ON THE OPTIMALITY OF ADAM UNDER HEAVY-TAILED NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Adam is the canonical optimizer in deep learning, and its decoupled weight decay variant AdamW is widely used for training large language models (LLMs). Meanwhile, Muon has recently achieved strong empirical gains over AdamW. One proposed explanation is robustness to heavy-tailed gradient noise: Muon admits sharp convergence rates, whereas Adam could behave poorly in such settings. This contrast motivates a basic question of whether Adam can converge optimally under heavy-tailed noise. We answer affirmatively through a stronger result for Adam-mini, a blockwise framework that encompasses Adam and Adam-Norm. Under heavy-tailed pth noise moment, we prove that Adam-mini attains the sharp rate $O(T^{-(p-1)/(3p-2)})$. Our theory also supports bias correction and weight decay, thereby covering AdamW as a special case. We further establish a matching lower bound that proves minimax optimality and, as a consequence, yields the first general stochastic ℓ_1 stationarity lower bound for smooth non-convex optimization with an unavoidable $\sqrt{d}$ dependence, where d is the problem dimension. Together, these results sharply characterize heavy-tailed convergence for Adam and its variants. Comparing the Adam directly with Muon under the same problem class, we show that, unlike prior justifications, heavy-tailed noise alone cannot explain Muon's empirical advantage.

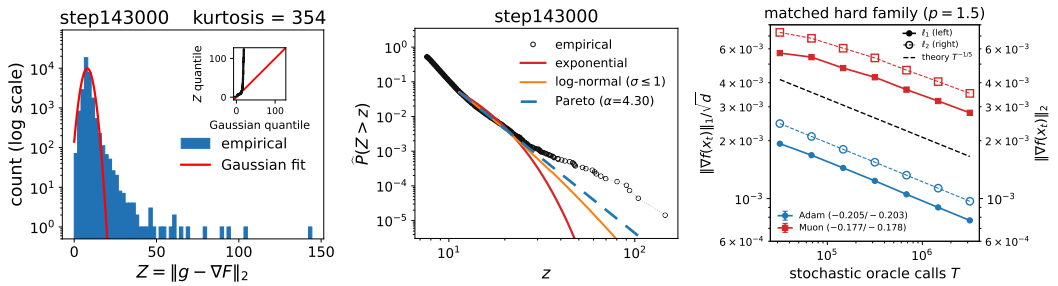
1 INTRODUCTION

Adam (Kingma & Ba, 2015) has been the de facto adaptive optimizer in deep learning for a decade. In modern foundation model training, its decoupled weight decay variant AdamW (Loshchilov & Hutter, 2019) is the standard practical form (Brown et al., 2020; Sun et al., 2021; Touvron et al., 2023a;b; Zeng et al., 2023; Grattafiori et al., 2024; DeepSeek-AI et al., 2024a;b; Sun et al., 2024; MiniMax et al., 2025). More recently, Muon (Jordan et al., 2024) has achieved strong empirical gains in LLM training, outperforming AdamW in several settings (Liu et al., 2025) and entering massive-scale industrial training recipes (Team et al., 2025; Zeng et al., 2025; 2026; DeepSeek-AI et al., 2026; Team et al., 2026b;a; Qiu et al., 2026).

This success of Muon naturally raises the question of why it can outperform the Adam family, and AdamW in particular, in LLM training. One proposed explanation is its robustness against *heavy-tailed gradient noise* (Yu et al., 2026d; Choudhury et al., 2026; Hübler et al., 2026). Muon's matrix-sign update discards singular value magnitudes, whereas Adam accumulates squared gradients that can retain the influence of large outliers (Kim, 2026). Heavy-tailed noise has been widely observed across modern learning systems in practice (Simsekli et al., 2019; Zhang et al., 2020; Zhou et al., 2020; Gurbuzbalaban et al., 2021; Garg et al., 2021; Battash et al., 2024; Ahn et al., 2024a; Yu et al., 2026d). Here, we directly test noise distributions along the pretraining trajectory of Pythia-6.9B model (Biderman et al., 2023). Figure 1a shows substantial departure from Gaussian, and Figure 1b shows pronounced heavier tails than exponential, lognormal or Pareto distributions.

Recent theory supports this explanation with sharp convergence rates for Muon under heavy-tailed noise (Yu et al., 2026d; Choudhury et al., 2026; Zhang & Lin, 2026; Hübler et al., 2026). Most existing Adam theories, however, require either finite variance or stronger noise assumptions (Zaheer et al., 2018; Guo et al., 2021; Zhang et al., 2022; Défossez et al., 2022; Li et al., 2023; Wang et al., 2023; Hong & Lin, 2024; Zhou et al., 2024b; Xie et al., 2025; Li et al., 2025a; Zhang et al., 2026).

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.



(a) Non-Gaussian noise. (b) Heavy-tailed noise. (c) Finite-sample tightness.

Figure 1: Heavy-tailed noise and finite-sample tightness. (a) Pythia-6.9B gradient deviations $Z_i = \|g_i - \widehat{\nabla F}_R\|_2$ at the final checkpoint have kurtosis **354**, versus **3** for a Gaussian. (b) Rare sample tails are heavier than exponential, lognormal, and Pareto references fitted above the 95th percentile. (c) Finite sample lower bound rates at $p = 1.5$, measured by average ℓ_1 (solid, left axis) and ℓ_2 (dashed, right axis) stationarity. Theorem-tuned Adam tracks $T^{-1/5}$, while Muon is flatter over this range. Error bars are 95% bootstrap intervals over 32 seeds. Full versions of (a)–(c) are in Figures 2, 3 and 6, respectively, with $p = 2$ results in Figure 7. Details are in Appendices G and H.

The heavy-tailed noise model, proposed in the seminal work of Zhang et al. (2020), relaxes finite variance to only a finite pth moment, as shown below.

Classical finite variance noise model
$\mathbb{E}_{t-1}\left[\|g_t - \nabla F(x_t)\|^2\right] \le \sigma^2$

Heavy-tailed noise model (Assumption 3)
$\mathbb{E}_{t-1}[\|g_t - \nabla F(x_t)\|^p] \le \sigma^p, \quad p \in (1, 2]$

For $p < 2$, heavy-tailed noise model generalizes finite variance model by permitting infinite variance. This allows rare event and large gradients to occur more often (Figures 1a and 1b), thus may better reflect the noise in practice than finite variance model. Under heavy-tailed noise model, Chezhegov et al. (2025) show that Adam-Norm can converge slowly to optima with high probability, and clipping remedies the issue. (Pang, 2026) also obtains sharp rates to first-order stationary points for clipped version of Adam, but without clipping, the convergence is significantly slower. (Liu, 2026b) analyzes vanilla Adam with randomized learning rates, but their rate is again far from optimal. These difficulties seem to suggest that there is a genuine barrier for vanilla Adam under heavy-tailed noise model, leading Yu et al. (2026c) to ask the following open problem.

OPEN PROBLEM Can the *standard, unclipped* Adam(W) recursion attain the minimax optimal convergence rate under heavy-tailed gradient noise?

We give an affirmative answer to this open problem through a stronger result for Adam-mini (Zhang et al., 2025b). Under the heavy-tailed noise model, we prove that Adam-mini attains the sharp $O(T^{-(p-1)/(3p-2)})$ rate. The practical result with bias correction further allows positive decoupled weight decay, so the guarantee directly covers AdamW as well. We also establish a matching lower bound against arbitrary randomized stochastic first-order methods that justifies the minimax optimality of our upper bounds. Experiments on this hard family make the predicted rate exponent visible at finite sample sizes (Figure 1c). These general results for Adam-mini immediately imply that **AdamW is also optimal under heavy-tailed noise**. In particular, our lower bound is the *first ℓ_1-norm lower bound* for non-convex stochastic optimization that extends the well-known ℓ_2 bound in Arjevani et al. (2023). This immediately confirms that prior upper bounds (under finite variance model) for AdamW (Li et al., 2025a), RMSProp (Li et al., 2025b), SignSGD (Jiang et al., 2025b), and Lion (Jiang et al., 2026) are optimal, resolving the conjecture in Li et al. (2025a).

