 2026; Yu et al., 2026). Our results combine conditional low-order batching, pre-polar momentum, direct nuclear stationarity, and logarithmic confidence dependence, with batch size and iteration count explicit in the bounds.

6 Experiments

We examine normalization and momentum on Iris and native gradient-noise tails in Qwen3-0.6B training. Section A presents the complete experimental results, figures, optimizer settings, and sampling protocols.

Controlled heavy-tailed noise. On Iris softmax regression, we vary a mean-one Pareto radial multiplier and compare 5,000 coupled trajectories per tail index, with learning rates selected on clean validation data and then fixed. At tail index $\alpha = 1.1$, the empirical p99 average-gradient score is 0.0887 for Muon, 0.2418 for SGD, and 0.1823–1.9458 across six Adam learning rates with comparable clean-validation loss. Exact-polar Muon with $\beta = 0$ retains its clean-oracle trajectory. Confidence profiles cover $\delta = 0.2$ to 0.01 with order-statistic intervals. Under directional noise, momentum ablations with a normalized-SGD control reveal a finite-horizon tradeoff between median and upper-tail performance (Section A.1).

Language-model training and matrix-gradient tails. Starting from a shared random initialization of Qwen3-0.6B, Adam, AdamW, and practical Nesterov/NS5 Muon use the same batch sequence and a budget of 20.48 million input tokens. After 1,250 updates, Muon achieves development NLL 5.0501, compared with 5.2440 for Adam and 5.2230 for AdamW. At the trained checkpoints, exploratory generalized Pareto fits yield tail-index estimates below two at all four thresholds for four of the 24 state–matrix pairs. Muon's layer-13 query projection gives estimates 1.724, 1.706, 1.380, and 1.675 at discovery quantiles 0.90, 0.925, 0.95, and 0.975, respectively. These matrix-level tail profiles accompany Muon's lower development loss in the matched run. Sections A.2 and A.3 provide the training curves, full tail diagnostics, and measurement procedure.

7 Conclusion

For $p \in (1, 2]$, tuning momentum to the given budget gives exact-polar Muon the fixed-batch rate $\mathcal{O}(T^{-(p-1)/(3p-2)}(1 + \log(1/\delta)))$, with an NS5 counterpart under its approximation-error condition. For Adam with fixed second-moment parameter and positive offset, a uniformly normalized two-dimensional family yields $c \min\{1, T^{-(p-1)/(3p-2)}\delta^{-q}\}$, $0 < q < (p-1)/(2p(3p-2))$, uniformly over first-moment parameters and positive constant learning rates. On these instances, the bounds match in time and separate in confidence even under budget-dependent Adam first-moment tuning.

AI USE STATEMENT

Generative AI tools assisted with language polishing and LaTeX typesetting. All mathematical derivations were carried out and all proofs were reviewed by the human authors; the proofs were also formally verified in Lean.

ETHICS STATEMENT

This work studies optimization theory and controlled benchmark experiments. It reports convergence guarantees and aggregate loss and gradient statistics, with experiments on UCI Iris and a randomly initialized Qwen3-0.6B model.

REPRODUCIBILITY STATEMENT

The assumptions are stated in Section 2.2, and complete proofs appear in Sections B and C. Section A collects the experimental results and protocols, including the Iris data split, radial and directional oracles, confidence intervals, momentum ablations, Qwen corpus preparation, optimizer settings, learning-rate selection, fixed-checkpoint sampling, and tail-fitting protocol. The accompanying package contains Iris data splits, per-trajectory outcomes, simulation and plotting code, together with Qwen training and sampling code, pinned data revisions, and numerical tables reproducing the figures.

REFERENCES

Aleksandar Armacki, Shuhua Yu, Pranay Sharma, Gauri Joshi, Dragana Bajović, Dušan Jakovetić, and Soummya Kar. High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 1774–1782. PMLR, 2025.

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 4780–4788. PMLR, 2024.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. *arXiv preprint arXiv:2410.21265*, 2024.

Da Chang, Yongxiang Liu, and Ganzhao Yuan. On the convergence of Muon and beyond. *arXiv preprint arXiv:2509.15816*, 2025.

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.

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026.

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. ISSN 2835-8856.

Thang Do, Steffen Dereich, and Arnulf Jentzen. On MUON optimization: From non-convergence to an error analysis with Polar Express and the Newton–Schulz polynomial from implementations. *arXiv preprint arXiv:2608.04607*, 2026.

Ronald A. Fisher. Iris. UCI Machine Learning Repository, 1936.

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

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pp. 3964–3975. PMLR, 2021.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *International Conference on Machine Learning*, pp. 4262–4274. PMLR, 2021.

Yusu Hong and Junhong Lin. High probability convergence of Adam under unbounded gradients and affine variance noise. *arXiv preprint arXiv:2311.02000*, 2023.

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.

Florian Hübler, Thomas Pethick, and Suvrit Sra. Free heavy-tailed lunch for Muon: A theoretical justification of empirical success. *arXiv preprint arXiv:2606.14560*, 2026.

Hideaki Iiduka. Muon converges under heavy-tailed noise: Nonconvex Hölder-smooth empirical risk minimization. *arXiv preprint arXiv:2603.15059*, 2026.

Dušan Jakovetić, Dragana Bajović, Anit Kumar Sahu, Soummya Kar, Nemanja Milošević, and Dušan Stamenković. Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423, 2023. doi: 10.1137/21M145896X.

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

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

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

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, pp. 45672–45718, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nikita Kornilov, Philip Zmushko, Andrei Semenov, Mark Ikonnikov, Alexander Gasnikov, and Alexander Beznosikov. Sign operator for coping with heavy-tailed noise in non-convex optimization: High probability bounds under (L_0, L_1)-smoothness. *arXiv preprint arXiv:2502.07923*, 2025.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why Adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30106–30148, 2024. doi: 10.52202/079017-0948.

Shaojie Li and Yong Liu. High probability analysis for non-convex stochastic optimization with clipping. In *European Conference on Artificial Intelligence*, pp. 1406–1413, 2023. doi: 10.3233/FAIA230418.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-SGD under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

Xun Qian and Peter Richtárik. Convergence analysis of Muon-type methods with inexact LMO in the degenerate case. *arXiv preprint arXiv:2606.21581*, 2026.

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

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact Muon update. *arXiv preprint arXiv:2510.19933*, 2025.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827–5837. PMLR, 2019.

Rachel Ward, Xiaoxia Wu, and Léon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686. PMLR, 2019.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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*, 2026.

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.

Minxin Zhang, Yuxuan Liu, and Hayden Schaeffer. AdaGrad meets Muon: Adaptive stepsizes for orthogonal updates. *arXiv preprint arXiv:2509.02981*, 2025.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, volume 35, pp. 28386–28399, 2022. doi: 10.52202/068431-2058.

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. ISSN 2835-8856.

A Detailed Experiments

This appendix presents the experimental results and protocols summarized in Section 6, including the Qwen matrix-level heavy-tail profiles with fitted exponents below two across all four thresholds.

A.1 Heavy-tailed noise on Iris

The real-data stress test uses the UCI Iris classification data (Fisher, 1936). A three-class softmax model has a parameter matrix $W \in \mathbb{R}^{5\times 3}$ after appending an intercept. A fixed stratified split assigns, per class, 30 examples to training, 10 to validation, and 10 to test; standardization uses training statistics only. The split uses seed 20260923 and a separate permutation within each class. The training objective is mean cross-entropy plus $0.01\|W_{1:4,:}\|_{\mathrm{F}}^2/2$. The data, split indices, standardization statistics, and numerical scripts are included in the experimental package. Computation uses float64 on CPU.

A.1.1 Radial shocks and confidence profiles

At step t, every method receives the controlled radial oracle

$$\widetilde{G}_t = C_{\alpha,t}\nabla F(W_t), \qquad C_{\alpha,t} = \frac{\alpha-1}{\alpha}V_{\alpha,t}^{-\frac{1}{\alpha}}, \qquad V_{\alpha,t} \sim \mathrm{Unif}(0,1).$$

Thus $C_{\alpha,t} > 0$, $\mathbb{E}C_{\alpha,t} = 1$, and $\mathbb{E}C_{\alpha,t}^q < \infty$ exactly when $q < \alpha$. Increasing α makes the stochastic gradient more concentrated; for $\alpha \le 2$ its second moment is infinite. The grid is $\alpha \in \{1.1, 1.2, 1.4, 1.6, 1.8, 2.5, 4, \infty\}$, where $\alpha = \infty$ denotes $C_t \equiv 1$. The theoretical moment order p is distinct from the Pareto tail index α: for $1 < \alpha \le 2$, any $1 < p < \alpha$ is admissible, while $\alpha > 2$ admits $p = 2$. Thus the full grid lies within the noise-moment range of the theory; $\alpha = \infty$ is the noiseless reference.

We compare canonical exact-polar Muon with $\beta = 0$, Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, and SGD for $T = 300$ updates. All start from $W_1 = 0$ and use the same training data and multiplier sequence within a coupled trial. Learning rates are selected by final clean-validation NLL from the common grid $\{0.003, 0.01, 0.03, 0.1, 0.3, 1, 2, 3\}$ and held fixed across all tail indices. The selected values are 0.03, 3, and 3 for Muon, Adam, and SGD, respectively; Adam uses $\epsilon = 10^{-8}$. The rank-deficient softmax gradient uses the canonical pseudoinverse polar, retaining only nonzero singular directions. For this Muon configuration, $C_{\alpha,t} > 0$ gives

$$\mathrm{polar}(C_{\alpha,t}\nabla F(W_t)) = \mathrm{polar}(\nabla F(W_t)),$$

so this $\beta = 0$ Muon trajectory is pathwise invariant to the radial shock. This radial scale invariance is shared by exactly scale-invariant directions, including normalized SGD. It provides a controlled setting for testing the response of each optimizer to changes in noise magnitude.

Each tail index uses 5,000 independent coupled trajectories, with a shared uniform array generated from seed 2026092301. At $\alpha = 1.1$, the empirical p99 average-gradient scores are 0.0887 for Muon, 1.9458 for Adam, and 0.2418 for SGD. Their p99 final-gradient norms are 0.00651, 2.9938, and 0.0438, respectively. The corresponding test NLLs are 0.2225, 184.07, and 0.2428, versus clean-oracle values between 0.2225 and 0.2239. Figure 1 shows the tail-index sweep.

To examine sensitivity to the learning-rate choice, we also evaluate every Adam candidate whose clean-validation NLL is within 10^{-3} of the minimum. These six rates, $\{0.03, 0.1, 0.3, 1, 2, 3\}$, give p99 average-gradient scores between 0.1823 and 1.9458 at $\alpha = 1.1$, all above Muon's 0.0887. All candidate results and per-trajectory outcomes are supplied.

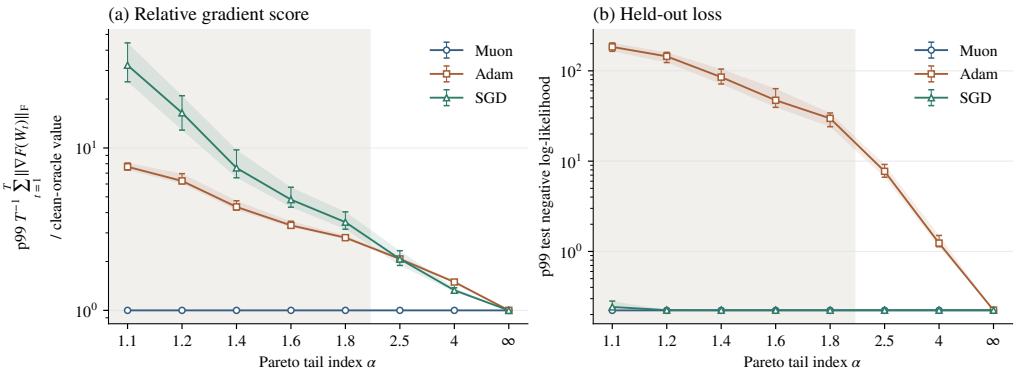

Figure 1: Radial heavy-tail stress test on UCI Iris softmax regression. Left: each method's p99 average clean-gradient score $T^{-1}\sum_{t=1}^{T}\|\nabla F(W_t)\|_{\mathrm{F}}$ divided by its own clean-oracle value. Right: p99 held-out cross-entropy. The shaded range has no finite second moment. Learning rates are selected on clean validation NLL and held fixed; each point uses 5,000 trajectories. Error bands are pointwise 95% binomial order-statistic intervals. Muon uses $\beta = 0$.

At $\alpha = 1.1$, Figure 2 (left) reports $(1-\delta)$ quantiles of

$$T^{-1}\sum_{t=1}^{T}\|\nabla F(W_t)\|_{\mathrm{F}}, \qquad \delta \in \{0.2, 0.1, 0.05, 0.02, 0.01\}.$$

Pointwise 95% intervals use binomial order statistics from the same 5,000 trajectories, conditional on the fixed data split and chosen learning rates. Muon's quantile is constant because the positive radial multiplier cancels pathwise; Adam's and SGD's curves measure their upper-tail sensitivity. These are finite-sample confidence profiles for the specified oracle.

A.1.2 Momentum under directional noise

To vary gradient direction as well as magnitude, we use the same objective and initialization with the additive oracle

$$\widetilde{G}_t = \nabla F(W_t) + \sigma C_{1.5,t} D_t, \qquad \sigma = \|\nabla F(0)\|_{\mathrm{F}}.$$

Here D_t is an independent Gaussian 5×3 matrix, centered across its three columns and normalized to unit Frobenius norm. Symmetry makes the noise mean zero; its pth moment is finite for $1 < p < 1.5$. We compare $\beta \in \{0, 0.5, 0.9, 0.99\}$ in the Muon recursion and in a normalized-SGD control that uses the same numerator N_t with direction $N_t/\|N_t\|_{\mathrm{F}}$. All configurations share their noise within each of 5,000 trials, using seed 2026092303 and $T = 300$.

Step lengths are selected separately for each method and β from $\{0.003, 0.01, 0.03, 0.1, 0.3\}$ by median final-validation NLL on 100 independent pilot trajectories (seed 2026092302). The grid specifies Frobenius displacement: Muon's learning rate is divided by $\sqrt{2}$ because the centered matrices have rank two almost surely. Both methods select displacements $0.3, 0.3, 0.1, 0.03$ in increasing order of β. We also compare a common displacement of 0.03.

At the common displacement, Muon's median $T^{-1}\sum_{t=1}^{T}\|\nabla F(W_t)\|_{\mathrm{F}}$ decreases from 0.2157 at $\beta = 0$ to 0.1873 at $\beta = 0.9$, while its p99 is smallest at $\beta = 0.5$ (0.2248) and reaches 0.4449 at $\beta = 0.99$. Normalized SGD has the same qualitative pattern. With independently selected step lengths, p99 increases with β for both methods (Figure 2). Thus the median and upper tail favor different memory lengths at this horizon, consistent with the balance between noise averaging and gradient tracking in the upper bound.

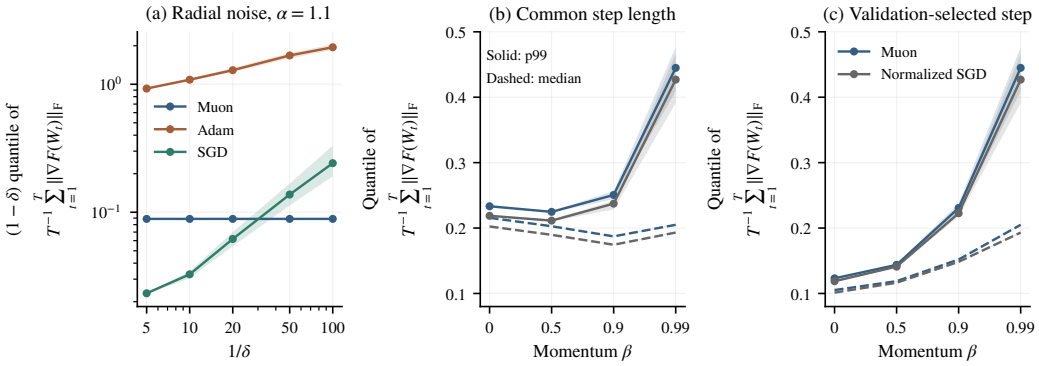

Figure 2: Confidence profiles and momentum on Iris. Left: radial-noise quantiles at $\alpha = 1.1$. Middle and right: median (dashed) and p99 (solid) under additive directional noise, with a common step length and independently selected step lengths, respectively. Each configuration uses 5,000 trajectories; bands are pointwise 95% order-statistic intervals.

A.2 Language-model training and native gradient-noise tails

We complement the controlled Iris experiments with training from random initialization using the Qwen3-0.6B architecture (Yang et al., 2025) (596 million parameters). Adam, AdamW, and Muon share the same initial checkpoint, training-batch sequence, and budget of 1,250 updates, or 20.48 million input tokens. Each batch contains 16 sequences of length 1,024 drawn from a common pool of 8,192 token blocks. Each optimizer receives a six-candidate learning-rate/decay search on development data; all three select a base learning rate of 6×10^{-4}. Muon uses Nesterov momentum 0.95 and five Newton–Schulz steps on attention and MLP projection matrices, with AdamW on the remaining parameters. Section A.3 gives the full training and sampling protocol.

Training performance. Muon attains the lowest endpoint fixed-training and development NLL in this matched run (Table 1). Its development NLL is 0.1940 below Adam and 0.1729 below AdamW at the same token budget. Figure 3 shows the development curves across the common token budget. Muon's best recorded development NLL is 5.0315.

Table 1: Qwen3-0.6B training from a common random initialization. NLL is evaluated on 256 fixed training blocks and 1,024 fixed development blocks; lower is better. Endpoint values use 1,250 updates for every method.