Another interesting implication of our theory is the direct comparison with Muon, which is a natural robustifier against heavy-tailed noise. Concretely, Yu et al. (2026d); Choudhury et al. (2026); Zhang & Lin (2026); Hübler et al. (2026) show that Muon attains the optimal $O(T^{-(p-1)/(3p-2)})$ rate under heavy-tailed noise model. Based on negative evidence for Adam-type methods and their lack of a sharp convergence guarantee (Chezhegov et al., 2025; Yu et al., 2026c; Kim, 2026), these works argue that Muon's robustness to heavy-tailed noise, which is prevalent in LLM training, may be one of the reasons to its empirical success. In contrast, we **refute this claim** by directly comparing Adam with Muon under the same problem class in Section 5. Both methods achieve the same minimax optimal rate in T, therefore heavy-tailed noise alone cannot explain Muon's empirical advantage.

Our contributions are listed as follows. See Table 2 for a detailed comparison with prior work.

- We prove the sharp $O(T^{-(p-1)/(3p-2)})$ heavy-tailed rate for Adam-mini, yielding the corresponding optimal guarantee for Adam. Our bias-corrected analysis additionally permits weight decay, thereby resolving the AdamW open problem of Yu et al. (2026c) within the same framework.
- We establish matching lower bounds against arbitrary randomized stochastic first-order methods, confirming Adam-mini's minimax optimality. At $p = 2$, we obtain the first tight ℓ_1 stationarity lower bound and prove the $\sqrt{d}$ factor necessary. This resolves the conjecture of Li et al. (2025a) and establishes the optimality of prior rates for AdamW, RMSProp, SignSGD, and Lion.
- We directly compare AdamW and Muon under a common Euclidean heavy-tailed oracle and show that both attain the same optimal rate (Section 5 and Table 3). This directly disproves prior believe and imply that heavy-tailed noise alone therefore cannot explain Muon's empirical advantage.

2 Problem Setting and Adam-mini Framework

In this section, we first state notations and assumptions used throughout this paper. Then, we introduce the Adam-mini algorithmic framework and the geometry induced by its blockwise adaptation.

Notation. For a positive integer n, let $[n] := \{1, \dots, n\}$. We use $\|\cdot\| = \|\cdot\|_2$ for the Euclidean norm and $\langle \cdot, \cdot \rangle$ for the Euclidean inner product. The expectation conditioned on the history $\mathcal{F}_{t-1}$ is denoted by $\mathbb{E}_{t-1}[\cdot]$. Given a partition $\mathcal{P} = \{1, \dots, J\}$ of $\mathbb{R}^d$ with block dimensions $d_1, \dots, d_J$, we write $z^{(j)} \in \mathbb{R}^{d_j}$ for block j of $z \in \mathbb{R}^d$, where $d = \sum_{j=1}^{J} d_j$ and $d_{\min} := \min_j d_j$. Constants in asymptotic notation may depend only on p, while $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ also hide logarithmic factors.

We employ the assumptions below. Assumptions 1 and 2 are standard in non-convex optimization (Lan, 2020; Arjevani et al., 2023), while Assumption 3 is the most common heavy-tailed noise model adopted in optimization literature (Zhang et al., 2020; Cutkosky & Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023; Liu et al., 2023; 2024; Liu & Zhou, 2025; Liu, 2026c; Pang, 2026).

Assumption 1 (Lower-bounded objective)**.** *The objective $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and bounded from below, with $F(x_1) - \inf_{x \in \mathbb{R}^d} F(x) \le \Delta$ for some $\Delta > 0$.*

Assumption 2 (Smoothness)**.** $\|\nabla F(x) - \nabla F(y)\| \le L \|x - y\|$ *for all $x, y \in \mathbb{R}^d$ and some $L > 0$.*

Assumption 3 (Heavy-tailed stochastic oracle)**.** *At iteration t, the algorithm observes a mini-batch $\{g_t^1, \dots, g_t^B\}$ of conditionally independent stochastic gradients. For some $p \in (1, 2]$ and $\sigma > 0$,*

$$\mathbb{E}_{t-1}\left[g_t^b\right] = \nabla F(x_t), \qquad \mathbb{E}_{t-1}\left[\left\|g_t^b - \nabla F(x_t)\right\|^p\right] \le \sigma^p, \qquad b \in [B].$$

Here $\mathcal{F}_0$ contains the initialization, hyperparameters, and the algorithm's internal random seed, and $\mathcal{F}_t := \sigma(\mathcal{F}_0, \{g_s^b : s \le t,\ b \in [B]\})$ is the natural filtration. In particular, x_t is $\mathcal{F}_{t-1}$-measurable.

2.1 Adam-mini

Adam-mini (Zhang et al., 2025b) keeps the Adam first-moment recursion but shares one common second-moment scale within each block, realized by $v_t^{(j)}$. The scale for block j uses $\|g_t^{(j)}\|^2/d_j$, so

Algorithm 1 Adam-mini (without bias correction) (Zhang et al., 2025b)

Require: $x_1 \in \mathbb{R}^d$, $\{\eta_t\}_{t \ge 1}$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon_{\text{alg}} > 0$, $\lambda \ge 0$ and $B \in \mathbb{N}_+$

Initialize $m_0 = 0$ and $v_0^{(j)} = 0$ for every $j \in [J]$
for $t = 1, \dots, T$ **do**
  Draw $g_t = B^{-1} \sum_{b=1}^{B} g_t^b$
  for $j = 1, \dots, J$ **do**
    $m_t^{(j)} = \beta_1 m_{t-1}^{(j)} + (1 - \beta_1) g_t^{(j)}, \quad v_t^{(j)} = \beta_2 v_{t-1}^{(j)} + (1 - \beta_2) \|g_t^{(j)}\|^2 / d_j$
    $x_{t+1}^{(j)} = (1 - \eta_t \lambda) x_t^{(j)} - \eta_t m_t^{(j)} / (\sqrt{v_t^{(j)}} + \epsilon_{\text{alg}})$
  end for
end for

Table 1: Partition endpoints of Adam-mini and stationarity criteria. Full algorithms in Appendix B.

Partition	Method	Stationarity criterion $\mathsf{S}_{\mathcal{P}}(h)$
Singleton blocks	Adam/AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019)	$\|h\|_1/\sqrt{d}$
One block	AdaSGD/Adam-Norm (Wang & Wiens, 2020; Wang et al., 2024a)	$\|h\|$
General $\mathcal{P}$	Adam-mini (Zhang et al., 2025b; Zhao et al., 2025)	$d^{-1/2}\sum_j \sqrt{d_j}\,\|h^{(j)}\|$

the partition controls both memory use and adaptivity. Algorithms 1 and 2 give the two formulations analyzed in Sections 3.1 and 3.2, respectively. When the blocks are singletons ($J = d$), the adaptive part is exactly Adam; with positive decoupled weight decay it becomes AdamW. If there is only one block ($J = 1$), Algorithm 1 reduces to AdaSGD (Wang & Wiens, 2020) or Adam-Norm (Wang et al., 2024a). Section 3.3 explains these reductions in detail. The complete algorithms are in Appendix B.

2.2 Natural Block Geometry

The average squared gradient in each block makes the update scale with $\sqrt{d_j}$. Motivated by Xie et al. (2025), we use the resulting mixed-norm geometry to measure directions and stationarity.

Definition 1 (Adam-mini natural geometry (Xie et al., 2025)). *For $u, h \in \mathbb{R}^d$, define*

$$\|u\|_{\mathcal{P}} := \sqrt{d}\max_{j\in[J]}\frac{\|u^{(j)}\|}{\sqrt{d_j}}, \qquad \mathsf{S}_{\mathcal{P}}(h) := \frac{1}{\sqrt{d}}\sum_{j=1}^{J}\sqrt{d_j}\,\|h^{(j)}\|. \tag{1}$$

We measure stationarity by $\mathsf{S}_{\mathcal{P}}(\nabla F(x))$, the dual norm matching the shared denominator scaling.

Lemma 1 (Exact duality). *The norms in Eq. (1) satisfy $\mathsf{S}_{\mathcal{P}}(h) = \sup_{\|u\|_{\mathcal{P}}\le 1}\langle h, u\rangle$.*

Proof. If $\|u\|_{\mathcal{P}} \le 1$, then $\|u^{(j)}\| \le \sqrt{d_j/d}$ for every j, and blockwise Cauchy–Schwarz gives $\langle h, u\rangle \le d^{-1/2}\sum_{j=1}^{J}\sqrt{d_j}\,\|h^{(j)}\| = \mathsf{S}_{\mathcal{P}}(h)$. Equality is attained by taking each nonzero $u^{(j)}$ parallel to $h^{(j)}$ with norm $\sqrt{d_j/d}$. □

Table 1 summarizes the interpolation between ℓ_1 to ℓ_2 stationarity for AdamW and Adam-Norm. The endpoint conventions and corresponding convergence guarantees are stated in Section 3.3.

3 Upper Bounds

This section presents our main upper bound results, with proofs deferred to Appendices C and D.

3.1 Convergence of Algorithm 1

To isolate the heavy-tailed difficulty, we first consider Algorithm 1 with $\lambda = 0$. We then analyze Algorithm 2 in Section 3.2, which includes bias correction and positive decoupled weight decay.

Theorem 1. *Suppose Assumptions 1 to 3 hold. Run Algorithm 1 with $\lambda = 0$ and $\eta_t \equiv \eta$, and set*

$$1-\beta_2 = \left(\frac{\Delta L B^{2(p-1)/p}}{2^{2/p}\sigma^2 T}\right)^{\frac{p}{3p-2}},\ \beta_2^2 \le \beta_1 \le \beta_2, \eta = \sqrt{\frac{\Delta}{LdT}},\ \epsilon_{\text{alg}} = \frac{2^{\frac{2p+1}{p}}\sigma(1-\beta_2)^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{d}\,B^{(p-1)/p}}. \tag{2}$$

Assume

$$\sqrt{\frac{d}{d_{\min}}}(1-\beta_2) \le \frac{1}{8}. \tag{3}$$

Then

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(x_t))\right] \le 42\sqrt{\frac{\Delta L}{T}} + 25\,2^{\frac{1}{3p-2}}\frac{(\Delta L)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}}{(BT)^{\frac{p-1}{3p-2}}}. \tag{4}$$

For singleton blocks, Theorem 1 resolves the open problem of Yu et al. (2026c) by proving the optimal $T^{-(p-1)/(3p-2)}$ rate (Liu & Zhou, 2025) for AdamW with $\lambda = 0$. As Yu et al. (2026c, Remark 1) emphasize, the key question is the optimal dependence on T, which we already establish.

Remark 1 (Hyperparameters). Theorem 1 requires the tuning in Eq. (2), which matches the finite-variance AdamW tuning of Li et al. (2025a) at $p = 2$ up to constants. However, similar to Li et al. (2025a), Theorem 1 excludes the common default $(\beta_1, \beta_2) = (0.9, 0.999)$ or $(0.9, 0.95)$, because it violates $\beta_2^2 \le \beta_1 \le \beta_2$. As discussed in Li et al. (2025a, Section 2.5), commonly used Adam parameters can fail to converge on constructed instances (Zhang et al., 2022; Wang et al., 2024b).

Remark 2 (Proof challenges). An old noise outlier can keep the denominator v_t inflated after its contribution to momentum has faded, suppressing later updates (Yu et al., 2026c; Kim, 2026). Heavy-tailed noise makes this effect more severe: when $p < 2$, large outliers occur more frequently and the second moment may be infinite, so v_t cannot be controlled through standard variance bounds. We therefore decompose $g_t = g_{t,\mathrm{core}} + g_{t,\mathrm{tail}}$ as in Eqs. (29) and (30). The **core** admits a controlled second moment, behaving similarly to finite variance gradients. While the **tail** event is rare (Lemma C.3 bounds its probability) and is handled directly through self-normalization (Lemma C.4).

Theorem 2 (Complexity of Theorem 1)**.** *Let $\varepsilon > 0$, and draw $\widehat{t}$ uniformly from $[T]$, independently of the algorithm and oracle, with $\widehat{x} := x_{\widehat{t}}$. It is sufficient to take*

$$T \ge \max\left\{ 84^2 \frac{\Delta L}{\varepsilon^2}, 50^{\frac{3p-2}{p-1}} 2^{\frac{1}{p-1}} \frac{\Delta L}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}} \frac{1}{B}, \frac{\Delta L B^{2(p-1)/p}}{2^{2/p}\sigma^2} \left(8\sqrt{\frac{d}{d_{\min}}}\right)^{\frac{3p-2}{p}} \right\}. \tag{5}$$

Then the tuning in Eq. (2) *gives $\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(\widehat{x}))\right] \le \varepsilon$. Moreover, suppose that the accuracy satisfies*

$$\varepsilon \le \frac{2^{1/p}\sigma}{B^{(p-1)/p}} \min\left\{ 50^{\frac{3p-2}{p}} 84^{-\frac{2(p-1)}{p}}, 50\left(8\sqrt{d/d_{\min}}\right)^{-\frac{p-1}{p}} \right\}. \tag{6}$$

Then all three requirements in Eq. (5) *are met, with complexity satisfying*

$$BT = B\left\lceil \frac{50^{\frac{3p-2}{p-1}} 2^{\frac{1}{p-1}}}{B} \frac{\Delta L}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}} \right\rceil \le B + 50^{\frac{3p-2}{p-1}} 2^{\frac{1}{p-1}} \frac{\Delta L}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}. \tag{7}$$

Theorem 2 implies that Algorithm 1 achieves the optimal complexity in ϵ (Zhang et al., 2020; Liu & Zhou, 2025). For tightness w.r.t. all parameters, we defer the results to Section 4.

Remark 3 (Geometric burn-in). Under Eq. (2), the geometric condition Eq. (3) is equivalent to $T \ge \frac{\Delta L B^{2(p-1)/p}}{2^{2/p}\sigma^2}(8\sqrt{d/d_{\min}})^{(3p-2)/p}$. This is the third requirement in Eq. (5), ensuring that the block-geometry term can be absorbed. Under Eq. (6), it is covered by the stochastic budget.

3.2 Convergence of Algorithm 2

Now we go a step further to consider Algorithm 2 with bias correction, with the update

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \widehat{v}_t^{(j)} = \frac{v_t^{(j)}}{1-\beta_2^t}, \quad x_{t+1}^{(j)} = (1-\eta_t\lambda)x_t^{(j)} - \eta_t \widehat{m}_t^{(j)} \Big/ \left(\sqrt{\widehat{v}_t^{(j)}} + \epsilon_{\mathrm{bc}}\right). \tag{8}$$

Its convergence guarantee is stated below.

Theorem 3. *Suppose Assumptions 1 to 3 hold and $\sigma > 0$. Run Algorithm 2 with $\eta_t \equiv \eta$ and*

$$\begin{aligned} 1-\beta_2 &= \left(\frac{\Delta L B^{2(p-1)/p}}{2^{2/p}\sigma^2 T}\right)^{\frac{p}{3p-2}}, \quad \beta_2^2 \le \beta_1 \le \beta_2, \quad \eta = \sqrt{\frac{\Delta}{LdT}}, \\ \epsilon_{\mathrm{bc}} &= \frac{2^{3+1/p}\sigma(1-\beta_2)^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{d}\, B^{(p-1)/p}}, \quad 0 \le \lambda \le \frac{\sqrt{1-\beta_2}}{8(R_1 + \sqrt{2}\eta T)}. \end{aligned} \tag{9}$$

where $R_1 := \max_j \left\|x_1^{(j)}\right\| / \sqrt{d_j}$. Assume also Eq. (3)*. Define*

$$\mathfrak{r}_T := \left(2 + \log\min\left\{T, (1-\beta_2)^{-1}\right\}\right)\left[\frac{1}{T(1-\beta_2)} + \sqrt{\frac{1-\beta_2}{T}}\right]. \tag{10}$$

Then

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(x_t))\right] \le 100\left\{(1+\mathfrak{r}_T)\sqrt{\frac{\Delta L}{T}} + \sqrt{1+\mathfrak{r}_T}\, 2^{\frac{1}{3p-2}} \frac{(\Delta L)^{\frac{p-1}{3p-2}} \sigma^{\frac{p}{3p-2}}}{(BT)^{\frac{p-1}{3p-2}}}\right\}. \tag{11}$$

For fixed $p, \Delta, L, \sigma, B, d$, and the partition, the exact tuning gives $\mathfrak{r}_T = o(1)$.

Table 2: Prior theory and our result. Rates give the leading T-dependence in each work's stationarity criterion, with other parameters fixed. We cover heavy-tailed noise, standard bias correction, and positive decoupled weight decay with the vanilla recursion.