Optimizer	Endpoint training NLL	Endpoint dev NLL	Best recorded dev NLL
Adam	4.9969	5.2440	5.2440
AdamW	4.9454	5.2230	5.2230
Muon	**3.4989**	**5.0501**	**5.0315**

Native matrix-gradient noise. At the common initialization and each final checkpoint, we measure batch-16 gradient radii around the full-pool mean for query and MLP down-projection matrices in layers 0, 13, and 27. Measurement precedes clipping and optimizer transformations. Each fixed state uses 2,048 threshold-discovery batches and 8,192 independent confirmation batches, with sampling plans shared across states.

Heavy-tail signatures across thresholds. We examine the excess radius $R - u$ using generalized Pareto distribution (GPD) fits with free scale and shape ξ. For $\widehat{\xi} > 0$, the fitted survival exponent is $\widehat{\alpha} = 1/\widehat{\xi}$. Four of the 24 state–matrix pairs yield $\widehat{\alpha} < 2$ at every discovery quantile $q \in \{0.90, 0.925, 0.95, 0.975\}$: Adam's layer-27 query projection, AdamW's layer-0 and layer-13 down projections, and Muon's layer-13 query projection (Figure 3). All four occur at trained checkpoints.

For Muon's layer-13 query projection, the four fitted exponents are 1.724, 1.706, 1.380, and 1.675, respectively. These profiles connect the optimizer comparison to native matrix-level tail structure over the observed range. Section A.3 gives the fitting protocol and tabulates all four profiles selected by the same all-threshold criterion.

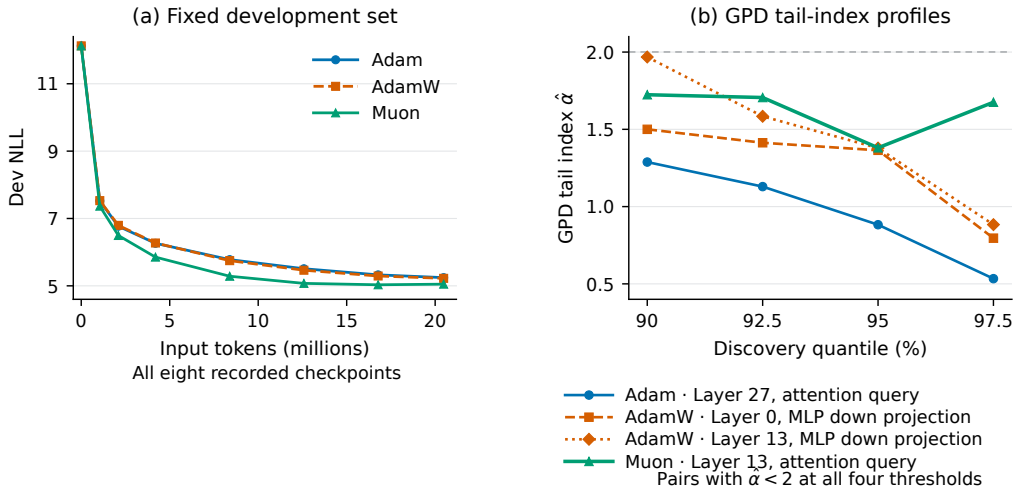

Figure 3: Qwen3-0.6B training and native gradient-noise diagnostics. Left: development NLL at all recorded checkpoints for the matched training runs. Right: GPD tail-index profiles for all four of the 24 state–matrix pairs whose fitted exponents remain below two at every threshold. Each profile uses the same confirmation sample across the four discovery quantiles. The horizontal line marks $\alpha = 2$; the selected profiles are tabulated in Section A.3.

A.3 Qwen training and native gradient-noise protocol

A.3.1 Training and sampling protocol

Model, data, and budget. We initialize the Qwen3-0.6B decoder-only architecture from its configuration, with 596,049,920 parameters and initialization seed 2026091501. Adam, AdamW, and Muon start from the same saved random weights and use the same complete sequence of training-block indices (sampling seed 2026091502). The corpus is FineWeb-Edu's 10-billion-token sample, tokenized with the Qwen3-0.6B-Base tokenizer. Raw documents are deduplicated by source identifier and SHA-256 of their unnormalized UTF-8 text. A deterministic document-level hash, using seed 2026091001, assigns residues modulo 11 to training (0–7), development (8), or an unused holdout (9–10). Tokenization adds no special tokens except one terminal EOS. Documents shorter than 1,024 tokens are excluded; an independently hashed starting position selects one contiguous block per retained document. The pools contain 8,192 training, 1,024 development, and 2,048 unused holdout blocks. Dataset and tokenizer revision identifiers, array checksums, and preprocessing code are supplied in the experimental package. Each of the 1,250 updates uses batch size 16 and microbatch size 1, giving 16,384 input tokens and 16,368 next-token targets per update, or 20,480,000 input tokens and 20,460,000 targets in total. Parameters are stored in FP32; forward passes use BF16 autocast and scaled dot-product attention, with TF32 disabled. Each run uses one CUDA GPU with Python 3.12.3, PyTorch 2.7.1, Transformers 5.15.1, and NumPy 2.2.6. Training uses global gradient-norm clipping at 1.

Optimizer configuration and selection. The common hidden-parameter group comprises the attention and MLP projection matrices: 196 tensors and 440,401,920 parameters. The remaining 114 tensors contain 155,648,000 parameters, including embeddings, the tied output head, and normalization parameters, and receive no weight decay. Adam and AdamW use $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. Muon uses momentum 0.95, Nesterov acceleration, five Newton–Schulz iterations, and update scaling $0.2\sqrt{\max(m, n)}$ for an $m \times n$ matrix; its remaining parameters use AdamW with

the same learning rate and the above AdamW coefficients. For each method, six candidates combine learning rates $\{10^{-4}, 3 \cdot 10^{-4}, 6 \cdot 10^{-4}\}$ with hidden-group decay $\{0, 10^{-4}\}$ for Adam or $\{0, 0.1\}$ for AdamW and Muon. Each candidate runs for 128 updates and is selected by endpoint NLL on 256 common development blocks. The selected learning rate is $6 \cdot 10^{-4}$ for all three methods; selected hidden-group decay is 10^{-4} for Adam (coupled L_2) and 0.1 for AdamW and Muon (decoupled). The full run uses 63 warmup updates (5%) followed by cosine decay to 0.1 times the base learning rate. Evaluation uses 256 fixed training blocks and 1,024 fixed development blocks. Reported NLL is natural-log cross-entropy, excluding Adam's L_2 penalty; the development set also serves model development and selection.

Fixed-state native gradient measurements. We examine the common initial weights and each method's weights after update 1,250. At each state, we measure the attention query and MLP down-projection matrices in layers 0, 13, and 27 (indexed from zero), giving 24 state–matrix pairs. For each selected matrix, the measured radius is

$$R = \left\| G_{\mathrm{native},B16}(W) - C_{\mathrm{fixed}}(W) \right\|_{\mathrm{F}}.$$

Here $G_{\mathrm{native},B16}$ is the native batch-16 gradient computed with microbatch size 1, before clipping or any optimizer transformation. The weights remain fixed throughout measurement. The center C_{fixed} is the deterministic mean of the 8,192 individual block gradients, accumulated in a fixed order in FP64. For each state, sampling with replacement from the same pool produces 2,048 threshold-discovery batches and 8,192 independent confirmation batches, using seeds 2026091702 and 2026091703, respectively. The four states share the batch-index schedules; pilot sampling is excluded. This diagnostic study follows training and pilot exploration; its sampling plan was fixed before these measurements.

A.3.2 Generalized Pareto profiles across thresholds

For each state–matrix pair, the independent discovery sample sets the threshold u at quantiles $q \in \{0.90, 0.925, 0.95, 0.975\}$. Using confirmation radii strictly above u, we fit a generalized Pareto distribution (GPD) to the excesses $Y = R - u$:

$$\Pr(Y > y \mid R > u) = \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \qquad \sigma > 0, \quad 1 + \xi y/\sigma > 0,$$

with the exponential limit at $\xi = 0$. The GPD has a free scale σ and shape ξ. Fitting requires at least 64 exceedances and constrains $\xi \geq -0.49$; $\widehat{\alpha} = 1/\widehat{\xi}$ is reported only when $\widehat{\xi} > 0$. Across the 24 state–matrix pairs and four thresholds, the 96 fits use 146–928 exceedances each. Each threshold profile uses the same confirmation sample at fixed checkpoint weights under finite-pool resampling.

There are 37 fitted exponents below two: 14 for Adam, 14 for AdamW, and 9 for Muon, all at trained checkpoints and all with successful numerical fits. We select the state–matrix pairs whose fitted exponents remain below two at every threshold. Exactly four of the 24 pairs satisfy this criterion: Adam's layer-27 query projection, AdamW's layer-0 and layer-13 down projections, and Muon's layer-13 query projection. Table 2 reports their fitted exponents, and Figure 3 shows their threshold profiles. These selected profiles exhibit matrix-level heavy-tail signatures across the full threshold grid over the observed radius range.

Table 2: The four of 24 state–matrix pairs with fitted GPD exponents $\widehat{\alpha} < 2$ at all four discovery quantiles. The table includes exactly the profiles satisfying this selection criterion. Each row uses the same confirmation sample across thresholds.

			GPD $\widehat{\alpha}$ at discovery quantile q			
State	Layer	Projection	0.90	0.925	0.95	0.975
Adam	27	Attention query	**1.288**	**1.129**	**0.883**	**0.534**
AdamW	0	MLP down	**1.500**	**1.413**	**1.364**	**0.796**
AdamW	13	MLP down	**1.968**	**1.584**	**1.379**	**0.883**
Muon	13	Attention query	**1.724**	**1.706**	**1.380**	**1.675**

B Proofs of Lemmas and Theorems

This appendix collects the formal proofs of all stated results.

B.1 Smoothness inequality

Under Assumption 2, for all $X, Y \in \mathbb{R}^{m\times n}$,

$$f(Y) \le f(X) + \langle \nabla f(X), Y - X\rangle + \frac{L}{2}\|Y - X\|_{\mathrm{F}}^2. \tag{B.1}$$

Indeed, let $D := Y - X$. Applying the fundamental theorem of calculus along the segment $X + sD$ and then Equation 2 gives

$$\begin{aligned} f(Y) - f(X) - \langle \nabla f(X), D\rangle &= \int_0^1 \langle \nabla f(X + sD) - \nabla f(X), D\rangle \, ds \\ &\le \int_0^1 Ls\|D\|_{\mathrm{F}}^2 \, ds = \frac{L}{2}\|D\|_{\mathrm{F}}^2. \end{aligned}$$

B.2 Unified Conditional p-Moment Momentum Result

This subsection proves Theorem 1 and Corollary 1. Put $r := \min\{m, n\}$. For a given iteration budget $T \in \mathbb{N}_+$, fix $p \in (1, 2]$, choose $0 \le \beta \le 1 - 1/T$, and use a deterministic constant step $\eta > 0$. We work under Assumptions 1 to 3. The initialization, batch size B, and all algorithmic parameters are deterministic. Before batch t is drawn, X_t and U_{t-1} are $\mathscr{F}_{t-1}$-measurable. Conditional on $\mathscr{F}_{t-1}$, the matrices $G_{t,1}, \dots, G_{t,B}$ are independent; their conditional unbiasedness and moment control are given by Assumption 3. Define

$$M_t := \frac{1}{B}\sum_{j=1}^{B} G_{t,j}.$$

No unconditional independence across times is used.

B.2.1 Polar geometry and conditional batch estimation

For a compact SVD $M = U_s\Sigma_s V_s^\top$, the canonical partial polar is

$$\mathrm{polar}(M) := M\big((M^\top M)^\dagger\big)^{\frac{1}{2}} = U_s V_s^\top, \qquad \mathrm{polar}(0) := 0.$$

The pseudoinverse formula makes this map Borel measurable, including at rank changes, so applying it to an adapted random matrix preserves measurability.

Lemma B.1 (Canonical partial-polar identities)**.** *For every $M \in \mathbb{R}^{m\times n}$,*

$$\|\mathrm{polar}(M)\|_{\mathrm{op}} \le 1, \qquad \|\mathrm{polar}(M)\|_{\mathrm{F}}^2 = \mathrm{rank}(M) \le r, \qquad \langle M, \mathrm{polar}(M)\rangle = \|M\|_*. \tag{B.2}$$

Moreover, for every iteration t and compatible matrix M,

$$\langle \nabla f(X_t), \mathrm{polar}(M)\rangle \ge \|\nabla f(X_t)\|_* - 2\|M - \nabla f(X_t)\|_*. \tag{B.3}$$

For every scalar $a > 0$, $\mathrm{polar}(aM) = \mathrm{polar}(M)$.

Proof. The nonzero singular values of $U_s V_s^\top$ are all one, which proves the first two assertions in equation B.2. Direct multiplication gives $\langle M, U_s V_s^\top\rangle = \mathrm{tr}(\Sigma_s) = \|M\|_*$. For any matrix in the operator-norm unit ball, each bilinear pairing between the unit singular vectors u_i, v_i has magnitude at most one. Its inner product with M is therefore at most $\sum_{i=1}^{s}\sigma_i(M) = \|M\|_*$. Thus $\mathrm{polar}(M)$ is a support maximizer over that ball. Consequently,

$$\begin{aligned} \langle \nabla f(X_t), \mathrm{polar}(M)\rangle &= \langle M, \mathrm{polar}(M)\rangle + \langle \nabla f(X_t) - M, \mathrm{polar}(M)\rangle \\ &\ge \|M\|_* - \|\nabla f(X_t) - M\|_* \ge \|\nabla f(X_t)\|_* - 2\|M - \nabla f(X_t)\|_*, \end{aligned}$$

which is equation B.3. Positive scalar invariance follows immediately from the compact SVD, including when $M = 0$. □

Lemma B.2 (Conditional type-p batch mean)**.** *Let $\mathscr{G}$ be a sigma-field and let $Z_1, \dots, Z_B$ be conditionally independent, conditionally mean-zero random matrices. For $p \in (1, 2]$,*

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{j=1}^{B} Z_j\right\|_{\mathrm{F}}^{p} \middle| \mathscr{G}\right] \le \frac{2^p}{B^p}\sum_{j=1}^{B}\mathbb{E}[\|Z_j\|_{\mathrm{F}}^{p} \mid \mathscr{G}] \tag{B.4}$$

In particular, under Assumption 3,

$$\mathbb{E}[\|M_t - \nabla f(X_t)\|_* \mid \mathscr{F}_{t-1}] \le 2\sqrt{r}\, B^{-\frac{p-1}{p}}\,(\sigma_0 + \sigma_1\|\nabla f(X_t)\|_*) \tag{B.5}$$

Proof. All calculations may be read under a regular conditional law given $\mathscr{G}$. When $p = 2$, conditional centering and independence eliminate the cross Frobenius inner products, so

$$\mathbb{E}\left\|\sum_{j=1}^{B} Z_j\right\|_{\mathrm{F}}^{2} = \sum_{j=1}^{B}\mathbb{E}\|Z_j\|_{\mathrm{F}}^{2}.$$

This proves equation B.4 for $p = 2$.

For $1 < p < 2$, on an enlarged conditional probability space, let $Z'_1, \dots, Z'_B$ be an independent copy of $Z_1, \dots, Z_B$. Conditional Jensen gives

$$\mathbb{E}\left\|\sum_{j=1}^{B} Z_j\right\|_{\mathrm{F}}^{p} = \mathbb{E}\left\|\mathbb{E}\left[\sum_{j=1}^{B}(Z_j - Z'_j)\middle| Z_1, \dots, Z_B\right]\right\|_{\mathrm{F}}^{p} \le \mathbb{E}\left\|\sum_{j=1}^{B}(Z_j - Z'_j)\right\|_{\mathrm{F}}^{p}. \tag{B.6}$$

Put $Y_j := Z_j - Z'_j$. The Y_j are conditionally independent and symmetric. Given them, independent Rademacher signs ε_j satisfy

$$\mathbb{E}_\varepsilon\left\|\sum_{j=1}^{B}\varepsilon_j Y_j\right\|_{\mathrm{F}}^{p} \le \left(\mathbb{E}_\varepsilon\left\|\sum_{j=1}^{B}\varepsilon_j Y_j\right\|_{\mathrm{F}}^{2}\right)^{\frac{p}{2}} = \left(\sum_{j=1}^{B}\|Y_j\|_{\mathrm{F}}^{2}\right)^{\frac{p}{2}} \le \sum_{j=1}^{B}\|Y_j\|_{\mathrm{F}}^{p}.$$

Symmetry and $\|x - y\|_{\mathrm{F}}^{p} \le 2^{p-1}(\|x\|_{\mathrm{F}}^{p} + \|y\|_{\mathrm{F}}^{p})$ therefore imply

$$\mathbb{E}\left\|\sum_{j=1}^{B}(Z_j - Z'_j)\right\|_{\mathrm{F}}^{p} \le 2^p\sum_{j=1}^{B}\mathbb{E}\|Z_j\|_{\mathrm{F}}^{p}.$$

Together with equation B.6, division by B^p proves equation B.4 for $1 < p < 2$.