Work	Algorithm	Rate in T	HT	BC	WD	Vanilla	Limitation
Li et al. (2025a)	AdamW	$O(T^{-1/4})$	✗	✗	✓	✗	finite variance
Xie et al. (2025)	Adam-mini	$\widetilde{O}(T^{-1/4})$	✗	✗	✗	✗	no momentum
Chezhegov et al. (2025)	Adam-Norm	$\widetilde{O}(T^{-\frac{p-1}{3p-2}})$	✓	✗	✗	✗	clipping
Liu (2026b)	Adam	$O(T^{-\frac{2(3p-4)}{19p-12}})$	✓	✓	✗	✗	random LR
Pang (2026)	Adam	$O(T^{-\frac{2(p-1)}{5p-3}})$	✓	✓	✗	✗	clipping
Ours (Theorem 3)	**Algorithm 2**	$O(T^{-\frac{p-1}{3p-2}})$	✓	✓	✓	✓	none above

HT, BC, and WD denote coverage of heavy-tailed noise with $p < 2$, standard bias correction, and positive decoupled weight decay, while Vanilla requires the unmodified recursion. Liu (2026b); Pang (2026) concern nonsmooth objectives and expected generalized ℓ_1 and ℓ_2 stationarity, respectively, with localization parameters fixed. Liu (2026b, Corollary 2) requires $\beta_1 = \beta_2$ and $p > 4/3$, giving $O(T^{-2/13})$ at $p = 2$. Pang (2026, Corollary 6.5) constrains its vector conversion learner to a known-radius ball and clips its output. The rate reduces to $O(T^{-2/7})$ at $p = 2$.

The rate in Eq. (11) is essentially the same as in Theorem 1. The extra factor $\mathfrak{r}_T$ comes from bias correction and vanishes under the stated tuning. Thus bias correction preserves the leading exponent, while Eq. (9) permits positive weight decay without changing the stationarity target.

Remark 4 (Weight decay and stationarity). The decay restriction in Eq. (9) allows the preceding theorem to control stationarity of the unregularized objective F. The parameter λ is fixed within each run, but its admissible upper bound decreases with T. For fixed problem parameters at $p = 2$, its $T^{-3/4}$ scaling matches Li et al. (2025a). The result therefore does not cover arbitrary fixed positive decay. As discussed in Li et al. (2025a, Section 2.6), this is consistent with the constrained interpretation of AdamW (Xie & Li, 2024): KKT stationarity on an ℓ_∞ ball need not imply $\nabla F = 0$.

Theorem 4 (Complexity of Theorem 3)**.** *Suppose Assumptions 1 to 3 hold and $\sigma > 0$. Let $\varepsilon > 0$, let $\widehat{t}$ be uniform on $[T]$ independently of the algorithm and oracle, and put $\widehat{x} := x_{\widehat{t}}$. Suppose*

$$\varepsilon \le \min\left\{ \frac{2^{1/p}\sigma}{B^{(p-1)/p}} \left(\frac{d_{\min}}{d}\right)^{\frac{p-1}{2p}}, \frac{100(p-1)\sqrt{\Delta L}}{p\left[1+\log\left(1+\frac{2^{3/2+1/p}p\sigma}{(p-1)B^{(p-1)/p}\sqrt{\Delta L}}\right)\right]} \right\}. \tag{12}$$

Under these conditions, the tuning in Eq. (9) *gives $\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(\widehat{x}))\right] \le \varepsilon$ with complexity*

$$BT = B\left\lceil \frac{2^{1/(p-1)}}{B}(200\sqrt{2})^{\frac{3p-2}{p-1}} \frac{\Delta L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}} \right\rceil \le B + 2^{\frac{1}{p-1}}(200\sqrt{2})^{\frac{3p-2}{p-1}} \frac{\Delta L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}. \tag{13}$$

The complexity of Algorithm 2 is now explicit and matches that of Algorithm 1 in Theorem 2. Thus the same heavy-tailed rate extends from Adam to the AdamW with bias correction and postive weight decay, formally settling the AdamW open problem in Yu et al. (2026c), with optimal dependence on ε (Liu & Zhou, 2025). Later in Section 4, we establish its optimality w.r.t. all parameters.

3.3 Reductions to Adam, Adam-Norm, and Adam-mini

We briefly discuss how the general framework specializes to Adam, Adam-Norm, and Adam-mini, and how the resulting rates compare with prior theory. See Table 2 for detailed comparisons.

Singleton blocks recover Adam. For $\mathcal{P} = \{\{1\}, \dots, \{d\}\}$, each coordinate has its own second moment, so Adam-mini reduces to coordinatewise Adam when $\lambda = 0$ and to AdamW with decoupled weight decay; in either case $\mathsf{S}_{\mathcal{P}}(\nabla F) = \|\nabla F\|_1/\sqrt{d}$. Theorems 1 and 3 give the sharp heavy-tail rate, including bias correction and admissible positive weight decay. Algorithm 3 is the singleton-block specialization of Algorithm 2.

The closest comparison is Li et al. (2025a). Their fixed coordinatewise variance bounds σ_i^2 are stronger than Assumption 3 at $p = 2$ with the same total budget $\sigma^2 = \sum_{i=1}^d \sigma_i^2$, since our condition controls only the total conditional variance. At $p = 2$ and $B = 1$, we recover their full rate dependence on d, T, Δ, L, σ in the stochastic regime, and additionally cover standard bias correction, which their analysis omits. Our guarantee also extends to $p < 2$.

One block gives AdaSGD and Adam-Norm. For $\mathcal{P} = \{[d]\}$, the second moment is shared across all coordinates and $\mathsf{S}_{\mathcal{P}}(\nabla F) = \|\nabla F\|_2$. This gives AdaSGD (Wang & Wiens, 2020) and, after rescaling, Adam-Norm (Wang et al., 2024a). Thus Theorem 3 apply directly to Algorithm 4, the one-block specialization of Algorithm 2.

Chezhegov et al. (2025, Theorem 2.1) identify a genuine failure mode of Adam-Norm: a rare early outlier can make the variance buffer v_t large and suppress many later updates. Indeed, a stochastic gradient g_s contributes proportionally to β_2^k to v_{s+k}. We call the constant-factor decay time $\tau_{\mathrm{mem}} = 1/(-\log \beta_2) \asymp (1-\beta_2)^{-1}$ the *memory length*. Their choice $\beta_2 = 1 - 1/T$ gives $\tau_{\mathrm{mem}} \asymp T$, so *one outlier can slow down essentially the whole run*. Our tuning in Eq. (9) instead gives $\tau_{\mathrm{mem}} = \Theta(T^{p/(3p-2)}) = o(T)$ for fixed problem parameters, so *an outlier only affects a sublinear part of the run*. Its remaining effect within this shorter window is controlled by the core–tail and self-normalization arguments in Remark 2; see also Lemmas C.3 and C.4 in Appendix C. Therefore, the high probability negative result in Chezhegov et al. (2025) does not contradict our theory, because we directly overcome such failure mode by our careful β_2 tuning and new proof techniques.

General blocks give Adam-mini. For general partitions, Theorems 1 and 3 measure stationarity in $\mathsf{S}_{\mathcal{P}}$. At $p = 2$, our $O(T^{-1/4})$ rate recovers the $T^{-1/4}$ dependence of the finite-variance analysis of Xie et al. (2025), without its *logarithmic factor* and with *nonzero momentum*. Our guarantee further covers heavy-tailed noise, standard bias correction, and positive decoupled weight decay.

4 Lower Bounds and Optimality

We establish minimax optimality through a partition-uniform lower bound against arbitrary randomized stochastic first-order methods. At the AdamW endpoint, it also yields a sharp stochastic ℓ_1 stationarity lower bound with an explicit $\sqrt{d}$ factor. All proofs are postponed to Appendix E.