Apply equation B.4 conditionally with $Z_j = G_{t,j} - \nabla f(X_t)$. Under Assumption 3,

$$\mathbb{E}[\|M_t - \nabla f(X_t)\|_{\mathrm{F}}^{p} \mid \mathscr{F}_{t-1}] \le 2^p B^{1-p}\,(\sigma_0^p + \sigma_1^p\|\nabla f(X_t)\|_{\mathrm{F}}^{p})\,. \tag{B.7}$$

Finally, $\|Z\|_* \le \sqrt{r}\|Z\|_{\mathrm{F}}$, conditional Jensen, and $(a^p + b^p)^{1/p} \le a + b$ give

$$\mathbb{E}[\|M_t - \nabla f(X_t)\|_* \mid \mathscr{F}_{t-1}] \le \sqrt{r}\,(\mathbb{E}[\|M_t - \nabla f(X_t)\|_{\mathrm{F}}^{p} \mid \mathscr{F}_{t-1}])^{\frac{1}{p}}\,.$$

Substituting equation B.7 into this inequality gives equation B.5. □

B.2.2 Pathwise radius and exact momentum recursions

The algorithm analyzed in the theorem is

$$\begin{aligned} &U_0 = 0, \qquad U_t = \beta U_{t-1} + (1-\beta)M_t, \qquad P_t = \mathrm{polar}(N_t), \\ &N_t = (1-\beta)M_t + \beta U_t = \beta^2 U_{t-1} + (1-\beta^2)M_t, \qquad X_{t+1} = X_t - \eta P_t. \end{aligned} \tag{B.8}$$

Lemma B.3 (Deterministic gradient radius)**.** *For every realization of equation B.8 and every $t \le T$,*

$$\|\nabla f(X_t)\|_* \le \sqrt{2Lr\Delta_1} + Lr\eta(t-1) < \sqrt{2Lr\Delta_1} + Lr\eta T. \tag{B.9}$$

Proof. Apply Equation B.1 with $Y = X - L^{-1}\nabla f(X)$. Since $f(Y) \geq f_\star$,

$$\|\nabla f(X)\|_{\mathrm{F}}^2 \leq 2L\big(f(X) - f_\star\big). \tag{B.10}$$

By Lemma B.1, $\|P_t\|_{\mathrm{F}} \leq \sqrt{r}$. Hence

$$\begin{aligned} f(X_{t+1}) - f_\star &\leq f(X_t) - f_\star + \eta\|\nabla f(X_t)\|_* + \frac{Lr}{2}\eta^2 \\ &\leq f(X_t) - f_\star + \eta\sqrt{2Lr\big(f(X_t) - f_\star\big)} + \frac{Lr}{2}\eta^2 = \left(\sqrt{f(X_t) - f_\star} + \eta\sqrt{\frac{Lr}{2}}\right)^2. \end{aligned}$$

Taking nonnegative square roots and iterating gives

$$\sqrt{f(X_t) - f_\star} \leq \sqrt{\Delta_1} + \sqrt{\frac{Lr}{2}}\,\eta(t-1). \tag{B.11}$$

Combining equation B.11 with equation B.10 and $\|Z\|_* \leq \sqrt{r}\|Z\|_{\mathrm{F}}$ proves equation B.9. □

B.2.3 Conditional martingale cancellation over memory blocks

The stochastic estimate is applied to a weighted sum of centered matrices before taking its norm. This retains cancellation across the memory window.

Lemma B.4 (Conditional Hilbert-space martingale bound)**.** *Let $(\mathscr{G}_j)_{j=0}^J$ be a filtration and let D_j be matrix-valued martingale differences with finite p-th moments, $1 < p \leq 2$. For deterministic real weights a_j,*

$$\mathbb{E}\left[\max_{1\leq l\leq J}\left\|\sum_{j=1}^{l} a_j D_j\right\|_{\mathrm{F}}^p \,\middle|\, \mathscr{G}_0\right] \leq 16\left(\frac{p}{p-1}\right)^p \sum_{j=1}^{J} |a_j|^p \mathbb{E}[\|D_j\|_{\mathrm{F}}^p \mid \mathscr{G}_0]. \tag{B.12}$$

Proof. For vectors in a finite-dimensional Hilbert space, with the derivative term interpreted as zero at $x = 0$,

$$\|x+y\|^p \leq \|x\|^p + p\|x\|^{p-2}\langle x, y\rangle + 16\|y\|^p. \tag{B.13}$$

The case $y = 0$ is immediate. For $y \neq 0$, a dimension-free verification is as follows. If $\|y\| \leq \|x\|/2$, the Hessian of $\|z\|^p$ along the segment has operator norm at most $p(\|x\|/2)^{p-2}$. Taylor's integral remainder is therefore at most $(p/2)\|y\|^p \leq \|y\|^p$. Otherwise $\|x\| < 2\|y\|$, and the remainder is at most $(3^p + p2^{p-1})\|y\|^p \leq 13\|y\|^p$. This proves equation B.13.

Apply this inequality to the successive partial sums and condition on $\mathscr{G}_{j-1}$. The linear term has conditional mean zero. Iteration and the tower property yield

$$\mathbb{E}\left[\left\|\sum_{j=1}^{J} a_j D_j\right\|_{\mathrm{F}}^p \,\middle|\, \mathscr{G}_0\right] \leq 16\sum_{j=1}^{J} |a_j|^p \mathbb{E}[\|D_j\|_{\mathrm{F}}^p \mid \mathscr{G}_0].$$

The partial-sum norms $V_l = \|\sum_{j=1}^{l} a_j D_j\|_{\mathrm{F}}$ form a nonnegative submartingale. Stopping at their first crossing of λ gives

$$\lambda\,\mathbb{P}\left(\max_{l\leq J} V_l \geq \lambda \mid \mathscr{G}_0\right) \leq \mathbb{E}\big[V_J \mathbf{1}_{\{\max_{l\leq J} V_l \geq \lambda\}} \mid \mathscr{G}_0\big].$$

Multiply by $p\lambda^{p-2}$, integrate over $\lambda > 0$, and use conditional Hölder's inequality. The resulting inequality is

$$\left(\mathbb{E}[(\max_{l\leq J} V_l)^p \mid \mathscr{G}_0]\right)^{\frac{1}{p}} \leq \frac{p}{p-1}\left(\mathbb{E}[V_J^p \mid \mathscr{G}_0]\right)^{\frac{1}{p}}.$$

Truncation followed by monotone convergence justifies the integration if the maximal moment has not yet been established. Combining the two bounds proves equation B.12. □

For the given budget T, put $\ell = 1$ if $\beta < 1/2$ and $\ell = \lceil 1/(1-\beta) \rceil$ otherwise. The choice $\beta \le 1 - 1/T$ ensures $\ell \le T$, so $J := \lfloor T/\ell \rfloor \ge 1$. Use $J-1$ blocks of length ℓ followed by one block containing the remaining updates. Write s_k for the first update in block k and ℓ_k for its length. Thus

$$\ell \le \ell_k \le 2\ell, \qquad \beta^{\ell_k} \le \frac{1}{2}, \qquad J\ell \le T, \qquad \ell \le \frac{2}{1-\beta}. \tag{B.14}$$

For each block define

$$Z_k := \max_{1 \le l \le \ell_k} \left\| (1-\beta) \sum_{j=1}^{l} \beta^{l-j} \big(M_{s_k+j-1} - \nabla f(X_{s_k+j-1}) \big) \right\|_*. \tag{B.15}$$

Only the exponent-zero term is present in this sum when a block is a singleton and $\beta = 0$.

Lemma B.5 (Conditional block-shock estimate)**.** *Suppose the update directions satisfy $\|P_t\|_{\mathrm{op}} \le K$ for a deterministic $K \ge 1$. Then*

$$\mathbb{E}[Z_k \mid \mathscr{F}_{s_k-1}] \le \frac{128 e^6 p}{p-1} \sqrt{r} \left(\frac{1-\beta}{B} \right)^{\frac{p-1}{p}} \big(\sigma_0 + \sigma_1 \|\nabla f(X_{s_k})\|_{\mathrm{F}} + 2\sigma_1 L \sqrt{r} K \eta \ell \big). \tag{B.16}$$

Proof. The step bound and gradient Lipschitzness imply, throughout block k,

$$\|\nabla f(X_t) - \nabla f(X_{s_k})\|_{\mathrm{F}} \le 2L\sqrt{r} K \eta \ell.$$

Together with equation B.7, this gives

$$\mathbb{E}[\|M_t - \nabla f(X_t)\|_{\mathrm{F}}^p \mid \mathscr{F}_{t-1}] \le 2^p B^{1-p} \big(\sigma_0 + \sigma_1 \|\nabla f(X_{s_k})\|_{\mathrm{F}} + 2\sigma_1 L \sqrt{r} K \eta \ell \big)^p. \tag{B.17}$$

The right-hand side is measurable before the block begins.

For $\beta \ge 1/2$, the partial weighted noise sums satisfy

$$(1-\beta) \sum_{j=1}^{l} \beta^{l-j} \big(M_{s_k+j-1} - \nabla f(X_{s_k+j-1}) \big) = (1-\beta)\beta^l \sum_{j=1}^{l} \beta^{-j} \big(M_{s_k+j-1} - \nabla f(X_{s_k+j-1}) \big).$$

The last sum is a martingale over the block, conditionally on $\mathscr{F}_{s_k-1}$. Moreover, $-\log \beta \le 2(1-\beta)$ and $(1-\beta)\ell \le 3/2$, so $\beta^{-\ell_k} \le \beta^{-2\ell} \le e^6$. Apply Lemma B.4, the tower property, and $\|A\|_* \le \sqrt{r}\|A\|_{\mathrm{F}}$. Conditional Jensen then bounds the left-hand side of equation B.16 by

$$2 \cdot 16^{\frac{1}{p}} \frac{p}{p-1} e^6 \sqrt{r}\, B^{-\frac{p-1}{p}} (1-\beta)(2\ell)^{\frac{1}{p}} \big(\sigma_0 + \sigma_1 \|\nabla f(X_{s_k})\|_{\mathrm{F}} + 2\sigma_1 L \sqrt{r} K \eta \ell \big).$$

Since $(1-\beta)\ell^{1/p} \le 2^{1/p}(1-\beta)^{(p-1)/p}$, the numerical multiplier is at most $2 \cdot 64^{1/p} e^6 p/(p-1) \le 128 e^6 p/(p-1)$. This proves the claim. For $\beta < 1/2$, each block has one update and $Z_k = (1-\beta)\|M_{s_k} - \nabla f(X_{s_k})\|_*$. The conditional batch estimate and $1 - \beta \le (1-\beta)^{(p-1)/p}$ give the same conclusion. □

B.2.4 Blockwise memory and logarithmic confidence

Lemma B.6 (Geometric memory with an initial residual)**.** *Let $H > 0$, $x_0, \dots, x_J \ge 0$, $d \ge 0$, and $z_k \ge 0$ satisfy $x_k \le x_{k-1}/2 + d + z_k$. Then*

$$\sum_{k=1}^{J} \min\{H, x_{k-1}\} \le 5x_0 + 4Jd + 4H \sum_{k=1}^{J} \log\left(1 + \frac{z_k}{H}\right). \tag{B.18}$$

Proof. Put $\phi(x) = \log(1+x)$. For $0 \le x \le 1$, integration gives $\phi(x) - \phi(x/2) \ge x/4$; for $x \ge 1$, monotonicity gives $\phi(x) - \phi(x/2) \ge \log(4/3) \ge 1/4$. Consequently $\min\{1, x\} \le 4[\phi(x) - \phi(x/2)]$. The recurrence implies

$$\phi\left(\frac{x_k}{H}\right) - \phi\left(\frac{x_{k-1}}{2H}\right) \le \frac{d}{H} + \log\left(1 + \frac{z_k}{H}\right).$$

Sum these inequalities for $1 \le k \le J-1$. The potential telescopes except for the nonnegative final term and the initial term $\phi(x_0/(2H)) \le x_0/H$. Add $\min\{H, x_0\} \le x_0$, and enlarge the nonnegative right-hand side to include block J. This gives equation B.18, also when $J = 1$. □

Lemma B.7 (Cumulative blockwise momentum discrepancy)**.** *Consider the momentum recursion with $U_0 = 0$ and compatible matrices g_t, $1 \le t \le T$. Suppose*

$$\|g_t\|_* \le H, \qquad \|g_t - g_{t-1}\|_* \le d \quad (t \ge 2)$$

for deterministic $H > 0$ and $d \ge 0$. With the blocks in equation B.14 and the shocks in equation B.15, with $\nabla f(X_t)$ replaced by g_t,

$$\sum_{t=1}^{T} \min\{H, \|N_t - g_t\|_*\} \le 20\left[\ell\|g_1\|_* + \frac{Td}{1-\beta} + \ell H \sum_{k=1}^{J} \log\left(1 + \frac{Z_k}{H}\right)\right]. \tag{B.19}$$

Proof. The initialization $U_0 = 0$ gives $U_1 - g_1 = -\beta g_1 + (1-\beta)(M_1 - g_1)$. The recursion also yields

$$\begin{aligned} U_t - g_t &= \beta(U_{t-1} - g_{t-1}) + \beta(g_{t-1} - g_t) + (1-\beta)(M_t - g_t) \quad (t \ge 2), \\ N_t - g_t &= \beta(U_t - g_t) + (1-\beta)(M_t - g_t) \quad (t \ge 1). \end{aligned} \tag{B.20}$$

Iterating from each block's entrance, with $U_0 - g_1 = -g_1$ for the first block, gives

$$\begin{aligned} \|U_{s_k+\ell_k-1} - g_{s_k+\ell_k-1}\|_* &\le \frac{1}{2}\|U_{s_k-1} - g_{\max\{1,s_k-1\}}\|_* + \frac{d}{1-\beta} + Z_k, \\ \|U_t - g_t\|_* &\le \|U_{s_k-1} - g_{\max\{1,s_k-1\}}\|_* + \frac{d}{1-\beta} + Z_k \qquad (s_k \le t < s_k + \ell_k). \end{aligned}$$

The difference between a block's consecutive partial noise sums is $(1-\beta)(M_t - g_t)$ after subtracting β times the previous sum. Thus $\|(1-\beta)(M_t - g_t)\|_* \le 2Z_k$, and

$$\|N_t - g_t\|_* \le \|U_{s_k-1} - g_{\max\{1,s_k-1\}}\|_* + \frac{d}{1-\beta} + 3Z_k.$$

Subadditivity of $x \mapsto \min\{H, x\}$ on nonnegative numbers gives

$$\begin{aligned} \sum_{t=1}^{T} \min\{H, \|N_t - g_t\|_*\} \le 2\ell \sum_{k=1}^{J} \min\{H, \|U_{s_k-1} - g_{\max\{1,s_k-1\}}\|_*\} \\ + \frac{Td}{1-\beta} + 6\ell \sum_{k=1}^{J} \min\{H, Z_k\}. \end{aligned}$$

Apply Lemma B.6 to the block-boundary residuals. Its initial value is $\|U_0 - g_1\|_* = \|g_1\|_*$. Finally use $J\ell \le T$ and

$$\min\{H, z\} \le \frac{H}{\log 2} \log\left(1 + \frac{z}{H}\right).$$

The coefficients of the three terms in equation B.19 are at most 10, 9, and $8 + 6/\log 2 < 20$, respectively. □

Lemma B.8 (Conditional log-shock supermartingale)**.** *Let $H > 0$, let $Z_k \ge 0$ be measurable at the end of deterministic block k, and suppose $\mathbb{E}[Z_k \mid \mathscr{F}_{s_k-1}] \le u_k$, where $u_k \ge 0$ is measurable before that block. For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$H \sum_{k=1}^{J} \log\left(1 + \frac{Z_k}{H}\right) \le \sum_{k=1}^{J} u_k + H \log\frac{1}{\delta}. \tag{B.21}$$

Proof. The products

$$\prod_{k=1}^{j} \frac{H + Z_k}{H + u_k}, \qquad 0 \le j \le J,$$

form a nonnegative supermartingale for the block-end filtration, starting at one. Markov's inequality bounds the final product by $1/\delta$ with probability at least $1-\delta$. Taking logarithms and using $H\log(1 + u/H) \le u$ proves the claim. □

Lemma B.9 (Relative-noise accounting across blocks)**.** *Under the conditions of Lemma B.5, let u_k be its upper bound in equation B.16. Then*

$$\frac{\ell}{T}\sum_{k=1}^{J} u_k \le \frac{128e^6 p}{p-1}\sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}}\left(\sigma_0 + \frac{\sigma_1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* + 4\sigma_1 L\sqrt{r}K\eta\ell\right). \tag{B.22}$$

Proof. For each t in block k, Lipschitzness gives

$$\|\nabla f(X_{s_k})\|_{\mathrm{F}} \le \|\nabla f(X_t)\|_{\mathrm{F}} + 2L\sqrt{r}K\eta\ell.$$

Average within each block and use $\ell/\ell_k \le 1$ to obtain

$$\ell\sum_{k=1}^{J}\|\nabla f(X_{s_k})\|_{\mathrm{F}} \le \sum_{t=1}^{T}\|\nabla f(X_t)\|_* + 2L\sqrt{r}K\eta\ell T.$$

Substitute this into the sum of equation B.16. The other terms use $J\ell/T \le 1$. □

B.2.5 Proofs of the exact-polar guarantees

Proof of Theorem 1. Lemma B.1 gives the realized alignment deficit

$$\epsilon_t := \|\nabla f(X_t)\|_* - \langle \nabla f(X_t), P_t\rangle, \qquad 0 \le \epsilon_t \le 2\min\{\|\nabla f(X_t)\|_*, \|N_t - \nabla f(X_t)\|_*\}.$$

Apply Lemma B.7 with

$$H = \sqrt{2Lr\Delta_1} + Lr\eta T, \qquad d = Lr\eta,$$

which is valid by Lemma B.3. Next apply Lemma B.8 using equation B.16 with $K = 1$, and then Lemma B.9. Consequently, with probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\epsilon_t \le 40\Bigg[&\frac{\ell\sqrt{2Lr\Delta_1}}{T} + \frac{Lr\eta}{1-\beta}\\ &+ \frac{128e^6 p}{p-1}\sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}}\left(\sigma_0 + \frac{\sigma_1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* + 4\sigma_1 L\sqrt{r}\eta\ell\right)\\ &+ \frac{\ell}{T}\left(\sqrt{2Lr\Delta_1} + Lr\eta T\right)\log\frac{1}{\delta}\Bigg].\end{aligned} \tag{B.23}$$

This uses one confidence event, at the block scale.

By Equation B.1,

$$f(X_{t+1}) \le f(X_t) - \eta\|\nabla f(X_t)\|_* + \eta\epsilon_t + \frac{Lr}{2}\eta^2.$$

Sum over t and use $f(X_{T+1}) \ge f_\star$. Division by ηT leaves

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* \le \frac{\Delta_1}{\eta T} + \frac{Lr\eta}{2} + \frac{1}{T}\sum_{t=1}^{T}\epsilon_t.$$

The choice

$$C_p = \frac{2^{14}e^6 p}{p-1}$$

is sufficient in the theorem. It exceeds the coefficient $40\cdot 128e^6 p/(p-1)$ in equation B.23, so the absorption condition makes the coefficient of the relative-gradient sum at most $1/2$. The remaining term containing $\sigma_1 L\sqrt{r}\eta\ell$ is then at most a numerical constant times $Lr\eta/(1-\beta)$, since $\ell \le 2/(1-\beta)$ and $\sqrt{r} \le r$. Move the relative-gradient sum to the left and divide by its remaining coefficient, which is at least $1/2$. Using equation B.14, the resulting coefficients are at most $80\cdot 128e^6 p/(p-1)$ for the additive-noise term and 160 for the radius and drift terms. They are all bounded by the displayed choice of C_p, proving Theorem 1. □

Proof of Corollary 1. For the given $T \in \mathbb{N}_+$, choose $\beta = 1 - T^{-p/(3p-2)}$ and $\eta = (1-\beta)^{(2p-1)/p}$. Since $0 < p/(3p-2) < 1$, $0 \le \beta = 1 - T^{-p/(3p-2)} \le 1 - 1/T$. Also $\eta = T^{-(2p-1)/(3p-2)}$, and hence

$$\frac{1}{\eta T} = \frac{\eta}{1-\beta} = (1-\beta)^{\frac{p-1}{p}} = T^{-\frac{p-1}{3p-2}}, \qquad \frac{1}{(1-\beta)T} = T^{-\frac{2(p-1)}{3p-2}}.$$

The noise condition in Corollary 1 supplies the absorption required by Theorem 1. Substitute the displayed identities into Theorem 1 to obtain the asserted rate. □

B.2.6 NS5 MOMENTUM DIRECTIONS

The NS5 branch changes only the polar-support step and the deterministic radius. Throughout this part, use the deterministic constants $1 \le K < 2$ and $\tau \ge 0$ in Assumption 4 and the final direction returned by Algorithm 1 in NS5 mode.

Proposition B.1 (Thresholded support transfer)**.** *Fix an iteration t, $1 \le K < 2$, and $\tau \ge 0$. Suppose*

$$\|P_t\|_{\mathrm{op}} \le K, \qquad \text{and either} \quad \|N_t\|_* \le \tau \quad \text{or} \quad \langle N_t, P_t\rangle \ge (2-K)\|N_t\|_*. \tag{B.24}$$

Then

$$\langle \nabla f(X_t), P_t\rangle \ge (2-K)\|\nabla f(X_t)\|_* - 2\|N_t - \nabla f(X_t)\|_* - 2\tau.$$

Proof. If $\|N_t\|_* > \tau$, the alignment branch in equation B.24 applies. If $\|N_t\|_* \le \tau$, nuclear–operator duality gives

$$\langle N_t, P_t\rangle \ge -K\|N_t\|_* \ge (2-K)\|N_t\|_* - 2\tau.$$

Thus the last inequality holds in both cases. Consequently,

$$\begin{aligned}\langle \nabla f(X_t), P_t\rangle &\ge (2-K)\|N_t\|_* - 2\tau - K\|\nabla f(X_t) - N_t\|_* \\ &\ge (2-K)\|\nabla f(X_t)\|_* - 2\|N_t - \nabla f(X_t)\|_* - 2\tau,\end{aligned}$$

where the two perturbation coefficients, $2-K$ and K, sum to two. □

Lemma B.10 (NS5 deterministic gradient radius)**.** *Under Assumption 4, every NS5 execution satisfies, for $t \le T$,*

$$\|\nabla f(X_t)\|_* \le \sqrt{2Lr\Delta_1} + LrK\eta(t-1) < \sqrt{2Lr\Delta_1} + LrK\eta T. \tag{B.25}$$

Proof. The gradient–gap inequality equation B.10 is independent of the update direction. Moreover, $\|P_t\|_{\mathrm{F}} \le \sqrt{r}\|P_t\|_{\mathrm{op}} \le \sqrt{r}K$. Equation B.1 and nuclear–operator duality therefore give

$$f(X_{t+1}) - f_\star \le f(X_t) - f_\star + K\eta\|\nabla f(X_t)\|_* + \frac{LrK^2}{2}\eta^2 \le \left(\sqrt{f(X_t) - f_\star} + K\eta\sqrt{\frac{Lr}{2}}\right)^2.$$

Taking square roots and iterating gives

$$\sqrt{f(X_t) - f_\star} \le \sqrt{\Delta_1} + K\sqrt{\frac{Lr}{2}}\,\eta(t-1). \tag{B.26}$$

Combining equation B.26 with equation B.10 and $\|Z\|_* \le \sqrt{r}\|Z\|_{\mathrm{F}}$ proves equation B.25. □

Lemma B.11 (NS5 alignment deficit)**.** *Under Assumption 4, for every deterministic H satisfying $\|\nabla f(X_t)\|_* \le H$, one has*

$$(2-K)\|\nabla f(X_t)\|_* - \langle \nabla f(X_t), P_t\rangle \le 2\min\{H, \|N_t - \nabla f(X_t)\|_*\} + 2\tau. \tag{B.27}$$

Proof. Proposition B.1 gives

$$(2-K)\|\nabla f(X_t)\|_* - \langle \nabla f(X_t), P_t\rangle \le 2\|N_t - \nabla f(X_t)\|_* + 2\tau.$$

On the other hand, nuclear–operator duality implies

$$(2-K)\|\nabla f(X_t)\|_* - \langle \nabla f(X_t), P_t\rangle \le \big((2-K)+K\big)\|\nabla f(X_t)\|_* = 2\|\nabla f(X_t)\|_* \le 2H.$$

Taking the smaller of the two upper bounds and using $\tau \ge 0$ proves equation B.27. □

Proof of Theorem 2. For the given budget T, choose β to satisfy the parameter conditions of Theorem 2. The momentum states follow Algorithm 1, with step bound $\|P_t\|_{\mathrm{op}} \le K < 2$. Consequently

$$\|\nabla f(X_t) - \nabla f(X_{t-1})\|_* \le LrK\eta.$$

Use Lemma B.7 with

$$H = \sqrt{2Lr\Delta_1} + LrK\eta T, \qquad d = LrK\eta.$$

Lemma B.10 supplies the required radius. Combine equation B.27 with that blockwise discrepancy bound, and apply Lemmas B.5, B.8, and B.9. On one event of probability at least $1-\delta$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\big[(2-K)\|\nabla f(X_t)\|_* - \langle \nabla f(X_t), P_t\rangle\big] \le 40\Bigg[&\frac{\ell\sqrt{2Lr\Delta_1}}{T} + \frac{LrK\eta}{1-\beta} \\ &+ \frac{128e^6 p}{p-1}\sqrt{r}\left(\frac{1-\beta}{B}\right)^{\frac{p-1}{p}}\left(\sigma_0 + \frac{\sigma_1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* + 4\sigma_1 L\sqrt{r}K\eta\ell\right) \\ &+ \frac{\ell}{T}\left(\sqrt{2Lr\Delta_1} + LrK\eta T\right)\log\frac{1}{\delta}\Bigg] + 2\tau.\end{aligned} \tag{B.28}$$

The discrepancy bound uses the numerator N_t itself. Accordingly, the threshold in the support condition contributes 2τ per update, independently of the momentum memory length.

Smoothness gives

$$f(X_{t+1}) \le f(X_t) - \eta\langle \nabla f(X_t), P_t\rangle + \frac{LrK^2}{2}\eta^2.$$

After summation and division by ηT, rearrangement gives

$$(2-K)\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_* \le \frac{\Delta_1}{\eta T} + \frac{LrK^2\eta}{2} + \frac{1}{T}\sum_{t=1}^{T}\big[(2-K)\|\nabla f(X_t)\|_* - \langle \nabla f(X_t), P_t\rangle\big].$$

Use the same choice $C_p = 2^{14}e^6 p/(p-1)$ as in the proof of Theorem 1. The noise condition in Theorem 2 leaves at least $(2-K)/2$ multiplying the gradient sum after relative-noise absorption. Its remaining drift contribution is bounded by a numerical constant times $LrK\eta/(1-\beta)$, using $\ell \le 2/(1-\beta)$ and $\sqrt{r} \le r$. Also $K^2/2 \le K$, because $K < 2$. Divide by the remaining coefficient and use equation B.14. After extracting $1/(2-K)$, the additive-noise coefficient is at most $80 \cdot 128e^6 p/(p-1)$, the radius and drift coefficients are at most 160, and the threshold coefficient is at most 4. These are all bounded by C_p. These estimates yield equation 5. □

Proof of Corollary 2. For the given $T \in \mathbb{N}_+$, the choice $\beta = 1 - T^{-p/(3p-2)}$ satisfies $0 \le \beta \le 1 - 1/T$. With $\eta = (1-\beta)^{(2p-1)/p}$, the four powers of T are those computed in the proof of Corollary 1. The noise condition in Corollary 2 supplies the absorption required by Theorem 2. Insert these parameters into equation 5. The common $1 \le K < 2$ and τ keep the approximation constants independent of T. With the problem parameters fixed, the resulting bound proves Corollary 2. □

C Adam's Uniform Confidence Lower Bound

We prove Theorem 3 for the recursion in Algorithm 2. Throughout this appendix, $p \in (1,2]$, $B \in \mathbb{N}_+$, $\beta_2 \in (0,1)$, $\epsilon > 0$ and $\nu \ge 0$ are fixed, and $h = 1 - \beta_1 \in (0,1]$. Constants may depend on these fixed quantities and on explicitly fixed auxiliary constants, but not on h, γ, T, δ.

Call an instance normalized if it has at most two scalar coordinates, a globally 1-Lipschitz gradient, infimum zero, initial objective value one, and conditionally independent, unbiased primitive gradients whose centered conditional p-moments are at most one. Scalar instances are embedded in $\mathbb{R}^{1\times 2}$ by adding a constant coordinate. This leaves the nonconstant Adam recursion unchanged and gives $m_0 = 0$, $v_0 = \nu\mathbf{1}$. The row-matrix nuclear and Frobenius norms coincide, so every lower bound below uses the same score

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}}.$$

Every constructed function and oracle law is specified before sampling. The oracle is defined at arbitrary query points, including those visited by Muon.

The proof uses deterministic interpolation to control long first-moment memories and stochastic constructions to control rare impulses and persistent noise. We first establish these constructions, then combine their parameter regimes with constants uniform over all positive learning rates.

C.1 Stochastic constructions

Throughout this subsection, fix $p \in (1, 2]$, $B \in \mathbb{N}_+$, $\beta_2 \in (0, 1)$, $\epsilon > 0$, and $\nu \geq 0$. The scalar Adam recursion is

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \qquad x_{t+1} = x_t - \gamma d_t, \quad d_t = \frac{m_t}{\sqrt{v_t} + \epsilon},$$

with $m_0 = 0$, $v_0 = \nu$, and g_t the average of the B primitive gradients. Write $h = 1 - \beta_1$. For two coordinates, the recursion is applied coordinatewise with $m_0 = (0, 0)$ and $v_0 = (\nu, \nu)$. Constants below are independent of $T, \delta, \beta_1, \gamma$, except for additional fixed quantities explicitly stated in a lemma. We use the globally 1-smooth profiles

$$\Phi_s(x) = \begin{cases} 0, & x \leq -s, \\ (x+s)^2/2, & -s < x < 0, \\ sx + s^2/2, & x \geq 0, \end{cases} \quad \Psi_s(x) = \begin{cases} x^2/2, & |x| \leq s, \\ s|x| - s^2/2, & |x| > s. \end{cases} \tag{C.1}$$

The parameters s in these profiles are gradient heights. All two-dimensional instances are represented as 1-by-2 row matrices, so their Frobenius, nuclear, and Euclidean gradient norms agree and the matrix rank parameter is one.

Lemma C.1 (A uniform confidence tail)**.** *For every fixed $\tau > 0$, there are constants $c_{\mathrm{tail}}, \delta_{\mathrm{tail}} > 0$ and one instance, independent of $T, \delta, \beta_1, \gamma$, such that*

$$\mathbb{P}\left\{ \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(X_t)\|_{\mathrm{F}} \geq c_{\mathrm{tail}} \min\left\{ 1, \frac{\delta^{-1/[2(p+\tau)]}}{T} \right\} \right\} > \delta \tag{C.2}$$

for every $T \geq 1$, $\beta_1 \in [0, 1)$, $\gamma > 0$, and $0 < \delta < \delta_{\mathrm{tail}}$. The instance has dimension two, a nonnegative globally 1-smooth objective with infimum zero and initial gap one, and conditionally independent and identically distributed primitive gradients with conditional centered p-moment at most one.

Proof. Put $\rho = \sqrt{\beta_2}$ and fix

$$b_* := \frac{1 + \beta_2^{1/4}}{2} \in (\beta_2^{1/4}, 1).$$

We construct two separable coordinates, one for $\beta_1 \leq b_*$ and one for $\beta_1 \geq b_*$.

Choose a fixed $0 < \kappa_0 \leq 1/32$. For an amplitude $A \geq 4$, set

$$\pi = \kappa_0 A^{-p}, \qquad s = \pi A/2, \qquad e = -2s + A\chi, \quad \chi \sim \mathrm{Bernoulli}(\pi).$$

Then $\mathbb{E}e = 0$ and

$$\mathbb{E}|e|^p \leq (\pi A)^p + \pi A^p = \kappa_0^p A^{-p(p-1)} + \kappa_0 \leq 1/8.$$

The first coordinate has objective Φ_s, initial point zero, and independent copies of e added to every primitive gradient at every update. If n_t counts the positive impulses in the batch and $I_t = \mathbf{1}_{\{n_t \geq 1\}}$, its observed gradient is

$$g_t = a_t + (A/B) n_t, \qquad a_t = \Phi_s'(x_t) - 2s \in [-2s, -s].$$

Increase A, if needed, so that $A/B \geq 4s$ and $s \leq 1/32$. Every input then has magnitude at least s, and an impulse batch has $g_t \geq A/(2B)$. In particular

$$\sqrt{v_t} + \epsilon \geq D_s := s\sqrt{1 - \beta_2} + \epsilon.$$

An impulse at update i contributes at most $A\beta_1^{t-i}$ to the positive numerator at update t. The same impulse implies $v_t \geq (1-\beta_2)\beta_2^{t-i}A^2/(4B^2)$. Consequently, uniformly over $\beta_1 \leq b_*$, its contribution to d_t is bounded by

$$F_k(A) := b_*^k \min\{c_2^{-1}\rho^{-k}, A/D_s\}, \qquad k = t-i, \qquad c_2 = \frac{\sqrt{1-\beta_2}}{2B}.$$

Let $H(A) = \sum_{k=0}^{\infty} F_k(A)$. Splitting this sum at the first index where $c_2^{-1}\rho^{-k} \geq A/D_s$, and summing the two geometric series, gives

$$H(A) \leq C\left(1 + (A/D_s)^{\eta}\right), \qquad \eta = 1 - \frac{\log b_*}{\log \rho} \in (0,1).$$

Since $D_s \to \epsilon > 0$ as $A \to \infty$, this implies $H(A) = o(A)$.

An impulse batch covers its own update and the next $K_A - 1$ updates, where

$$K_A = \max\left\{1, \left\lceil \frac{4\log\big(A/[s(1-\beta_2^{1/4})]\big)}{-\log\beta_2} \right\rceil \right\}.$$

At an uncovered update t, every earlier impulse has age $k \geq K_A$, and hence $A\beta_2^{k/4} \leq s(1-\beta_2^{1/4})$. Summing over its possible ages gives

$$\sum_{i=1}^{t} \beta_2^{t-i} A^2 I_i \leq s^2 \frac{(1-\beta_2^{1/4})^2}{1-\sqrt{\beta_2}} \leq s^2.$$

Also $g_i^2 \leq 4s^2 + A^2 I_i$. Choose a fixed integer K_0 with $b_*^{K_0} \leq 1/2$ and $\beta_2^{K_0}\nu \leq s^2$. At an uncovered $t \geq K_0$, we therefore have $v_t \leq 6s^2$. The baseline contribution to m_t is at most $-s(1-\beta_1^t)$, so at these updates its normalized contribution is at most

$$-c_s, \qquad c_s := \frac{s}{2(\sqrt{6}s + \epsilon)}.$$

At all other updates that contribution is nonpositive. Expanding the numerator into its baseline and impulses, and using the impulse envelope for each summand with the actual denominator, proves the pathwise prefix bound

$$\sum_{t=1}^{j} d_t \leq -c_s j + c_s K_0 + \big(c_s K_A + H(A)\big) \sum_{t=1}^{j} I_t. \tag{C.3}$$

This argument uses only $a_t \in [-2s, -s]$, and therefore holds along every resulting trajectory.