Oracle model. Fix p, Δ, L, σ, the dimension d, and a partition $\mathcal{P}$. Let $\mathfrak{C}_p(\Delta, L, \sigma; d)$ be the class of objective and oracle pairs satisfying the following conditions. The objective $F : \mathbb{R}^d \to \mathbb{R}$ is bounded below, has L-Lipschitz gradient, and obeys $F(0) - \inf_x F(x) \le \Delta$. For every $x \in \mathbb{R}^d$, the stochastic gradient oracle satisfies

$$\mathbb{E}_\xi\left[\mathsf{O}(x,\xi)\right] = \nabla F(x), \qquad \mathbb{E}_\xi\left[\|\mathsf{O}(x,\xi) - \nabla F(x)\|^p\right] \le \sigma^p. \tag{14}$$

Each call at x_t uses an independent oracle seed ξ_t and returns only $G_t := \mathsf{O}(x_t, \xi_t)$. Write $\mathsf{H}_t := ((x_1, G_1), \dots, (x_t, G_t))$, with $\mathsf{H}_0 = \varnothing$. A randomized method has an internal seed R_{A}, independent of the oracle seeds, chooses x_t measurably from $(R_{\mathsf{A}}, \mathsf{H}_{t-1})$, and after n calls returns $\widehat{x}_{\mathsf{A}}$ measurably from $(R_{\mathsf{A}}, \mathsf{H}_n)$. An unqueried output counts as one extra oracle call, so $\mathrm{cost}(\mathsf{A}) := n + \mathbb{I}_{\{\widehat{x}_{\mathsf{A}} \notin \{x_1, \dots, x_n\}\}}$. The origin point $0 \in \mathbb{R}^d$ only fixes the initial-gap reference. Replacing $F(x)$ by $F(x - x_{\mathrm{init}})$ and shifting the oracle likewise gives the same model around any prescribed x_{init}. Let $\mathfrak{A}_N$ denote all such methods with $\mathrm{cost}(\mathsf{A}) \le N$ almost surely. Define

$$N^*(\varepsilon; \mathcal{P}) := \inf\left\{ N \in \mathbb{N} : \inf_{\mathsf{A} \in \mathfrak{A}_N} \sup_{(F,\mathsf{O}) \in \mathfrak{C}_p(\Delta, L, \sigma; d)} \mathbb{E}_{\mathsf{A},\mathsf{O}}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(\widehat{x}_{\mathsf{A}}))\right] \le \varepsilon \right\}. \tag{15}$$

Theorem 5. *For suitable p-dependent constants $c_p, C_p > 0$ and any partition, suppose*

$$0 < \varepsilon \le c_p \min\{\sigma, \sqrt{\Delta L}\}, \qquad d \ge C_p R_\varepsilon \log(2 + R_\varepsilon Q_\varepsilon), \tag{16}$$

where $R_\varepsilon = \Delta L/\varepsilon^2$ and $Q_\varepsilon = (\sigma/\varepsilon)^{p/(p-1)}$. Then

$$N^*(\varepsilon; \mathcal{P}) \ge c_p R_\varepsilon Q_\varepsilon = c_p \frac{\Delta L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}. \tag{17}$$

We defer the refinement for smaller dimensions to Appendix E.5. The key geometric ingredient in Theorem 5 is the following lemma: *every partition admits a large subspace on which $\mathsf{S}_{\mathcal{P}}$ behaves like the Euclidean norm*. The isometry $U : \mathbb{R}^r \to \mathbb{R}^d$ has orthonormal columns spanning this subspace, so $z \mapsto Uz$ embeds $\mathbb{R}^r$ without changing Euclidean lengths. Lifting a hard objective as $F(x) = f(U^\top x)$ places all gradients $U\nabla f(U^\top x)$ in this subspace.

Lemma 2 (Isometry subspace embedding). *There are universal constants $c_{\dim}, c_{\sec} > 0$ such that, for every partition $\mathcal{P}$ of $\mathbb{R}^d$ and every integer $r \le c_{\dim} d$, there is a isometry $U : \mathbb{R}^r \to \mathbb{R}^d$ satisfying*

$$c_{\sec} \|z\| \le \mathsf{S}_{\mathcal{P}}(Uz) \le \|z\| \qquad \text{for every } z \in \mathbb{R}^r. \tag{18}$$

Here isometry means $U^\top U = I_r$, or equivalently $\|Uz\| = \|z\|$ for every $z \in \mathbb{R}^r$.

Remark 5 (Why the partition causes no loss). A global comparison between $\mathsf{S}_{\mathcal{P}}$ and the Euclidean norm can incur a dimension-dependent loss. Lemma 2 avoids this by embedding the Euclidean hard instance in a large subspace where the two norms are equivalent up to universal constants. This isometric lift preserves the gap, smoothness, and noise moments, and arbitrary ambient queries can be simulated by projection without extra oracle calls (Appendix E.4).

Combining Theorem 5 with the complexity upper bounds in Theorems 2 and 4, using $x_1 = 0$ to match the initial-gap convention, yields the following minimax characterization.

Theorem 6 (Minimax optimality). *Set $B = 1$ and fix a partition $\mathcal{P}$. Under Eq.* (16)*, and under the corresponding accuracy condition Eq.* (6) *for Algorithm 1 or Eq.* (12) *for Algorithm 2,*

$$N^*(\varepsilon; \mathcal{P}) = \Theta\Big(\Delta L \sigma^{\frac{p}{p-1}} \varepsilon^{-\frac{3p-2}{p-1}}\Big).$$

Thus, in the stated regime, we completely characterize the minimax complexity in $\varepsilon, \Delta, L, \sigma$ for every partition, up to constants depending only on p. Algorithms 1 and 2 both attain this complexity, without logarithmic or partition-dependent losses. We also test this rate on the hard family, with details in Appendix H. Figures 1c and 7 show that AdamW's fitted slopes of -0.205 at $p = 1.5$ and -0.253 at $p = 2$ closely follow the predicted $-1/5$ and $-1/4$, making the tightness visible at finite sample sizes.

The first general ℓ_1 stationarity lower bound. Li et al. (2025a, Section 2.1) conjecture that the ℓ_1 stationarity lower bound for constant-step-size SGD (Jiang et al., 2025a) extends to general stochastic first-order methods, making the stochastic $\sqrt{d}\,T^{-1/4}$ scaling unavoidable. The following corollary resolves this stochastic dimension-dependence question in our Euclidean bounded-variance model. To our knowledge, it gives the **first tight** ℓ_1 stationarity lower bound against arbitrary randomized stochastic first-order methods.

Corollary 7 (ℓ_1 stationarity lower bound). *There are universal constants $c, C > 0$ such that, for $\Delta, L, \sigma > 0$ with $T \ge C \max\{\Delta L/\sigma^2, \sigma^2/(\Delta L)\}$ and $d \ge C\sqrt{\Delta L T}\log(2+T)/\sigma$, every randomized stochastic first-order method of cost at most T admits an L-smooth instance with initial gap at most Δ and unbiased noise variance at most σ^2 satisfying*

$$\mathbb{E}\,\|\nabla F(\widehat{x})\|_1 \ge c\sqrt{d}\left(\frac{\Delta L \sigma^2}{T}\right)^{1/4}. \tag{19}$$

This follows from Theorem 5 with $p = 2$, singleton blocks, and $\varepsilon = \kappa(\Delta L \sigma^2/T)^{1/4}$ for a sufficiently small $\kappa > 0$, using $\mathsf{S}_{\mathcal{P}}(\nabla F) = \|\nabla F\|_1/\sqrt{d}$. The stochastic $\sqrt{d}\,T^{-1/4}$ scaling is therefore tight in the stated oracle model and regime: it matches the rates of momentum SignSGD (Jiang et al., 2025b) and Lion (Jiang et al., 2026), as well as AdamW (Li et al., 2025a) and RMSProp (Li et al., 2025b) (up to logarithmic factors for RMSProp).

5 Comparison of AdamW and Muon, and Implications

We now specialize the general Adam-mini theory to its AdamW endpoint and assess whether heavy-tailed noise can explain Muon's empirical advantage. We establish matching upper and lower bounds for Muon under the same assumptions, then compare the two methods' minimax rates.

For $X \in \mathbb{R}^{m\times n}$, vectorization turns Assumptions 1 to 3 into equivalent Frobenius assumptions (Assumption 4) with the same constants. Put $d = mn$, $r = \min\{m, n\}$, and $\mathcal{N}(G) = \|G\|_*/\sqrt{r}$. For a gradient matrix G, the comparison uses $\|\operatorname{vec}(G)\|_1/\sqrt{d}$ for AdamW and $\mathcal{N}(G)$ for Muon. We analyze exact-polar Muon under this common oracle using

$$M_t = \beta M_{t-1} + (1-\beta)G_t, \qquad X_{t+1} = X_t - \eta\,\operatorname{msign}(M_t), \qquad M_0 = 0. \tag{20}$$

Here G_t is the mini-batch gradient and $\operatorname{msign}(M)$ denotes the compact polar factor of M, with $\operatorname{msign}(0) = 0$. The complete algorithm using Newton–Schulz is listed in Algorithms 5 and 6.

Table 3: AdamW and Muon under the common Euclidean oracle with $B = 1$. The joint regime is stated in Appendix F.4. Here $G = \nabla F(X)$, $d = mn$, and $r = \min\{m, n\}$.