The variables I_t are independent Bernoulli variables of parameter $\Pi = 1 - (1-\pi)^B \leq B\pi$. Write $J_A = c_s K_A + H(A)$. Here $K_A = \mathcal{O}(\log A)$, with a constant depending only on p, β_2, κ_0, while c_s is of order A^{1-p} and $H(A) = o(A)$, so

$$\frac{\Pi J_A}{c_s} \leq B\pi K_A + \frac{B\pi H(A)}{c_s} \longrightarrow 0.$$

Fix A large enough that $\Pi J_A < c_s/4$. Choose a fixed $C_0 > c_s K_0$ large enough that

$$\mathbb{E}\exp\left\{\frac{\log 16}{C_0 - c_s K_0}(J_A I_t - c_s/2)\right\} \leq 1.$$

Such a choice exists because the moment-generating function of $J_A I_t - c_s/2$ has negative derivative at zero. The constant C_0 depends only on the fixed structural parameters and the fixed amplitude A, not on $T, \delta, \beta_1, \gamma$. The exponential product is a nonnegative supermartingale. Ville's maximal inequality gives, with probability at least $15/16$, simultaneously for every $j \geq 1$,

$$J_A \sum_{t=1}^{j} I_t \leq c_s j/2 + C_0 - c_s K_0.$$

On this event, Equation C.3 implies

$$x_{j+1} = -\gamma \sum_{t=1}^{j} d_t \geq \gamma(c_s j/2 - C_0).$$

Thus every sufficiently late iterate is nonnegative and has gradient s. More precisely, set $J_0 = \max\{1, \lceil 2C_0/c_s \rceil + 1\}$. If $T \geq 2(J_0+1)$, at least half the evaluated gradients equal s. For smaller T, the initial gradient gives $T^{-1}\sum_{t=1}^{T} |\Phi_s'(x_t)| \geq s/T$. Consequently a fixed $c_{\mathrm{low}} > 0$ satisfies

$$\mathbb{P}\left\{ \frac{1}{T}\sum_{t=1}^{T} |\Phi_s'(x_t)| \geq c_{\mathrm{low}} \right\} \geq 15/16 \quad (\beta_1 \leq b_*,\ \gamma > 0,\ T \geq 1). \tag{C.4}$$

For the second coordinate, set $s_+ = \sqrt{2 - s^2}$, use objective Ψ_{s_+}, and start at $y_1 = s_+$. We first establish a deterministic bound when its first observed gradient is $-M$, with $M \geq s_+$, and every later gradient is exact. Suppose $\beta_1 \geq b_*$, put $u = M/s_+ \geq 1$, and let $\kappa = -\log\rho > 0$, $h_* = 1 - b_*$. Until the momentum becomes nonnegative, the positions remain at least s_+. On any such initial interval the exact recursion is

$$m_t = -hM\beta_1^{t-1} + s_+(1 - \beta_1^{t-1}). \tag{C.5}$$

Every use of this formula below certifies negative momentum and hence the location required for its next use.

If $h \leq u^{-1/2}$, take $n = 1 + \lfloor \sqrt{u}/4 \rfloor$. For $t \leq n$, Bernoulli's inequality gives $\beta_1^{t-1} \geq 3/4$, and

$$\frac{m_t}{h} = -M\beta_1^{t-1} + s_+ \sum_{j=0}^{t-2} \beta_1^j \leq -3s_+u/4 + s_+\sqrt{u}/4 < 0.$$

The first n evaluated gradients therefore equal s_+, so

$$\frac{1}{T}\sum_{t=1}^{T} |\Psi_{s_+}'(y_t)| \geq \frac{s_+}{4}\min\{1, \sqrt{u}/T\}.$$

If $h \geq u^{-1/2}$, define

$$\zeta(h) = 1 + \frac{\log(1-h)}{\kappa}, \qquad W = hu^{\zeta(h)}.$$

The function $\log h + (\log u)\log(1-h)/\kappa$ is concave on $[u^{-1/2}, h_*]$. At the right endpoint, $\zeta(h_*) > 1/2$ because $b_* > \beta_2^{1/4}$, giving $W \geq h_*\sqrt{u}$. At the left endpoint, $\log(1 - u^{-1/2}) \geq -u^{-1/2}/b_*$, and $(\log u)/\sqrt{u} \leq 2/e$. Thus throughout this interval

$$W \geq c_*\sqrt{u}, \qquad c_* := \min\{h_*, \exp[-2/(e\kappa b_*)]\} > 0. \tag{C.6}$$

Set

$$k = 1 + \left\lceil \frac{\log u}{\kappa} \right\rceil, \quad D_+ = \sqrt{\nu + 2s_+^2} + \epsilon, \quad W_0 = 2/b_*.$$

If $W \geq W_0$, then

$$hM\beta_1^{k-1} \geq \beta_1 s_+ W \geq b_* s_+ W \geq 2s_+.$$

Equation C.5 implies $m_t < 0$ for all $t \leq k$ and $-m_k \geq b_* s_+ W/2$. Moreover $\beta_2^{k-1}M^2 \leq s_+^2$, whence

$$v_k = \beta_2^k \nu + (1-\beta_2)\beta_2^{k-1}M^2 + s_+^2(1 - \beta_2^{k-1}) \leq \nu + 2s_+^2.$$

The k-th update gives

$$y_{k+1} \geq s_+ + \frac{\gamma b_* s_+ W}{2D_+}.$$

After update k, the exact gradients are in $[-s_+, s_+]$, and $m_k < 0$ implies $m_t \leq s_+$ for every later t. Whenever $y_t \geq s_+$, the current exact gradient is s_+, so its leftward update has length at

most $\gamma s_+/D_-$, where $D_- = s_+\sqrt{1-\beta_2}+\epsilon$. Let $a_0 = b_* D_-/(4D_+) \in (0,1)$ and $n = \lfloor a_0 W \rfloor$. Inductively, for $j = 0, \dots, n$,

$$y_{k+1+j} \geq s_+ + \frac{\gamma b_* s_+ W}{2D_+} - n\frac{\gamma s_+}{D_-} \geq s_+ + \frac{\gamma b_* s_+ W}{4D_+} > s_+.$$

This certifies one initial interval of constant gradient containing at least $n+1 \geq a_0 W$ positions, including horizons that end earlier. It follows that

$$\frac{1}{T}\sum_{t=1}^{T} |\Psi'_{s_+}(y_t)| \geq s_+ a_0 \min\{1, W/T\}.$$

If $W < W_0$, the initial gradient instead gives

$$\frac{1}{T}\sum_{t=1}^{T} |\Psi'_{s_+}(y_t)| \geq s_+/T \geq (s_+/W_0)\min\{1, W/T\}.$$

Combining these estimates with Equation C.6 and the case $h \leq u^{-1/2}$ proves that a fixed $c_+ > 0$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T} |\Psi'_{s_+}(y_t)| \geq c_+ \min\{1, \sqrt{M/s_+}/T\} \quad (\beta_1 \geq b_*). \tag{C.7}$$

The same bound applies if the first observed gradient is at most $-M$, by using its actual magnitude in the argument.

We now realize that first observation with one fixed innovation law. Let Z have distribution

$$\mathbb{P}(Z = 2^j) = \frac{2^{-pj}/j^2}{\sum_{i=1}^{\infty} 2^{-pi}/i^2}, \qquad j \geq 1,$$

let ω be an independent fair sign, and put $\xi = a_\xi \omega Z$, with $a_\xi > 0$ small enough that $\mathbb{E}|\xi|^p \leq 1/8$. This law is centered, and for all sufficiently large z,

$$\mathbb{P}(\xi \leq -z) \geq c z^{-p} (\log z)^{-2}.$$

Indeed choose the smallest j with $a_\xi 2^j \geq z$ and use that atom, for which 2^j is within a fixed factor of z and j is of order $\log z$. At update one only, add independent copies of ξ to the B primitive y-gradients; at later updates use exact gradients. Choose fixed b_0 with $\mathbb{P}(|\xi| \leq b_0) > 0$. The event

$$\xi_1 \leq -B(M + s_+) - (B-1)b_0, \qquad |\xi_j| \leq b_0 \quad (2 \leq j \leq B)$$

makes the first average gradient at most $-M$. Its probability is at least $c_B M^{-p}(\log M)^{-2}$ for large M, and hence at least $c_\tau M^{-(p+\tau)}$. Taking $M = (c_\tau/(2\delta))^{1/(p+\tau)}$ and restricting δ to a fixed sufficiently small interval makes this probability at least 2δ, while ensuring $M \geq s_+$. Equation C.7 gives the asserted confidence rate for every $\beta_1 \geq b_*$.

Finally use

$$F(x,y) = \Phi_s(x) + \Psi_{s_+}(y), \qquad (x_1, y_1) = (0, s_+).$$

This function is nonnegative, has infimum zero, global smoothness one, and initial gap $(s^2 + s_+^2)/2 = 1$. Its two coordinate recursions decouple. Use the two independent coordinate innovations just specified, independently across primitive members. For $p \leq 2$,

$$\|(e_x, e_y)\|_2^p \leq |e_x|^p + |e_y|^p,$$

so every centered primitive p-moment is at most $1/4$. The complete objective, initialization, and oracle law are fixed before $T, \delta, \beta_1, \gamma$ are selected. For $\beta_1 \leq b_*$, use Equation C.4; for $\beta_1 \geq b_*$, use the first-innovation event above. Reducing the common rate constant and taking $\delta_{\mathrm{tail}} \leq 1/4$ proves Equation C.2. □

Lemma C.2 (A persistent bias floor)**.** *There are constants $c_{\mathrm{bias}}, C_{\mathrm{bias}} > 0$ and one scalar instance, independent of T, β_1, γ, such that*

$$\mathbb{P}\left\{\frac{1}{T}\sum_{t=1}^{T} |f'(x_t)| \geq c_{\mathrm{bias}}(1-\beta_1)\right\} \geq 1 - T\exp\{-C_{\mathrm{bias}}(1-\beta_1)^2 T\} \tag{C.8}$$

for every $T \geq 2$, $\beta_1 \in [0,1)$, and $\gamma > 0$. The objective is nonnegative and globally 1-smooth with infimum zero and initial gap one. Its primitive gradients are conditionally independent and identically distributed, unbiased, and have centered conditional p-moment at most one.

Proof. Set $A = 1/2$, $a = A/(4B)$, and use independent primitive innovations

$$\xi = \begin{cases} -a, & \text{with probability } A/(A+a), \\ A, & \text{with probability } a/(A+a). \end{cases}$$

They are centered and satisfy $\mathbb{E}|\xi|^p \le 1$. Let X_t be their batch average at update t. It has a single negative support point $-a$, with probability $(A/(A+a))^B$; every positive support point is at least

$$a_+ = (A+a)/B - a > a.$$

Thus $a \le |X_t| \le A$, $\mathbb{E}X_t = 0$, and $\mathbb{E}[(X_t)_+] = a(A/(A+a))^B$.

Define auxiliary moments m_t^0, v_t^0 from the inputs X_t alone, with the same initial state, and let $d_t^0 = m_t^0/(\sqrt{v_t^0} + \epsilon)$. Write

$$V = \max\{\nu, A^2\}, \qquad D_0 = a\sqrt{1-\beta_2} + \epsilon.$$

Then $a^2(1-\beta_2) \le v_t^0 \le V$ and $|m_t^0| \le A$. Condition on every batch innovation except X_i, for $i \le t$, and express $v_t^0 = z + wX_i^2$, where $w = (1-\beta_2)\beta_2^{t-i} > 0$. The reciprocal denominator decreases with the positive magnitude of its argument. Therefore

$$\mathbb{E}\left[\frac{X_i}{\sqrt{z + wX_i^2} + \epsilon} \,\middle|\, (X_j)_{j\ne i}\right] \le a\left(\frac{A}{A+a}\right)^B \left(\frac{1}{\sqrt{z + wa_+^2} + \epsilon} - \frac{1}{\sqrt{z + wa^2} + \epsilon}\right) \le 0. \tag{C.9}$$

At the current lag $i = t$, rationalizing the difference of square roots bounds the right side above by $-c_*$, where

$$c_* = \frac{a(A/(A+a))^B(1-\beta_2)(a_+^2 - a^2)}{2\sqrt{V}(\sqrt{V} + \epsilon)^2} > 0.$$

The expansion $m_t^0 = h\sum_{i=1}^t \beta_1^{t-i} X_i$ and Equation C.9 consequently imply

$$\mathbb{E}d_t^0 \le -c_* h, \qquad \mathbb{E}F_j^0 \le -c_* hj, \quad F_j^0 := \sum_{t=1}^j d_t^0.$$

Replacing one innovation X_i while keeping all others fixed changes the auxiliary moments at time $i+k$ by at most

$$|\Delta m_{i+k}^0| \le h\beta_1^k(A+a), \qquad |\Delta v_{i+k}^0| \le (1-\beta_2)\beta_2^k(A^2 - a^2).$$

Both compared square roots are at least $a\sqrt{1-\beta_2}$. It follows that

$$|\Delta d_{i+k}^0| \le \frac{h\beta_1^k(A+a)}{D_0} + C_v\beta_2^k, \qquad C_v = \frac{A(1-\beta_2)(A^2-a^2)}{2a\sqrt{1-\beta_2}D_0^2}.$$

Summing over k shows that each innovation changes F_j^0 by at most the fixed constant

$$L_0 = \frac{A+a}{D_0} + \frac{C_v}{1-\beta_2},$$

independently of h and j. The bounded-differences inequality therefore gives

$$\mathbb{P}\{F_j^0 > -c_* hj/2\} \le \exp\left\{-\frac{c_*^2 h^2 j}{2L_0^2}\right\}. \tag{C.10}$$

Fix $s_0 \in (0, \min\{a/2, 1\}]$. For any predictable sequence $u_i \in [0, s_0]$, compare the actual moments from inputs $X_i + u_i$ with the auxiliary moments. Every actual input has magnitude at least $a/2$, and hence its denominator is at least $D_* = (a/2)\sqrt{1-\beta_2} + \epsilon$. Moment expansion and the reverse triangle inequality for the weighted Euclidean input vectors give

$$|m_t - m_t^0| \le h\sum_{i=1}^t \beta_1^{t-i} u_i, \qquad |\sqrt{v_t} - \sqrt{v_t^0}| \le \sqrt{1-\beta_2}\sum_{i=1}^t (\sqrt{\beta_2})^{t-i} u_i.$$

The initial-variance component is identical in the two vectors. Using $|m_t^0| \le A$, and summing both geometric convolutions, proves the pathwise inequality

$$\sum_{t=1}^{j} |d_t - d_t^0| \le \left(\frac{1}{D_*} + \frac{A\sqrt{1-\beta_2}}{D_* D_0 (1-\sqrt{\beta_2})} \right) \sum_{i=1}^{j} u_i. \tag{C.11}$$

This coefficient is independent of h.

Set

$$D_s = 1/s_0 - s_0/2, \qquad f(x) = \Phi_{s_0}(x + D_s), \qquad x_1 = 0.$$

Then $f(0) = 1$, $\inf f = 0$, and $0 \le f' \le s_0$, with $f'(x) = s_0$ for $x \ge 0$. The primitive oracle is $G_{t,j} = f'(x_t) + \xi_{t,j}$, so its batch gradient has the form above with $u_t = f'(x_t)$. Choose

$$0 < c_{\mathrm{bias}} \le \min \left\{ \frac{s_0}{3}, \frac{c_*}{8 \left(\frac{1}{D_*} + \frac{A\sqrt{1-\beta_2}}{D_* D_0 (1-\sqrt{\beta_2})} \right)} \right\}.$$

By Equation C.10 and a union bound, with probability at least

$$1 - T \exp\left\{ -\frac{c_*^2 h^2 T}{4 L_0^2} \right\},$$

every $j \in \{\lceil T/2 \rceil, \dots, T-1\}$ satisfies $F_j^0 \le -c_* h j/2$. Suppose on this event that

$$\frac{1}{T} \sum_{t=1}^{T} |f'(x_t)| < c_{\mathrm{bias}} h.$$

All gradients are nonnegative, so $\sum_{i=1}^{j} u_i < c_{\mathrm{bias}} h T$ for every $j \le T$. Equation C.11 yields, for the selected prefixes,

$$\sum_{t=1}^{j} d_t < -c_* h T/4 + \left(\frac{1}{D_*} + \frac{A\sqrt{1-\beta_2}}{D_* D_0 (1-\sqrt{\beta_2})} \right) c_{\mathrm{bias}} h T \le -c_* h T/8 < 0.$$

Thus $x_{j+1} > 0$, and at least $\lfloor T/2 \rfloor$ of the evaluated gradients equal s_0. For $T \ge 2$, this implies $T^{-1} \sum_{t=1}^{T} |f'(x_t)| \ge s_0/3 \ge c_{\mathrm{bias}} h$, a contradiction. Taking $C_{\mathrm{bias}} = c_*^2/(4L_0^2)$ proves the lemma. □

Lemma C.3 (A floor at a fixed effective step)**.** *For every fixed $\alpha_0 > 0$, there is a constant $c_{\mathrm{eff}} > 0$ such that, for every $T \ge 1$, $\beta_1 \in [0, 1)$, and $\gamma > 0$ with $\gamma(1-\beta_1) \ge \alpha_0$, a preselected instance satisfies*

$$\mathbb{P}\{ \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(X_t)\|_{\mathrm{F}} \ge c_{\mathrm{eff}} \} \ge 1/8. \tag{C.12}$$

The instance has dimension at most two, a nonnegative globally 1-smooth objective with infimum zero and initial gap one, and conditionally independent unbiased primitive gradients, each with conditional centered p-moment at most one. The constant may depend on α_0 and the fixed structural parameters, but not on T, β_1, γ.