Method	Natural stationarity	Minimax complexity
AdamW	$\|\operatorname{vec}(G)\|_1/\sqrt{d}$	$\Theta(\Delta L\sigma^{p/(p-1)}\varepsilon^{-(3p-2)/(p-1)})$
Muon	$\|G\|_*/\sqrt{r}$	$\Theta(\Delta L\sigma^{p/(p-1)}\varepsilon^{-(3p-2)/(p-1)})$

Theorem 8 (Heavy-tailed Muon, full version in Theorem 11)**.** *Under Assumptions 1 to 3 after vectorization, run Eq.* (20) *for $T \geq 1$ iterations with the hyperparameters in Eq.* (163)*. Then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\mathcal{N}(\nabla F(X_t)) = O\Bigg(\sqrt{\frac{\Delta L}{T}} + \frac{2^{\frac{1}{3p-2}}(\Delta L)^{\frac{p-1}{3p-2}}\sigma^{\frac{p}{3p-2}}}{(BT)^{\frac{p-1}{3p-2}}} + \frac{2^{\frac{2}{3p-2}}\sigma^{\frac{2p}{3p-2}}}{(\Delta L)^{\frac{2-p}{2(3p-2)}}(BT)^{\frac{2(p-1)}{3p-2}}}\Bigg). \quad (21)$$

The leading stochastic rate matches AdamW's. We also prove a matching lower bound for normalized nuclear-norm stationarity against arbitrary randomized stochastic first-order methods under the same oracle (Theorem 13). Together with the upper bound in Theorem 11 and the AdamW result in Theorem 6, this yields the following minimax comparison.

Corollary 9 (Common-oracle minimax complexity, full details in Appendix F.4)**.** *Under the joint conditions in Appendix F.4 with $B = 1$, AdamW and Muon both attain minimax complexity $\Theta(\Delta L\sigma^{p/(p-1)}\varepsilon^{-(3p-2)/(p-1)})$ in their normalized natural stationarity criteria.*

Prior explanations (Yu et al., 2026d; Choudhury et al., 2026; Hübler et al., 2026) on why Muon outperforms AdamW in practice appeal to Muon's *optimal T-dependence*, whereas Table 3 shows that both methods attain the same *optimal T-dependence (equivalently, ε-complexity)* under the same oracle. This directly refutes their *convergence-rate-based explanation*. However, dimension dependent factors may still favor either method after conversion to a common Frobenius target (Remark 6).

Remark 6. For a common Frobenius target, *the norm conversion favors Muon on diagonal gradients and AdamW on flat rank-one gradients.* Appendix F.7 gives the exact comparison and uniformity conditions. The same rates extend to collections of Muon matrices under the product Frobenius oracle (Appendix F.6). Appendix F.5 also compares the assumptions in prior analyses.

6 Related Work

Due to space constraints, we summarize the most closely related work here and defer the full discussion to Appendix A. Existing analyses typically analyze Adam under finite variance noise (Zaheer et al., 2018; Zhang et al., 2022; Défossez et al., 2022; Li et al., 2023; Wang et al., 2023), while Li et al. (2025a) obtain an $O(\sqrt{d}\,T^{-1/4})$ rate for AdamW in the ℓ_1 gradient norm. The blockwise variant Adam-mini (Zhang et al., 2025b) is also studied under finite variance (Xie et al., 2025). The challenge is to extend these guarantees to heavy-tailed noise. Although methods based on clipping and normalization achieve sharp rates in this setting (Zhang et al., 2020; Cutkosky & Mehta, 2021; Liu & Zhou, 2025), existing analyses of Adam (Chezhegov et al., 2025; Pang, 2026; Liu, 2026b) do not establish the optimal rate for the standard, unclipped recursion, leaving the AdamW open problem (Yu et al., 2026c) unresolved. By contrast, Muon admits sharp heavy-tailed rates (Yu et al., 2026d; Choudhury et al., 2026; Hübler et al., 2026), motivating a direct comparison between the two optimizers. We close this gap for Adam and its variants and establish matching lower bounds.

7 Conclusion

We establish sharp heavy-tailed convergence guarantees for Adam through a more general analysis of Adam-mini. The same framework covers standard bias correction and admissible positive decoupled weight decay, and therefore includes AdamW as the practically relevant singleton-block endpoint. Matching lower bounds prove minimax optimality and show that the $\sqrt{d}$ dependence in stochastic ℓ_1 stationarity is unavoidable. These results also let us revisit the proposed heavy-tail explanation for Muon's empirical advantage. By comparing the AdamW endpoint with Muon under the same oracle, we show that the two methods attain the same minimax rate. Heavy-tailed noise alone therefore cannot explain Muon's advantage, refuting the convergence-rate explanation suggested by Yu et al. (2026d); Choudhury et al. (2026); Hübler et al. (2026).

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and the preparation of auxiliary scripts for processing experimental logs and producing figures. All experimental results are based on recorded outputs from actual training runs, and the reported plots are generated from these numerical records deterministically. AI tools were also used to improve the presentation and organization of some mathematical arguments, while all theorem statements, proofs, references, and quantitative claims were independently checked by the authors. The authors take full responsibility for the correctness and final presentation of the manuscript and all accompanying artifacts.

ETHICS STATEMENT

This work focuses on theoretical and empirical questions in stochastic optimization. It does not involve human participants, private or sensitive data, or the development of application-specific systems with direct societal impact. We are not aware of ethical concerns beyond the standard requirements of responsible and reproducible research.

REPRODUCIBILITY STATEMENT

The authors try their best to make this work reproducible. All assumptions and theorem statements are given explicitly in the main text, with complete proofs and technical details deferred to the appendix. The appendix also documents the experimental setup, model checkpoints, data processing, and hyperparameter choices used in our empirical study. Code for reproducing the reported experiments is included in the supplementary material.

REFERENCES

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 94909–94933, 2024.

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations (ICLR)*, pp. 16193–16205, 2024a.

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 619–640, 2024b.

Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free SGD is also effective for nonconvex optimization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 772–795, 2025.

Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. A new regret analysis for Adam-type algorithms. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 202–210, 2020.

Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory efficient adaptive optimization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 9746–9755, 2019.

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

Søren Asmussen and Jaakko Lehtomaa. Distinguishing log-concavity from heavy tails. *Risks*, 5(1): 10, 2017.

Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1):244–274, 2021.

Barak Battash, Lior Wolf, and Ofir Lindenbaum. Revisiting the noise model of stochastic gradient descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4780–4788, 2024.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SignSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 560–569, 2018.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 2397–2430, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 1877–1901, 2020.

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

Hongxu Chen, Ke Wei, Xiaoming Yuan, and Luo Luo. Stability and generalization of nonconvex optimization with heavy-tailed noise. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026a.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research (TMLR)*, 2026b. ISSN 2835-8856.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 49205–49233, 2023.

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

Savelii Chezhegov, Klyukin Yaroslav, Andrei Semenov, Aleksandr Beznosikov, Alexander Gasnikov, Samuel Horváth, Martin Takáč, and Eduard Gorbunov. Clipping improves Adam-Norm and AdaGrad-Norm when the noise is heavy-tailed. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 10269–10333, 2025.

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.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M. Gower. An exploration of non-euclidean gradient descent: Muon and its many variants. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

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

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 6643–6670, 2023.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024a.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024b.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad. *Transactions on Machine Learning Research (TMLR)*, 2022.

Kuangyu Ding, Nachuan Xiao, and Kim-Chuan Toh. Adam-family methods with decoupled weight decay in deep learning. *Transactions on Machine Learning Research (TMLR)*, 2025.

Paul Embrechts and Hanspeter Schmidli. Modelling of extremal events in insurance and finance. *Zeitschrift für Operations Research*, 39(1):1–34, 1994.

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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 3610–3619, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 3964–3975, 2021.

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

Feihu Huang, Junyi Li, and Heng Huang. SUPER-ADAM: Faster and universal framework of adaptive gradients. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 9074–9085, 2021.

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.

Ruichen Jiang, Devyani Maladkar, and Aryan Mokhtari. Provable complexity improvement of AdaGrad over SGD: Upper and lower bounds in stochastic non-convex optimization. In *Proceedings of the 38th Conference on Learning Theory (COLT)*, pp. 3124–3158, 2025a.

Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved analysis for sign-based methods with momentum updates. *arXiv preprint arXiv:2507.12091*, 2025b.

Wei Jiang, Mao Xu, Wenhao Yang, Yibo Wang, Zechao Li, and Lijun Zhang. Convergence analysis of the Lion optimizer in centralized and distributed settings. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

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

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. URL `https://kellerjordan.github.io/posts/muon/`.

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

Jeonseong Kim. Second-moment memory in coordinatewise Adam. *arXiv preprint arXiv:2608.15824*, 2026.