Proof. Fix $K = 1/(4B)$. We first study a stationary coordinate with objective Ψ_λ, initial point $x_1 = 0$, and a small fixed gradient height $0 < \lambda \le K/2$. At each update, draw

$$\zeta_t = \begin{cases} K, & \text{with conditional probability } (1 + \Psi_\lambda'(x_t)/K)/2, \\ -K, & \text{otherwise.} \end{cases}$$

It is conditionally unbiased, and $\zeta_t^2 = K^2$ on every path. Thus its Adam denominator is deterministic:

$$D_t = \sqrt{\beta_2^t \nu + (1-\beta_2^t) K^2} + \epsilon,$$
$$0 < D_- := K\sqrt{1-\beta_2} + \epsilon \le D_t \le D_+ := \sqrt{\max\{\nu, K^2\}} + \epsilon.$$

Fix $c_L \in (0, K^2/128]$ and let $\ell = \lfloor c_L/\lambda^2 \rfloor$. We will decrease λ so that ℓ is sufficiently large and $\ell \geq c_L/(2\lambda^2)$. Fix a block of updates $a+1, \dots, a+\ell$, and condition on any entrance history $\mathscr{F}_a$. Compare its actual conditional sign law $\mathcal{L}((\zeta_{a+i}/K)_{i=1}^{\ell} \mid \mathscr{F}_a)$ with $\mathrm{Unif}(\{-1,1\})^{\otimes \ell}$, the product law of independent fair signs. The fair-sign experiment retains the same entrance state and deterministic denominator sequence, replacing only the adaptive coins. For $|u| \leq 1/2$,

$$D_{\mathrm{KL}}\left(\mathrm{Bern}\big((1+u)/2\big) \,\big\|\, \mathrm{Bern}(1/2)\right) \leq u^2.$$

For example, the second derivative of the left side as a function of u is $1/(1-u^2) \leq 4/3$, and its value and derivative vanish at zero. Since $|\Psi'_\lambda| \leq \lambda$, the conditional chain rule for relative entropy gives

$$D_{\mathrm{KL}}\big(\mathcal{L}((\zeta_{a+i}/K)_{i=1}^{\ell} \mid \mathscr{F}_a) \,\big\|\, \mathrm{Unif}(\{-1,1\})^{\otimes \ell}\big) \leq \ell\lambda^2/K^2 \leq 1/128.$$

Pinsker's inequality, with total variation defined as the supremum over events, therefore gives

$$\big\|\mathcal{L}((\zeta_{a+i}/K)_{i=1}^{\ell} \mid \mathscr{F}_a) - \mathrm{Unif}(\{-1,1\})^{\otimes \ell}\big\|_{\mathrm{TV}} \leq 1/16.$$

Suppose $h\ell \leq 1/2$. The terminal position in the fair-sign experiment has the form

$$X = \mu - \sum_{i=1}^{\ell} K w_i \omega_i, \qquad w_i = \gamma h \sum_{j=i}^{\ell} \frac{\beta_1^{j-i}}{D_{a+j}},$$

where the ω_i's are independent fair signs and μ absorbs the arbitrary entrance position and first moment. Bernoulli's inequality gives $\beta_1^{\ell} \geq 1 - h\ell \geq 1/2$, so

$$w_i \leq \frac{\gamma h \ell}{D_-}, \qquad w_i \geq \frac{\gamma h \ell}{4D_+} \quad (i \leq \lfloor \ell/2 \rfloor).$$

For $\ell \geq 4$, the variance $K^2 \sum_{i=1}^{\ell} w_i^2$ and the sum of third absolute moments satisfy

$$K\sqrt{\sum_{i=1}^{\ell} w_i^2} \geq \frac{K\gamma h \ell^{3/2}}{8D_+}, \qquad \frac{\sum_{i=1}^{\ell} |Kw_i|^3}{\big(K\sqrt{\sum_{i=1}^{\ell} w_i^2}\big)^3} \leq \frac{\max_i |Kw_i|}{K\sqrt{\sum_{i=1}^{\ell} w_i^2}} \leq \frac{8D_+}{D_-\sqrt{\ell}}.$$

The Berry–Esseen inequality for independent centered summands bounds the distance of the standardized sum from a standard normal distribution by a universal constant times this last ratio. A normal distribution of standard deviation $K\sqrt{\sum_{i=1}^{\ell} w_i^2}$ assigns mass at most $2\lambda/(K\sqrt{2\pi \sum_{i=1}^{\ell} w_i^2})$ to any interval of length 2λ. It follows, uniformly in the entrance state, that the conditional probability in the fair-sign experiment satisfies

$$\mathbb{P}_{\omega \sim \mathrm{Unif}(\{-1,1\})^{\otimes \ell}} \left\{ \left| \mu - \sum_{i=1}^{\ell} K w_i \omega_i \right| < \lambda \right\} \leq \frac{C_1 \lambda}{\gamma h \ell^{3/2}} + \frac{C_2}{\sqrt{\ell}}. \tag{C.13}$$

Here C_1, C_2 depend only on $K, \beta_2, \epsilon, \nu$. Choose λ small enough that the second term is at most $1/8$. Since $\ell \geq c_L/(2\lambda^2)$, there is a fixed $C_{\mathrm{step}} > 0$, independent of λ, such that the first term is at most $1/8$ whenever $\gamma h \geq C_{\mathrm{step}} \lambda^4$. Returning to the actual experiment, the total-variation estimate then gives

$$\mathbb{P}\{|x_{a+\ell+1}| \geq \lambda \mid \mathscr{F}_a\} \geq 1/2. \tag{C.14}$$

Decrease the fixed λ further so that $C_{\mathrm{step}} \lambda^4 \leq \alpha_0$. For $h\ell \leq 1/2$ and $T \geq 2\ell$, apply Equation C.14 to the ℓ updates preceding each position $t \geq \ell + 1$. Since $|\Psi'_\lambda(x)| = \lambda$ for $|x| \geq \lambda$, this gives

$$\mathbb{E}R_x \geq \lambda/4, \qquad R_x := \frac{1}{T} \sum_{t=1}^{T} |\Psi'_\lambda(x_t)| \in [0, \lambda].$$

Boundedness gives

$$\mathbb{E}R_x \leq \lambda/8 + (7\lambda/8)\mathbb{P}\{R_x \geq \lambda/8\}.$$

Thus $\mathbb{P}\{R_x \geq \lambda/8\} \geq 1/7 \geq 1/8$.

To realize this coordinate with the prescribed primitive batch size and initial gap, take

$$F(x,z) = \Psi_\lambda(x) + z^2/2, \qquad (x_1, z_1) = (0, \sqrt{2}).$$

At a query (x, z), put $b = \Psi'_\lambda(x)$, draw ζ as above, and return

$$G_1 = (B\zeta - (B-1)b, z), \qquad G_j = (b, z) \quad (2 \le j \le B). \tag{C.15}$$

All these vectors have conditional mean $\nabla F(x, z)$. They are conditionally independent, since only the first is random, and their average is (ζ, z). The centered noise in the first primitive has norm

$$B|\zeta - b| \le B(K + \lambda) \le 3/8,$$

while all the other centered noises vanish. In particular every primitive has conditional centered p-moment at most one. When $B > 1$, these primitive distributions need not be identical. The first-coordinate recursion is exactly the scalar process above. The objective has infimum zero, initial gap one, and global gradient Lipschitz constant one. Since $T^{-1}\sum_{t=1}^{T}\|\nabla F(x_t, z_t)\|_2 \ge R_x$, it proves the required event when $h\ell \le 1/2$ and $T \ge 2\ell$.

If $h\ell > 1/2$, use instead Lemma C.2. It supplies a normalized scalar witness with

$$\mathbb{P}\left\{\frac{1}{T}\sum_{t=1}^{T}|f'(x_t)| \ge c_{\mathrm{bias}}/(2\ell)\right\} \ge 1 - T\exp\{-C_{\mathrm{bias}}T/(4\ell^2)\}.$$

Choose one fixed integer $T_* \ge 2\ell$ such that the right side is at least $3/4$ for every $T \ge T_*$. For $T < T_*$, the deterministic scalar quadratic $f(z) = z^2/2$, started at $z_1 = \sqrt{2}$, gives $T^{-1}\sum_{t=1}^{T}|f'(z_t)| \ge \sqrt{2}/T \ge \sqrt{2}/T_*$ from the initial gradient. Scalar witnesses can be padded with a constant coordinate. Selecting the applicable instance before sampling and taking

$$c_{\mathrm{eff}} = \min\{\lambda/8, c_{\mathrm{bias}}/(2\ell), \sqrt{2}/T_*\} > 0$$

proves Equation C.12 for every displayed parameter choice. □

C.2 Deterministic objectives for long first-moment memory

For scalar objectives, the average gradient norm is

$$\frac{1}{T}\sum_{t=1}^{T}|f'(x_t)|.$$

Throughout this subsection, $\beta_2 \in (0,1)$, $\epsilon > 0$, and $\nu \ge 0$ are fixed. Write $h = 1 - \beta_1$, $\beta = 1 - h$, and

$$m_t = \beta m_{t-1} + h g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2, \qquad x_{t+1} = x_t - \frac{\gamma m_t}{D_t}, \qquad D_t = \sqrt{v_t} + \epsilon,$$

with $x_1 = 0$, $m_0 = 0$, and $v_0 = \nu$. All the oracles in this subsection return exact gradients. In particular, any fixed batch consists of conditionally independent exact gradients and has zero centered moment of every order.

Lemma C.4 (Oscillatory objective with a fixed learning-rate cap)**.** *For every fixed $\Gamma \in (0, \infty)$, there are constants $C_0(\Gamma), c_\Gamma > 0$, depending only on $\Gamma, \beta_2, \epsilon, \nu$, with the following property. For every $T \ge 1$, $h \in (0,1)$, and*

$$0 < \gamma \le \Gamma, \qquad \gamma/h \ge C_0(\Gamma),$$

there is a scalar objective $f \ge 0$ with $\mathrm{Lip}(f') \le 1$, $\inf f = 0$, and $f(0) = 1$, such that

$$\frac{1}{T}\sum_{t=1}^{T}|f'(x_t)| \ge \frac{c_\Gamma}{\sqrt{1+hT}}.$$

The objective may depend on the numerical parameters T, h, γ.

Proof. Put $D_{\max} = \sqrt{\nu} + 1 + \epsilon$, and fix a sufficiently small $c_{\mathrm{per}} > 0$, with

$$c_{\mathrm{per}} \le \min\{1/8, (16\sqrt{D_{\max}})^{-1}\}.$$

Choose

$$\overline{C} \ge (1024/c_{\mathrm{per}})^2, \qquad C_0(\Gamma) \ge \max\{\overline{C}, 4\Gamma, 64\Gamma^2\}.$$

Thus $h \le 1/4$ and $\gamma h \le 1/64$. Define the period, with angular frequency $2\pi/K$, by

$$K = 4\left\lceil \frac{2\pi}{4c_{\mathrm{per}}\sqrt{\gamma h}} \right\rceil. \tag{C.16}$$

Increasing $\overline{C}$, if needed, gives

$$\frac{c_{\mathrm{per}}}{2}\sqrt{\gamma h} \le 2\pi/K \le c_{\mathrm{per}}\sqrt{\gamma h} \le \frac{1}{8}, \qquad \frac{2\pi}{Kh} \ge 512, \qquad hK \le \frac{1}{16}. \tag{C.17}$$

These estimates use only $\gamma h \le 1/64$, $\gamma/h \ge \overline{C}$, and the fixed choice of c_{per}. Constants denoted by C or implicit in $\mathcal{O}$ below depend only on $\beta_2, \epsilon, \nu, c_{\mathrm{per}}$; the dependence on γ is retained explicitly in the work estimates.

Let $\varrho \in (0, 1/4]$ be an amplitude to be chosen later, and set

$$U = \frac{h\varrho(1 - \beta\cos(2\pi/K))}{h^2 + 2\beta(1 - \cos(2\pi/K))}, \qquad V = \frac{h\varrho\beta\sin(2\pi/K)}{h^2 + 2\beta(1 - \cos(2\pi/K))}, \qquad b = 8V. \tag{C.18}$$

The bounds in equation C.17 imply

$$\frac{h\varrho}{8(2\pi/K)} \le V \le \frac{4h\varrho}{2\pi/K}, \qquad 0 \le U \le V/2, \qquad b \le \frac{32h\varrho}{2\pi/K} \le \frac{\varrho}{16}. \tag{C.19}$$

For example, $U/V = (1 - \beta\cos(2\pi/K))/(\beta\sin(2\pi/K)) \le 4h/(2\pi/K) + 2(2\pi/K) \le 1/2$. The other inequalities follow from $\left[h^2 + 2\beta(1 - \cos(2\pi/K))\right] \asymp (2\pi/K)^2$, $\beta \ge 1/2$, and $(2\pi/K)/2 \le \sin(2\pi/K) \le (2\pi/K)$.

Prescribe $g_t = -(b + \varrho)$ until the first update n at which $m_n \le -2b$. This update exists because $b < \varrho$, and

$$m_t = -(b + \varrho)(1 - \beta^t)$$

on this initial interval. Minimality of n and $h(b + \varrho) \le b$ give

$$-3b \le m_n \le -2b.$$

Moreover, $n \le CK/(2\pi)$: the integer $\lceil 4b/[h(b + \varrho)] \rceil$ suffices, by $1 - e^{-z} \ge z/2$ for $0 \le z \le 1$, and $b = \mathcal{O}(h\varrho K/(2\pi))$. Continue the prescribed gradients by

$$g_{n+j} = -b - \varrho\cos(2\pi j/K), \qquad j \ge 0. \tag{C.20}$$

The definitions agree at $j = 0$. Substitution in the first-moment recursion gives the exact formula

$$\begin{aligned} m_{n+j} &= -b - U\cos(2\pi j/K) - V\sin(2\pi j/K) + E\beta^j, \\ E &= m_n + b + U < 0, \qquad |E| \le 2b. \end{aligned} \tag{C.21}$$

In particular,

$$-4b \le m_{n+j} \le -b/2 < 0. \tag{C.22}$$

Generate the positions and second moments from these prescribed inputs using the actual Adam recursion. Every position increment is positive. Since $|g_t| \le 2\varrho \le 1/2$, $\epsilon \le D_t \le D_{\max}$. On the periodic interval, write $d_t = x_{t+1} - x_t$; then

$$\frac{\gamma b}{2D_{\max}} \le d_t \le \frac{4\gamma b}{\epsilon}, \qquad |g_{t+1} - g_t| \le \varrho(2\pi/K) \le d_t/4. \tag{C.23}$$

The last inequality follows from $b \ge h\varrho/(2\pi/K)$ and the choice of c_{per}. Thus the linear interpolation of the gradients at consecutive positions has slope of magnitude at most $1/4$ on this interval, and is constant during the initial interval.

We bound the decrease of the function obtained by integrating this interpolation. The initial constant-gradient interval costs at most

$$\frac{\gamma h(b+\varrho)^2}{\epsilon}\sum_{t=1}^{n} t \le C\frac{\gamma h\varrho^2}{(2\pi/K)^2} \le C\varrho^2. \tag{C.24}$$

On the periodic interval, first replace D_t by $\epsilon + |g_t|$. Define

$$F_b(z) = \frac{-b-\varrho z}{\epsilon + |-b-\varrho z|}.$$

The term $-V\sin(2\pi j/K)$ in equation C.21 cancels over a full period under the pairing $j \mapsto K-j$. At $b=0$, the sum of $F_0(\cos(2\pi j/K))$ is zero by the half-period pairing. Since $z \mapsto z/(\epsilon+|z|)$ is $1/\epsilon$-Lipschitz,

$$\left|\sum_{j=0}^{K-1} F_b(\cos(2\pi j/K))\right| \le Kb/\epsilon.$$

Consequently the ideal left-point work in one full period is at most

$$\begin{aligned}\gamma\sum_{j=0}^{K-1}\frac{g_{n+j}\big(-b-U\cos(2\pi j/K)-V\sin(2\pi j/K)\big)}{\epsilon+|g_{n+j}|} &\le \frac{\gamma K}{\epsilon}(b^2+2U\varrho)\\ &\le CK(1+\gamma)h\varrho^2.\end{aligned} \tag{C.25}$$

Here $b^2+U\varrho \le C\varrho^2(h^2/(2\pi/K)^2+h)$. An incomplete period has absolute work at most $C\gamma Kb\varrho \le C\varrho^2$.

The additional term $E\beta^j$ in the actual momentum contributes only $C\varrho^2$ over every prefix. Indeed, the periodic sequence $F_b(\cos(2\pi j/K))$ has mean of magnitude at most b/ϵ; its centered partial sums are bounded by $C\varrho K/(2\pi)$. Abel summation against the decreasing weights β^j therefore shows, for every integer $N \ge 1$,

$$\left|\gamma E\sum_{j=0}^{N-1}\beta^j F_b(\cos(2\pi j/K))\right| \le C\gamma b\left(\frac{\varrho K}{2\pi}+\frac{b}{h}\right) \le C\varrho^2. \tag{C.26}$$

The actual second moment can be controlled without replacing its state. Let $\varpi = \sqrt{\beta_2}$. For the prescribed inputs, including the initial interval,

$$|g_t - g_{t-k}| \le \varrho(2\pi/K)k \qquad (0 \le k < t).$$

The reverse triangle inequality for the weighted Euclidean vectors representing $\sqrt{v_t}$ and $|g_t|$ gives

$$\begin{aligned}\big|\sqrt{v_t}-|g_t|\big| &\le \varpi^t(\sqrt{\nu}+|g_t|)+\left((1-\beta_2)\sum_{k=0}^{t-1}\beta_2^k|g_{t-k}-g_t|^2\right)^{1/2}\\ &\le (\sqrt{\nu}+2\varrho)\varpi^t + C\varrho(2\pi/K).\end{aligned} \tag{C.27}$$

The comparison vector has initial component $\varpi^t g_t$, followed by $\sqrt{1-\beta_2}\varpi^k g_t$ at each lag $0 \le k < t$. Its norm is exactly $|g_t|$, including the initial mass. Both denominators are at least ϵ. The error over the first N periodic updates is therefore at most

$$\begin{aligned}\gamma\sum_{t=n}^{n+N-1}|g_t m_t|\left|D_t^{-1}-(\epsilon+|g_t|)^{-1}\right| &\le C\gamma\varrho b\left(N\varrho(2\pi/K)+(\sqrt{\nu}+2\varrho)\sum_{t=1}^{\infty}\varpi^t\right)\\ &\le C\gamma h\varrho^3 N + C\varrho^2.\end{aligned} \tag{C.28}$$

The constant term uses $\gamma\varrho b = \mathcal{O}(\varrho^2\sqrt{\gamma h})$.

Set the function value at zero to one, and integrate the piecewise linear derivative through (x_t, g_t), for $1 \le t \le T+1$. On each segment its exact change is

$$f(x_{t+1}) - f(x_t) = \tfrac{1}{2}(g_t+g_{t+1})d_t.$$

Its difference from the left-point expression is at most $C\gamma h\varrho^2$ per periodic update, by equation C.23; it is zero on the initial interval. A partial segment has variation at most $2\varrho d_t \le C\varrho^2$. Equations (C.24) to (C.28) therefore bound every prefix, including points between consecutive iterates, by

$$f(x) \ge 1 - C\varrho^2\big(1 + (1+\gamma)hT\big) \ge 1 - C(1+\Gamma)\varrho^2(1+hT). \tag{C.29}$$

Choose

$$\varrho = \frac{c_*}{\sqrt{(1+\Gamma)(1+hT)}}, \tag{C.30}$$

where $c_* > 0$ is fixed and sufficiently small. Then the constructed function is at least $3/4$ on the visited interval.

Extend this function to the whole real line. To the left of zero use

$$f'(x) = -(x + b + \varrho) \quad \text{for } -(b+\varrho) \le x \le 0,$$

and then a constant continuation. This matches the initial derivative, has slope of magnitude one, and keeps the values at least one. At the right endpoint, join the terminal derivative linearly to zero over a length equal to its magnitude. This decreases the value by at most $g_{T+1}^2/2 \le 2\varrho^2$, leaving a positive value H. Starting from this zero-derivative endpoint, let the derivative decrease with slope -1 from zero to $-\sqrt{H}$, then increase with slope 1 back to zero, over a total length $2\sqrt{H}$. The value decreases monotonically by exactly H. Continue with the constant value zero. All derivatives join continuously and are 1-Lipschitz. The resulting global objective is nonnegative, attains zero, and has initial gap one. Its gradient at every prescribed position is g_t, so induction recovers the same Adam trajectory with its original moments.

There is a fixed H_0 such that, if $T(2\pi/K) \ge H_0$, the bound $n \le CK/(2\pi)$ and the period K ensure that complete periods occupy at least $T/2$ of the evaluated positions. In each period, at least $K/4$ indices have $\cos(2\pi j/K) \ge 1/2$, and hence $|g_{n+j}| \ge \varrho/2$. Thus $T^{-1}\sum_{t=1}^{T}|f'(x_t)| \ge \varrho/16$, proving the claimed bound in this case.

If $T(2\pi/K) < H_0$, equation C.17 gives $\gamma hT^2 \le C_{\mathrm{short}}$ for a fixed constant. Choose a fixed $\zeta \in (0, 1/2]$ so that $C_{\mathrm{short}}\zeta^2/\epsilon \le 1/4$, put $\ell = (1-\zeta^2/2)/\zeta$, and use instead

$$f(x) = \begin{cases} \zeta(\ell - x) + \zeta^2/2, & x \le \ell, \\ (\ell + \zeta - x)^2/2, & \ell \le x \le \ell + \zeta, \\ 0, & x \ge \ell + \zeta. \end{cases} \tag{C.31}$$

This function is nonnegative, 1-smooth, and has initial gap one. Before leaving its linear part, $|m_t| \le \zeta ht$, so

$$\zeta\sum_{t=1}^{T}(x_{t+1} - x_t) \le \frac{\gamma h\zeta^2 T(T+1)}{2\epsilon} \le \frac{C_{\mathrm{short}}\zeta^2}{\epsilon} \le \frac{1}{4} < \zeta\ell.$$

All evaluated positions remain on that part, and $T^{-1}\sum_{t=1}^{T}|f'(x_t)| = \zeta$. Decreasing c_Γ proves the statement for every T. □

Lemma C.5 (A constant lower bound for large learning rates)**.** *There are constants $\alpha_0 > 0$, $\Gamma_0 \ge \max\{1, 2\alpha_0\}$, and $c_{\mathrm{tent}} > 0$, depending only on β_2, ϵ, ν, such that the following holds. For every $T \ge 1$, $h \in (0,1)$, and*

$$\gamma \ge \Gamma_0, \qquad \gamma h \le \alpha_0,$$

there is a scalar nonnegative 1-smooth objective with infimum zero and initial gap one on which exact-gradient Adam satisfies

$$\frac{1}{T}\sum_{t=1}^{T}|f'(x_t)| \ge c_{\mathrm{tent}}.$$

Proof. Write $\alpha = \gamma h$, and retain $D_{\max}, c_{\mathrm{per}}, \overline{C}, \varpi$ from the preceding construction. Choose $\alpha_0 \le 1/64$ and then Γ_0 large enough that $\gamma/h = \gamma^2/\alpha \ge \overline{C}$. The frequency construction equation C.16, the transfer coefficients equation C.18, and the short initial interval used in the preceding proof consequently satisfy equation C.17–equation C.23, without an upper bound on γ. Keep the amplitude

ϱ fixed for now. In particular, the actual periodic-interval momentum is $-b - U\cos(2\pi j/K) - V\sin(2\pi j/K) + E\beta^j$, with $E < 0$, $|E| \le 2b$, and $-4b \le m_{n+j} \le -b/2$.

We use the exact trapezoidal work of the linear derivative interpolant. Let

$$C_j = \cos(2\pi j/K), \qquad w_b(C) = \frac{1}{\epsilon + |b + \varrho C|}.$$

For the forward gradient average,

$$\bar{g}_j = \tfrac{1}{2}(g_{n+j} + g_{n+j+1}) = -b - \varrho\frac{1+\cos(2\pi/K)}{2}C_j + \varrho\frac{\sin(2\pi/K)}{2}\sin(2\pi j/K).$$

Pairing j with $K - j$ cancels the terms odd in $\sin(2\pi j/K)$, and expanding the remaining terms gives the exact identity

$$\begin{aligned}\sum_{j=0}^{K-1} \bar{g}_j\big(-b - U\cos(2\pi j/K) - V\sin(2\pi j/K)\big)w_b(C_j) &= b^2\sum_{j=0}^{K-1} w_b(C_j)\\ &+ b\left(U + \varrho\frac{1+\cos(2\pi/K)}{2}\right)\sum_{j=0}^{K-1} C_j w_b(C_j)\\ &+ \varrho\frac{\sin(2\pi/K)}{2}V\sum_{j=0}^{K-1}\big(C_j^2 - \sin^2(2\pi j/K)\big)w_b(C_j)\\ &+ \varrho\left(\frac{1+\cos(2\pi/K)}{2}U - \frac{\sin(2\pi/K)}{2}V\right)\sum_{j=0}^{K-1} C_j^2 w_b(C_j).\end{aligned} \tag{C.32}$$

The transfer coefficients satisfy

$$\frac{1+\cos(2\pi/K)}{2}U - \frac{\sin(2\pi/K)}{2}V = \frac{h^2\varrho(1+\cos(2\pi/K))}{2\big[h^2 + 2\beta(1-\cos(2\pi/K))\big]}. \tag{C.33}$$

Also,

$$\sum_{j=0}^{K-1}\frac{C_j^2 - \sin^2(2\pi j/K)}{\epsilon + \varrho|C_j|} \le 0. \tag{C.34}$$

Indeed, the involution $j \mapsto K/4 - j$ exchanges $|C_j|$ and $|\sin(2\pi j/K)|$, and each paired sum is

$$(C_j^2 - \sin^2(2\pi j/K))\left(\frac{1}{\epsilon + \varrho|C_j|} - \frac{1}{\epsilon + \varrho|\sin(2\pi j/K)|}\right) \le 0.$$

This is why the period in equation C.16 was chosen to be divisible by four.

The weight perturbation and the half-period pairing give

$$|w_b - w_0| \le b/\epsilon^2, \qquad \sum_{j=0}^{K-1} C_j w_0(C_j) = 0.$$

Using equation C.33 and equation C.34 in equation C.32, we obtain

$$\gamma\sum_{j=0}^{K-1}\bar{g}_j\big(-b - U\cos(2\pi j/K) - V\sin(2\pi j/K)\big)w_b(C_j) \le CK\big(h\varrho^2 + \alpha\varrho^3\big). \tag{C.35}$$

The first two terms in equation C.32 are bounded by CKb^2, and $\gamma b^2 = \mathcal{O}(h\varrho^2)$. The last term has the same bound by equation C.33 and $\big[h^2 + 2\beta(1-\cos(2\pi/K))\big] \asymp (2\pi/K)^2 \asymp \alpha$. The loss from replacing w_0 by w_b in the third term is at most

$$CK\gamma\varrho(2\pi/K)Vb \le CK\alpha\varrho^2 b \le CK\alpha\varrho^3.$$

Thus no term of order $K\gamma h\varrho^2$ remains in the trapezoidal work.

The incomplete-period work is bounded by $C\gamma K\varrho b \le C\varrho^2$. The momentum transient also costs at most $C\varrho^2$ over every prefix. To see this, the periodic sequence $\bar{g}_j w_b(C_j)$ has mean $\mathcal{O}(b)$: its sine term cancels, and $\sum_{j=0}^{K-1} C_j w_b(C_j) = \mathcal{O}(Kb)$. Its centered partial sums are $\mathcal{O}(\varrho K/(2\pi))$, so Abel summation yields

$$\left|\gamma E \sum_{j=0}^{N-1} \beta^j \bar{g}_j w_b(C_j)\right| \le C\gamma b\left(\frac{\varrho K}{2\pi} + \frac{b}{h}\right) \le C\varrho^2 \qquad (N \ge 1). \tag{C.36}$$

Finally, equation C.27, $|\bar{g}_j| \le 2\varrho$, and $|m_{n+j}| \le 4b$ give the absolute actual-denominator error

$$\begin{aligned}\gamma \sum_{j=0}^{N-1} |\bar{g}_j m_{n+j}| \left|D_{n+j}^{-1} - w_b(C_j)\right| &\le C\gamma\varrho b\left(N\varrho(2\pi/K) + (\sqrt{\nu} + 2\varrho)\sum_{t=1}^{\infty}\varpi^t\right) \\ &\le C\alpha\varrho^3 N + C\varrho^2.\end{aligned} \tag{C.37}$$

Together with the initial-interval bound equation C.24, these estimates show that the decrease of the linear-interpolant potential through any prefix of N periodic segments is at most

$$C_*\varrho^2 + C_*\varrho^2(h + \alpha\varrho)N, \tag{C.38}$$

where C_* is independent of T, h, γ, ϱ in the stated parameter range.

We now change the derivative between the visited positions while preserving its values at those positions. On each periodic segment of length d_t, add to the linear derivative interpolant the nonnegative function

$$\omega_t(z) = \tfrac{1}{2}\min\{z, d_t - z\}, \qquad 0 \le z \le d_t. \tag{C.39}$$

It vanishes at both endpoints, has slopes of magnitude $1/2$, and has integral $d_t^2/8$. By equation C.23, the resulting derivative is continuous and has Lipschitz constant at most $3/4$ on the periodic interval. Moreover, for a fixed $\kappa > 0$,

$$\frac{d_t^2}{8} \ge \frac{\gamma^2 b^2}{32 D_{\max}^2} \ge \kappa\alpha\varrho^2. \tag{C.40}$$

Here $b \ge h\varrho/(2\pi/K)$ and $(2\pi/K)^2 \le c_{\mathrm{per}}^2\alpha$. Adding these integrals to the function value turns equation C.38 into the upper bound

$$C_*\varrho^2 + \alpha\varrho^2\left(C_*/\gamma + C_*\varrho - \kappa\right)N \tag{C.41}$$

on the potential decrease. Choose $\Gamma_0 \ge 4C_*/\kappa$, in addition to its preceding requirements, and choose a fixed $\varrho = \varrho_0 \le \kappa/(4C_*)$. The coefficient of N in equation C.41 is then negative. At a partial final segment the added integral is nonnegative, whereas the absolute variation of the linear interpolant is at most $C\gamma\varrho_0 b \le C\varrho_0^2$. Consequently the modified function, initialized with $f(0) = 1$, satisfies

$$f(x) \ge 1 - C\varrho_0^2$$

at every point of the entire visited interval. Reducing ϱ_0 makes this lower bound at least $3/4$.

Use the same left continuation and the same terminal extension to a zero minimum as in the preceding proof. The added functions in equation C.39 vanish at the visited endpoints, including the terminal one; hence the terminal derivative still has magnitude at most $2\varrho_0$. Its value is finite and positive, so that extension produces a global nonnegative 1-smooth objective with initial gap exactly one. The derivative at every sampled point remains the prescribed g_t. Therefore the actual Adam trajectory, including both moments, remains unchanged.

When $T(2\pi/K) \ge H_0$, the same complete-period counting as above gives $T^{-1}\sum_{t=1}^{T}|f'(x_t)| \ge \varrho_0/16$. When $T(2\pi/K) < H_0$, αT^2 is bounded by a fixed constant, and the objective equation C.31 gives a fixed positive rate instead. Taking the smaller constant proves the lemma for every T. □

C.3 Travel, rare shocks, and bounded learning rates

Throughout this part, fix $p \in (1, 2]$, $B \in \mathbb{N}_+$, $\beta_2 \in (0, 1)$, $\epsilon > 0$, and $\nu \ge 0$, and write $h = 1 - \beta_1 \in (0, 1]$. Adam uses

$$m_t = (1-h)m_{t-1} + hg_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^{\odot 2}, \qquad x_{t+1} = x_t - \gamma\frac{m_t}{\sqrt{v_t} + \epsilon},$$

with coordinatewise operations, $m_0 = 0$, and $v_0 = \nu\mathbf{1}$. We study the average gradient norm

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|_2.$$

A normalized instance below has dimension at most two, a nonnegative objective with infimum zero, globally 1-Lipschitz gradient, initial gap one, and conditionally independent unbiased primitive gradients whose centered conditional p-moments are at most one. Scalar instances can be padded by a constant coordinate. Regarding the variable as a 1-by-2 matrix then makes the Euclidean and nuclear gradient norms equal, with rank parameter one.

Lemma C.6 (Deterministic travel)**.** *There is $c_{\mathrm{tr}} > 0$, depending only on ϵ, such that, for every $T \geq 1$, $\gamma > 0$, and $h \in (0, 1]$, a normalized scalar instance with an exact oracle satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}|f'(x_t)| \geq c_{\mathrm{tr}}\min\{1, (\gamma T)^{-1/2}\}$$

deterministically.

Proof. Set

$$r = \min\left\{\frac{1}{2}, \sqrt{\frac{\epsilon}{4\gamma T}}\right\}, \qquad \ell = \frac{1 - r^2/2}{r},$$

and define

$$f(x) = \begin{cases} 0, & x \leq -\ell - r, \\ (x + \ell + r)^2/2, & -\ell - r \leq x \leq -\ell, \\ r(x + \ell) + r^2/2, & x \geq -\ell. \end{cases} \qquad x_1 = 0.$$

The derivative is continuous and 1-Lipschitz, and $f(0) = 1$. Use exact gradients in every primitive sample. Until the trajectory leaves $[-\ell, \infty)$, its gradients equal r, its first moments belong to $[0, r]$, and each update moves left by at most $\gamma r/\epsilon$. Since

$$\ell r = 1 - r^2/2 \geq 7/8, \qquad \frac{\gamma r^2 T}{\epsilon} \leq 1/4,$$

a first exit during the first T updates is impossible. All evaluated gradients therefore equal r. The conclusion follows with $c_{\mathrm{tr}} = \min\{1/2, \sqrt{\epsilon}/2\}$. □

Put $s = 1/16$, and use the notation

$$\log_+ z = \log\max\{1, z\}, \qquad k(M) = \left\lceil \frac{\log_+((\nu + M^2)/s^2)}{-\log\beta_2} \right\rceil \quad (M > 0).$$

Lemma C.7 (A rare shock and an uphill bank)**.** *There is $c_{\mathrm{bank}} > 0$, depending only on the fixed parameters, with the following property. For $T \geq 1$, $0 < \delta \leq 1/4$, $\gamma > 0$, and $h \in (0, 1]$, set*

$$M = B^{1/p-1}\left(\frac{T}{256\delta}\right)^{1/p}.$$

If

$$h\big(k(M) + 1\big) \leq \frac{1}{64}, \qquad hT \geq 8, \tag{C.42}$$

there is a normalized two-dimensional instance for which

$$\mathbb{P}\left\{\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|_2 \geq c_{\mathrm{bank}}\min\left\{1, \sqrt{\frac{\gamma M\min\{s, hM, \gamma M\}}{1 + \gamma T/\epsilon}}\right\}\right\} > \delta. \tag{C.43}$$

The primitive gradients can be chosen conditionally identically distributed within each batch.

Proof. We first construct an objective that turns a single negative shock of size M into sustained gradients. Define

$$\phi(z) = \begin{cases} 0, & z \le 0, \\ 3z^2 - 2z^3, & 0 \le z \le 1, \\ 1, & z \ge 1. \end{cases}$$

This function has $0 \le \phi \le 1$, $0 \le \phi' \le 3/2$, and a 6-Lipschitz derivative. Set

$$D_0 = \sqrt{2}s + \epsilon, \qquad D = \frac{\gamma M}{64 D_0}, \qquad G = \frac{1}{128}\min\{s, hM, D\}, \qquad E = \frac{GD}{8}. \tag{C.44}$$

Let $U(x) = E\phi(x/D)$. It is zero on $(-\infty, 0]$, equals E on $[D, \infty)$, and satisfies

$$0 \le U'(x) \le G, \qquad |U''(x)| \le \frac{3G}{4D}$$

almost everywhere. Define the nonincreasing cutoff

$$\chi(z) = \begin{cases} 3/2, & z \le -1, \\ 3/2 - (z+1)^2/2, & -1 \le z \le 0, \\ 1 - z, & 0 \le z \le 1/2, \\ (z - 3/2)^2/2, & 1/2 \le z \le 3/2, \\ 0, & z \ge 3/2. \end{cases}$$

It satisfies $0 \le \chi \le 3/2$, $-1 \le \chi' \le 0$, and $\mathrm{Lip}(\chi') \le 1$. Fix $\eta = 1/16$, and set

$$r = \eta \min\left\{1, \sqrt{\frac{E}{1 + \gamma T/\epsilon}}\right\}, \qquad \psi(y) = \chi(ry/E), \qquad F(x, y) = \phi((x+4)/4) + U(x)\psi(y). \tag{C.45}$$

Initialize at $(x_1, y_1) = (0, 0)$. The objective is nonnegative, vanishes whenever $x \le -4$, and has $F(0, 0) = 1$ and $\nabla F(0, 0) = 0$.