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

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 17343–17363, 2023.

Lexiao Lai, Tianyi Lin, and Jiayu Zhang. Nonsmooth optimization via orthogonalized momentum. *arXiv preprint arXiv:2609.13677*, 2026.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Jaakko Lehtomaa. Limiting behaviour of constrained sums of two variables and the principle of a single big jump. *Statistics & Probability Letters*, 107:157–163, 2015.

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

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

Huan Li, Yiming Dong, and Zhouchen Lin. On the $O(\sqrt{d}/T^{1/4})$ convergence rate of RMSProp and its momentum extension measured by ℓ_1 norm. *Journal of Machine Learning Research (JMLR)*, 26(131):1–25, 2025b.

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon and further. *arXiv preprint arXiv:2502.02900*, 2025.

Jichu Li, Xuan Tang, and Difan Zou. The implicit bias of steepest descent with mini-batch stochastic gradient. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 32122–32138, 2024.

Zhuanghua Liu, Menglian Wang, and Luo Luo. Musec: Momentum spectral clipping for stable Muon-type training. *arXiv preprint arXiv:2609.11655*, 2026.

Zijian Liu. Can adaptive gradient methods converge under heavy-tailed noise? a case study of AdaGrad. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026a.

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

Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory (ALT)*, pp. 1–47, 2026c.

Zijian Liu. Clipped gradient methods for nonsmooth convex optimization under heavy-tailed noise: A refined analysis. In *International Conference on Learning Representations (ICLR)*, pp. 116560–116608, 2026d.

Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations (ICLR)*, pp. 92529–92554, 2025.

Zijian Liu, Jiawei Zhang, and Zhengyuan Zhou. Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *Proceedings of 36th Conference on Learning Theory (COLT)*, pp. 2266–2290, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations (ICLR)*, 2019.

Vitali D. Milman and Gideon Schechtman. *Asymptotic Theory of Finite-Dimensional Normed Spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1986.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, et al. MiniMax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

Quan M. Nguyen. How to set β_1, β_2 in Adam: An online learning perspective. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory (ALT)*, pp. 1–16, 2026.

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

Rui Pan, Yang Luo, Yuxing Liu, Yang You, and Tong Zhang. Unbiased gradient low-rank projection. *arXiv preprint arXiv:2510.17802*, 2025.

Yijiang Pang. The convergence behavior of Adam under heavy-tailed noise. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 5244–5265, 2026.

Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M Gower. Muon does not converge on convex lipschitz functions. *arXiv preprint arXiv:2605.08980*, 2026.

Iosif Pinelis. Best possible bounds of the von Bahr–Esseen type. *Annals of Functional Analysis*, 6 (4):1–29, 2015.

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtarik. Muon is provably faster with momentum variance reduction. *arXiv preprint arXiv:2512.16598*, 2025.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

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

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. From Muon to Gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

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

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Convergence bound and critical batch size of Muon optimizer. *arXiv preprint arXiv:2507.01598*, 2025.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and Muons: Optimization via stochastic frank-wolfe under heavy-tailed noise. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=4nH4CulGaP.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations (ICLR)*, 2021.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 5167–5175, 2026.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5827–5837, 2019.

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

Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, et al. Hunyuan-Large: An open-source MoE model with 52 billion activated parameters by Tencent. *arXiv preprint arXiv:2411.02265*, 2024.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

Xuan Tang, Han Zhang, Yuan Cao, and Difan Zou. Understanding the generalization of stochastic gradient adam in learning neural networks. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 18120–18187, 2025.

Xuan Tang, Jichu Li, and Difan Zou. A convergence analysis of adaptive optimizers under floating-point quantization. In *International Conference on Learning Representations (ICLR)*, pp. 4069–4127, 2026.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. ADOPT: Modified Adam can converge with any β_2 with the optimal rate. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 72438–72474, 2024.

Hongyi Tao, Dingzhi Yu, and Lijun Zhang. When and Why SignSGD Outperforms SGD: A Theoretical Study Based on ℓ_1-norm Lower Bounds. *arXiv preprint arXiv:2605.06615*, 2026.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026a.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*, 2026b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Sharan Vaswani, Yifan Sun, and Reza Babanezhad Harikandeh. Convergence of steepest descent and Adam under non-uniform smoothness. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the r-th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299 – 303, 1965.

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

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

Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of Adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2960–2969, 2024b.

Guanghui Wang, Shiyin Lu, Quan Cheng, Wei-Wei Tu, and Lijun Zhang. SAdam: A variant of Adam for strongly convex functions. In *International Conference on Learning Representations (ICLR)*, 2020.

Jiaxuan Wang and Jenna Wiens. AdaSGD: Bridging the gap between SGD and Adam. *arXiv preprint arXiv:2006.16541*, 2020.

Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Weinan E, and Lei Wu. The sharpness disparity principle in transformers for accelerating language model pre-training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 64859–64879, 2025.

James G. Wendel. Note on the gamma function. *The American Mathematical Monthly*, 55(9): 563–564, 1948.

Nachuan Xiao, Xiaoyin Hu, Xin Liu, and Kim-Chuan Toh. Adam-family methods for nonsmooth optimization with convergence guarantees. *Journal of Machine Learning Research (JMLR)*, 25 (48):1–53, 2024.

Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 54488–54510, 2024.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *International Conference on Learning Representations (ICLR)*, pp. 7937–7965, 2025.

Yan-Feng Xie, Yu-Jie Zhang, Peng Zhao, and Zhi-Hua Zhou. Dynamic regret via discounted-to-dynamic reduction with applications to curved losses and Adam optimizer. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, pp. to appear, 2026.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations (ICLR)*, 2020.

Adams Wei Yu, Lei Huang, Qihang Lin, Ruslan Salakhutdinov, and Jaime Carbonell. Block-normalized gradient method: An empirical study for training deep neural network. *arXiv preprint arXiv:1707.04822*, 2017.

Dingzhi Yu, Yuxing Liu, Rui Pan, Lijun Zhang, Yang You, and Difan Zou. Finite-time nonsmooth convergence of Signum and Muon, 2026a. URL `https://www.lamda.nju.edu.cn/yudz/pdf/Finite-Time_Nonsmooth_Convergence_of_Signum_and_Muon.pdf`.

Dingzhi Yu, Rui Pan, Yuxing Liu, Difan Zou, and Tong Zhang. StoSignSGD: Unbiased Structural Stochasticity Fixes SignSGD for Training Large Language Models. *arXiv preprint arXiv:2604.15416*, 2026b.

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

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*, 2026d.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 9793–9803, 2018.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: An open bilingual pre-trained model. In *International Conference on Learning Representations (ICLR)*, 2023.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. GLM-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of Adam on separable data. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 23988–24021, 2024a.

Jiayu Zhang and Tianyi Lin. Scale-invariant neural network optimization: Norm geometry and heavy-tailed noise. *arXiv preprint arXiv:2605.18528*, 2026.

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

Qi Zhang, Yi Zhou, and Shaofeng Zou. Convergence guarantees for RMSProp and Adam in generalized-smooth non-convex optimization with affine noise variance. *Transactions on Machine Learning Research (TMLR)*, 2025a. ISSN 2835-8856.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 58780–58799, 2024.

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 35 (NeurIPS)*, pp. 28386–28399, 2022.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why Transformers need Adam: A Hessian perspective. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 131786–131823, 2024b.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik (Durk) Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations (ICLR)*, pp. 28033–28063, 2025b.

Yushun Zhang, Bingran Li, Congliang Chen, Zhi-Quan Luo, and Ruoyu Sun. Adam converges without any modification on update rules. *arXiv preprint arXiv:2603.02092*, 2026.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *International Conference on Learning Representations (ICLR)*, pp. 2830–2850, 2025.

Shuai Zheng and James T Kwok. Blockwise adaptivity: Faster training and better generalization in deep learning. *arXiv preprint arXiv:1905.09899*, 2019.

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 (TMLR)*, 2024a.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, and Weinan E. Towards theoretically understanding why SGD generalizes better than Adam in deep learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 21285–21296, 2020.

Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of AdamW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6486–6493, 2024b.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations (ICLR)*, 2019.

Zhenxun Zhuang, Mingrui Liu, Ashok Cutkosky, and Francesco Orabona. Understanding AdamW through proximal methods and scale-freeness. *Transactions on Machine Learning Research (TMLR)*, 2022.

Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of Adam in learning neural networks with proper regularization. In *International Conference on Learning Representations (ICLR)*, 2023.