The objective is globally smooth. When $x \le 0$, the second term vanishes and the only nonzero Hessian entry has absolute value at most $6/16$. When $x \ge 0$, the first term is constant and

$$|F_{xx}| \le \frac{9G}{8D}, \qquad |F_{xy}| \le \frac{3r}{2D}, \qquad |F_{yy}| \le \frac{r^2}{E}.$$

The choices above give

$$\frac{G}{D} \le \frac{1}{128}, \qquad \frac{E}{D^2} \le \frac{1}{1024}, \qquad \frac{r^2}{E} \le \eta^2.$$

Consequently the two absolute Hessian row sums are bounded by $9/1024 + 3\eta/64$ and $3\eta/64 + \eta^2$, respectively, both less than one. The first derivatives match at every piecewise boundary. Thus F has globally 1-Lipschitz gradient.

Consider a noise sequence whose first nonzero value is $-M(1, 0)$ at an update $u \le \lfloor T/8 \rfloor$, with all later noise zero. There is no noise in the y-coordinate. Globally, $-r \le F_y \le 0$, so $-r \le m_{t,y} \le 0$, the positions y_t are nondecreasing, and each of their increments is at most $\gamma r/\epsilon$. For all $t \le T + 1$,

$$0 \le \frac{r y_t}{E} \le \frac{\gamma r^2 T}{\epsilon E} \le \eta^2 < \frac{1}{2}. \tag{C.46}$$

Hence throughout the horizon

$$\psi(y_t) = 1 - r y_t/E \in [1 - \eta^2, 1], \qquad F_y(x_t, y_t) = -rU(x_t)/E.$$

Before update u, the iterate stays at the stationary initial point and its first moment remains zero. The favorable shock gives $m_{u,x} = -hM$ and $v_{u,x} \le \nu + M^2$. Put $J = \lfloor 1/(4h) \rfloor$. For as long as $x \ge 0$, Equation C.46 implies $0 \le F_x \le G$. For local update indices $0 \le j \le J$, induction gives

$$m_{u+j,x} \le -hM(1-h)^j + G\big(1 - (1-h)^j\big) \le -\frac{hM}{2}. \tag{C.47}$$

Indeed, $(1-h)^j \geq 1-hj \geq 3/4$ and $G \leq hM/128$. The resulting updates all move right, starting at $x_u = 0$, which verifies the location condition used in this induction. Their second moments obey

$$v_{u+j,x} \leq \beta_2^j(\nu + M^2) + G^2.$$

For $j \geq k(M)$, this is at most $2s^2$. Each update with $k(M) \leq j \leq J$ therefore increases x by at least $\gamma hM/(2D_0)$. The condition in equation C.42 implies $h \leq 1/64$ and

$$J - k(M) + 1 \geq \frac{1}{8h}.$$

These updates produce total displacement at least $\gamma M/(16D_0) = 4D$; in particular the trajectory reaches $x \geq D$ by position $u + J + 1$. Its first moment is still negative at that crossing. Subsequently $F_x = 0$, so the carried first moment decays without changing sign and the positions never leave this plateau. This argument uses the actual first and second moments across the entire uphill interval.

The conditions also imply $T \geq 512$ and $J \leq T/32$, whence $u + J + 1 \leq T/4$. At all later evaluated points $U = E$ and $F_y = -r$. In particular $T^{-1}\sum_{t=1}^{T}\|\nabla F(x_t, y_t)\|_2 \geq r/2$. Finally,

$$E \geq c_E\gamma M \min\{s, hM, \gamma M\}, \qquad c_E = \frac{1}{65536D_0}\min\left\{1, \frac{1}{64D_0}\right\} > 0.$$

Substituting into Equation C.45 gives the threshold in Equation C.43, with a constant independent of T, δ, h, γ, for every such favorable shock sequence.

We now realize this sequence using normalized primitive oracles. Let

$$w = \lfloor T/8 \rfloor, \qquad \varpi = \frac{256\delta}{BT}, \qquad H = \varpi^{-1/p}, \qquad M = H/B.$$

At an active update, independently for each primitive index $j \in \{1, \ldots, B\}$, add to the exact gradient a random vector $\xi_{t,j}$ equal to $-H(1,0)$, $+H(1,0)$, or zero, with probabilities $\varpi/2, \varpi/2, 1-\varpi$. Updates are active until the first batch containing a nonzero primitive noise, or until the end of update w, whichever occurs first. All subsequent primitive gradients are exact. The active indicator is measurable before the current batch is drawn. Thus every primitive is conditionally unbiased, the batch is conditionally independent and identically distributed, and

$$\mathbb{E}\big[\|\xi_{t,j}\|_2^p \mid \mathscr{F}_{t-1}\big] \leq \varpi H^p = 1.$$

These properties hold at arbitrary query points.

Let $\mathcal{E}$ be the event that the first nonempty batch in the active window has exactly one negative primitive noise and all other primitive noises zero. On $\mathcal{E}$, its average noise is $-M(1,0)$, and the preceding pathwise argument applies. Set

$$\pi_0 = (1-\varpi)^B, \qquad \pi_- = (B\varpi/2)(1-\varpi)^{B-1}.$$

Summing over the possible first nonempty batch gives

$$\mathbb{P}(\mathcal{E}) = \frac{\pi_-}{1-\pi_0}(1-\pi_0^w).$$

Since $T \geq 512$ and $\delta \leq 1/4$, we have $B\varpi \leq 1/8$, while $w \geq T/16$. In particular,

$$\frac{\pi_-}{1-\pi_0} \geq \frac{1}{2}(1-B\varpi) \geq \frac{3}{8}, \qquad \pi_0^w \leq \exp(-B\varpi w) \leq e^{-16\delta}.$$

Concavity of $1-e^{-16\delta}$ on $[0, 1/4]$ now yields

$$\mathbb{P}(\mathcal{E}) \geq \frac{3}{8}(1-e^{-16\delta}) \geq \frac{3}{2}(1-e^{-4})\delta > \delta.$$

The objective was chosen before the draws and is the same for every trigger time in this event, proving the lemma. □

C.3.1 A UNIFORM RATE ON EVERY FIXED BOUNDED LEARNING-RATE INTERVAL

Fix a finite number $\Gamma > 0$. All constants in this part may depend on $p, B, \beta_2, \epsilon, \nu, \Gamma$, and on any displayed fixed confidence exponent, but never on T, δ, h, or the actual $\gamma \in (0, \Gamma]$. Choose the washboard constants from Lemma C.4, denote its lower constant by $c_{\mathrm{wb}}(\Gamma) > 0$, and increase its threshold, if needed, to

$$C_0 = C_0(\Gamma) \geq \max\{1, 4\Gamma\}.$$

Thus, whenever $\gamma > C_0 h$ and $\gamma \leq \Gamma$, one has $h < 1/4$, within that lemma's momentum domain.

Lemma C.8 (Balance of travel and the bank)**.** *There is $c_{\mathrm{res}}(\Gamma) > 0$ such that, for every $0 < \gamma \leq \Gamma$, $0 < \delta \leq 1/4$, and T, h satisfying Equation C.42 with $M = B^{1/p-1}(T/(256\delta))^{1/p}$, some normalized instance satisfies*

$$\mathbb{P}\left\{ \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|_2 \geq c_{\mathrm{res}}(\Gamma) \min\{1, T^{-(p-1)/(2p)} \delta^{-1/(2p)}\} \right\} > \delta.$$

Proof. Write $z = \sqrt{M/T}$. If $\gamma T \leq 1$, Lemma C.6 supplies a positive constant lower bound. Otherwise $1 + \gamma T/\epsilon \leq \gamma T(1 + 1/\epsilon)$, so Lemma C.7 supplies, with probability greater than δ, a lower bound

$$c_b \min\{1, z\sqrt{\min\{1, hM, \gamma M\}}\}, \tag{C.48}$$

where $c_b > 0$ is fixed; here the constant $s = 1/16$ has been absorbed into c_b.

Suppose first that $\gamma \leq C_0 h$, and put $x = \gamma M$. Then $hM \geq x/C_0$, so Equation C.48 is at least $c_b' \min\{1, z\sqrt{\min\{1, x\}}\}$ for a fixed $c_b' > 0$. Lemma C.6 supplies the alternative bound $c_{\mathrm{tr}} \min\{1, z/\sqrt{x}\}$. If $x \geq 1$, choose the bank instance; if $x < 1$, choose the travel instance. In either case the rate is at least a fixed multiple of $\min\{1, z\}$.

If $\gamma > C_0 h$, put instead $x = hM$. The bank lower is again at least a fixed multiple of $\min\{1, z\sqrt{\min\{1, x\}}\}$. Since $hT \geq 8$, Lemma C.4 gives a deterministic lower bound

$$\frac{c_{\mathrm{wb}}(\Gamma)}{\sqrt{1 + hT}} \geq c_w(\Gamma)(hT)^{-1/2} = c_w(\Gamma) z/\sqrt{x}.$$

The same division at $x = 1$ selects a witness with lower bound a fixed multiple of $\min\{1, z\}$. Every selection depends only on the numerical parameters and precedes sampling. The deterministic witnesses have probability one, and the bank witness has probability greater than δ. Since

$$z = B^{1/(2p)-1/2} 256^{-1/(2p)} T^{-(p-1)/(2p)} \delta^{-1/(2p)},$$

the claimed rate follows after taking a minimum of the finitely many positive constants. □

Lemma C.9 (Uniform rate with a fixed learning-rate cap)**.** *Let*

$$a = \frac{p-1}{3p-2}, \qquad 0 < q < \frac{a}{2p}, \qquad 0 < \Gamma < \infty$$

be fixed. There are $c_\Gamma > 0$ and $\delta_\Gamma \in (0, 1/4]$ such that, for every $T \geq 1$, $\beta_1 \in [0, 1)$, $0 < \gamma \leq \Gamma$, and $0 < \delta < \delta_\Gamma$, a normalized instance selected before sampling satisfies

$$\mathbb{P}\left\{ \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|_2 \geq c_\Gamma \min\{1, T^{-a} \delta^{-q}\} \right\} > \delta. \tag{C.49}$$

Proof. Choose a fixed $\tau > 0$ sufficiently small that

$$q < a q_0, \qquad q_0 = \frac{1}{2(p + \tau)},$$

and set $K = 1/q_0$ and $\zeta = a - Kq > 0$. Lemma C.1 supplies constants $c_{\mathrm{tail}}, \delta_{\mathrm{tail}} > 0$ and normalized instances satisfying

$$\mathbb{P}\left\{ \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|_2 \geq c_{\mathrm{tail}} \min\{1, \delta^{-q_0}/T\} \right\} > \delta \tag{C.50}$$

for every momentum and learning rate and every $0 < \delta < \delta_{\mathrm{tail}}$. Fix $\delta_\Gamma = \min\{\delta_{\mathrm{tail}}, 1/4\}$. When $\delta \le T^{-K}$, the minimum in Equation C.50 equals one. This proves the desired bound throughout that confidence range.

Now suppose $\delta > T^{-K}$. Then

$$T^{-a}\delta^{-q} \le T^{-\zeta} \le 1. \tag{C.51}$$

For the amplitude used in Lemma C.7,

$$M \le C_{\mathrm{amp}} T^{(1+K)/p}, \qquad C_{\mathrm{amp}} = B^{1/p-1}256^{-1/p}.$$

The definition of $k(M)$ therefore gives a fixed constant $C_k \ge 1$ such that

$$k(M) + 1 \le C_k \log(eT). \tag{C.52}$$

For example, one may take

$$C_k = 2 + \frac{\log_+((\nu + C_{\mathrm{amp}}^2)/s^2) + 2(1+K)/p}{-\log \beta_2}.$$

Write $c_h, C_h > 0$ for the constants of Lemma C.2, so its normalized instance satisfies

$$\mathbb{P}\{\frac{1}{T}\sum_{t=1}^{T} |f'(x_t)| \ge c_h h\} \ge 1 - T\exp(-C_h h^2 T), \qquad T \ge 2.$$

Put $c_k = 1/(64C_k)$. Choose a fixed integer $T_0 \ge 2$ so that

$$T\exp\left(-\frac{C_h c_k^2 T}{\log^2(eT)}\right) \le \frac{1}{4} \quad \text{for all } T \ge T_0. \tag{C.53}$$

Such a choice exists since the left side tends to zero. If $h(k(M)+1) > 1/64$, Equation C.52 implies $h > c_k/\log(eT)$. For $T \ge T_0$, the bias-floor event thus has probability at least $3/4 > \delta$, with rate at least

$$\frac{c_h c_k}{\log(eT)} \ge \frac{c_h c_k}{L_\zeta} T^{-\zeta} \ge \frac{c_h c_k}{L_\zeta} T^{-a}\delta^{-q}, \qquad L_\zeta = 1 + \frac{1}{e\zeta}.$$

Here $\log(eT) \le L_\zeta T^\zeta$ for $T \ge 1$, and the last inequality is Equation C.51.

If $h(k(M)+1) \le 1/64$ and $hT \ge 8$, apply Lemma C.8. For $1 < p \le 2$,

$$\frac{p-1}{2p} \le a, \qquad q < \frac{1}{2p}.$$

As $T \ge 1$ and $\delta < 1$, its capped rate dominates the capped rate in Equation C.49.

It remains to treat $hT < 8$. If $\gamma \le C_0 h$, then $\gamma T < 8C_0$, and Lemma C.6 gives the positive deterministic lower bound

$$c_{\mathrm{tr}} \min\{1, (8C_0)^{-1/2}\}.$$

If $\gamma > C_0 h$, Lemma C.4 gives the positive deterministic lower bound $c_{\mathrm{wb}}(\Gamma)/3$. The choice $C_0 \ge 4\Gamma$ ensures that this use has $h < 1/4$.

Finally, the finitely many horizons $T < T_0$ have a direct normalized deterministic witness: take $f(x) = x^2/2$ and $x_1 = \sqrt{2}$, with an exact oracle. The first evaluated gradient alone gives $T^{-1}\sum_{t=1}^{T} |f'(x_t)| \ge \sqrt{2}/T \ge \sqrt{2}/T_0$, uniformly in the numerical parameters.

Choose

$$c_\Gamma = \min\left\{c_{\mathrm{tail}}, \frac{\sqrt{2}}{T_0}, \frac{c_h c_k}{L_\zeta}, c_{\mathrm{res}}(\Gamma), c_{\mathrm{tr}}\min\{1, (8C_0)^{-1/2}\}, \frac{c_{\mathrm{wb}}(\Gamma)}{3}\right\} > 0.$$

The preceding cases cover every numerical tuple. In each case choose one corresponding objective and oracle before sampling; the dimension, gap, smoothness, and primitive moment bounds remain those of that instance. All constants were fixed after Γ and before T, δ, h, γ were chosen, which proves the required uniformity and Equation C.49. □

C.4 Uniform parameter coverage and Muon comparison

Proof of Theorem 3. Let $\Gamma_0, \alpha_0, c_{\mathrm{tent}} > 0$ be the fixed constants from Lemma C.5, enlarging Γ_0 if necessary so that $\Gamma_0 \geq \max\{1, 2\alpha_0\}$. Apply Lemma C.9 at the single fixed cap $\Gamma = \Gamma_0$, and denote its constants by c_{bd} and δ_{bd}. Apply Lemma C.3 at the fixed threshold α_0, and denote its positive lower constant by c_{eff}. Define

$$c = \min\{c_{\mathrm{bd}}, c_{\mathrm{tent}}, c_{\mathrm{eff}}\}, \qquad \bar{\delta} = \min\{\delta_{\mathrm{bd}}, 1/16\}.$$

All these constants depend only on the fixed quantities in the theorem.

Fix $T, \delta, \beta_1, \gamma$ in its stated domain. If $\gamma \leq \Gamma_0$, Lemma C.9 supplies the required normalized instance and the event in Equation 6. If $\gamma > \Gamma_0$ and $\gamma h \leq \alpha_0$, then $h < 1/2$. The deterministic instance in Lemma C.5 gives $T^{-1}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} \geq c_{\mathrm{tent}}$ with probability one. If $\gamma > \Gamma_0$ and $\gamma h > \alpha_0$, Lemma C.3 gives $T^{-1}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} \geq c_{\mathrm{eff}}$ with probability at least $1/8$. The last two thresholds dominate $c\min\{1, T^{-a}\delta^{-q}\}$ and their probabilities are strictly greater than $\delta < \bar{\delta}$. These three cases cover the entire domain, including $\beta_1 = 0$ and every positive learning rate.

Only one normalized instance is selected in each case. Consequently its dimension, initial gap and noise budget satisfy the common class bounds; none are added across constructions. Padding scalar instances gives the stated 1×2 row-matrix class and completes the proof. □

For the given budget T, choose Muon's parameters as $\beta = 1 - T^{-p/(3p-2)}$ and $\eta = (1-\beta)^{(2p-1)/p}$. Then $0 \leq \beta \leq 1 - 1/T$, and the noise-absorption condition holds since $\sigma_1 = 0$. Substituting the normalized class bounds into Equation 4 gives, with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(X_t)\|_{\mathrm{F}} \leq C_p\left[(1 + B^{-(p-1)/p})T^{-a} + (\sqrt{2}T^{-2a} + T^{-a})(1 + \log(1/\delta))\right].$$

Using $T^{-2a} \leq T^{-a}$ gives the Muon bound in the comparison paragraph of Section 3.3, with the uniform constant stated there. For $T^{-a/q} \leq \delta < \bar{\delta}$, the minimum in Equation 6 is unsaturated, and dividing the two thresholds gives the rate-threshold ratio stated at the end of Section 3.3.