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

A Extended Discussion of Related Work

Convergence theory for Adam. The original convergence analysis of Adam (Kingma & Ba, 2015) was challenged by the counterexamples of Reddi et al. (2018), motivating convergent variants and a broad theory of Adam-type methods (Zaheer et al., 2018; Chen et al., 2019; Zhou et al., 2019; Luo et al., 2019; Zou et al., 2019; Wang et al., 2020; Alacaoglu et al., 2020; Shi et al., 2021; Huang et al., 2021; Barakat & Bianchi, 2021; Taniguchi et al., 2024; Wang et al., 2024a; Zhou et al., 2024a; Yu et al., 2026b). For vanilla Adam, convergence under standard smoothness and finite- or affine-variance noise has been established under increasingly realistic assumptions, with several works obtaining near-optimal or SGD-comparable finite-time and last-iterate guarantees (Défossez et al., 2022; Zhang et al., 2022; Wang et al., 2023; Hong & Lin, 2024; Jin et al., 2025; Zhang et al., 2026). Convergence under generalized or non-uniform smoothness has also been studied (Li et al., 2023; Wang et al., 2024b; Zhang et al., 2025a; Vaswani et al., 2026). For AdamW (Loshchilov & Hutter, 2019), theoretical work has studied its proximal and scale-free interpretation (Zhuang et al., 2022), convergence and generalization under decoupled weight decay (Zhou et al., 2024b; Ding et al., 2025), and its implicit bias toward ℓ_∞-constrained optimization (Xie & Li, 2024). More recently, Li et al. (2025a) establish an $O(\sqrt{d}T^{-1/4})$ rate measured by the ℓ_1 gradient norm. Beyond stationarity guarantees, complementary work studies the generalization and implicit bias of Adam and AdamW, including the effects of weight decay, ℓ_∞ geometry, and mini-batch stochasticity (Zou et al., 2023; Zhang et al., 2024a; Tang et al., 2025).

For nonsmooth objectives, online-to-nonconvex conversion translates online regret guarantees into generalized stationarity guarantees (Cutkosky et al., 2023; Zhang & Cutkosky, 2024; Ahn et al., 2025; Yu et al., 2026b). Adam admits an FTRL interpretation in this framework (Ahn et al., 2024b), and Ahn & Cutkosky (2024) use related ideas to obtain optimal smooth and nonsmooth rates for clipped Adam with model EMA. Subsequent work applies online-learning techniques to Adam's momentum parameters, dynamic regret, and clip-free Adam (Nguyen, 2026; Xie et al., 2026; Liu, 2026b). A complementary nonsmooth analysis based on two-timescale stochastic approximation is given by Xiao et al. (2024).

Heavy-tailed gradient noise. Heavy-tailed stochastic gradients have been observed and modeled in deep learning and stochastic optimization (Simsekli et al., 2019; Zhang et al., 2020; Gurbuzbalaban et al., 2021; Garg et al., 2021; Ahn et al., 2024a; Yu et al., 2026d). Under a finite pth noise moment, methods based on clipping, normalization, and momentum have been developed under increasingly weak assumptions (Cutkosky & Mehta, 2021; Liu & Zhou, 2023; Liu et al., 2023; Koloskova et al., 2023; Nguyen et al., 2023; Liu et al., 2024; Liu & Zhou, 2025; Sun et al., 2025; Liu, 2026d), with recent work also studying stability and generalization under heavy-tailed noise (Chen et al., 2026a). Heavy-tailed adaptive methods have only recently received direct analyses, including clipped Adam-Norm and AdaGrad-Norm (Chezhegov et al., 2025), unclipped AdaGrad (Liu, 2026a;c), and plain/clipped Adam (Pang, 2026; Liu, 2026b). These results leave open whether the standard Adam recursion can attain the optimal heavy-tailed rate; Yu et al. (2026c) formulated this question explicitly for AdamW. Sign-based optimizers such as SignSGD and Lion (Bernstein et al., 2018; Chen et al., 2023) also have heavy-tail guarantees (Yu et al., 2026d).

Blockwise Adam. Coarse group-structured scaling predates modern large-model optimization, appearing in block-normalized gradients, layer-wise scaling, and shared adaptive statistics (Yu et al., 2017; You et al., 2017; 2020; Zheng & Kwok, 2019; Anil et al., 2019). Related ideas have recently re-emerged in Transformer optimization (Zhao et al., 2025; Wang et al., 2025). Zhang et al. (2024b) show that Transformer Hessians become approximately block diagonal with substantial curvature heterogeneity across blocks. Motivated by this structure, Adam-mini (Zhang et al., 2025b) uses a Hessian-informed partition and shares one second-moment scale within each block, substantially reducing optimizer-state memory. Xie et al. (2025) provide convergence theory for essentially the same block-mean second-moment recursion under block-adapted smoothness and finite variance.

Convergence theory for Muon. Following the introduction of Muon (Jordan et al., 2024), a growing literature has developed convergence analyses, non-Euclidean interpretations, and geometric characterizations of its matrix-aware updates (Li & Hong, 2025; Sato et al., 2025; Chang et al., 2025; Qian et al., 2025; Pan et al., 2025; Riabinin et al., 2026; Shen et al., 2026; Chen et al., 2026b; Crawshaw et al., 2026; Tang et al., 2026; Li et al., 2026; Yu et al., 2026d; Tao et al., 2026). Several

works further bridge the gap between idealized polar updates and practical approximate orthogonalization, including finite-step Newton–Schulz and more general inexact updates (Shulgin et al., 2026; Kim & Oh, 2026). Recent work also studies Muon beyond smooth objectives, identifying nonsmooth failure modes and conditions or modifications that restore convergence (Yu et al., 2026a; Parshakova et al., 2026; Lai et al., 2026; Liu et al., 2026). Heavy-tailed stochasticity has received particular attention through stochastic Frank–Wolfe, matrix sign, inexact-polar, and non-Euclidean analyses (Sfyraki & Wang, 2026; Yu et al., 2026d; Choudhury et al., 2026; Hübler et al., 2026).

B Supporting Algorithms

We give the complete specification of Algorithm 2, the reductions in Section 3.3, and a Newton–Schulz implementation of the Muon recursion in Section 5.

B.1 Adam-mini

Algorithm 2 is the algorithm analyzed in Theorems 3 and 4.

Algorithm 2 Adam-mini with bias correction and decoupled weight decay (Zhang et al., 2025b)

Require: Partition $\mathcal{P}$ with block dimensions $\{d_j\}_{j=1}^J$, $x_1 \in \mathbb{R}^d$, $T, B \in \mathbb{N}_+$, $\{\eta_t\}_{t=1}^T$, $\beta_1, \beta_2 \in [0,1)$, $\epsilon_{\mathrm{bc}} > 0$, $\lambda \geq 0$

Initialize $m_0 = 0$ and $v_0^{(j)} = 0$ for every $j \in [J]$
for $t = 1, \dots, T$ **do**
  Compute $g_t = B^{-1} \sum_{b=1}^B g_t^b$ at x_t
  for $j = 1, \dots, J$ **do**
    $m_t^{(j)} = \beta_1 m_{t-1}^{(j)} + (1-\beta_1) g_t^{(j)}$
    $v_t^{(j)} = \beta_2 v_{t-1}^{(j)} + (1-\beta_2) \|g_t^{(j)}\|^2 / d_j$
    $\widehat{m}_t^{(j)} = m_t^{(j)} / (1-\beta_1^t)$ and $\widehat{v}_t^{(j)} = v_t^{(j)} / (1-\beta_2^t)$
    $x_{t+1}^{(j)} = (1-\eta_t\lambda) x_t^{(j)} - \eta_t \widehat{m}_t^{(j)} / (\sqrt{\widehat{v}_t^{(j)}} + \epsilon_{\mathrm{bc}})$
  end for
end for
Return x_{T+1}

B.2 AdamW

The singleton partition $\mathcal{P} = \{\{1\}, \dots, \{d\}\}$ in Algorithm 2 gives Algorithm 3, with $\epsilon_{\mathrm{alg}} = \epsilon_{\mathrm{bc}}$. Vector products, square roots, and divisions in this algorithm are coordinatewise.

Algorithm 3 AdamW (Loshchilov & Hutter, 2019)

Require: $x_1 \in \mathbb{R}^d$, $T, B \in \mathbb{N}_+$, $\{\eta_t\}_{t=1}^T$, $\beta_1, \beta_2 \in [0,1)$, $\epsilon_{\mathrm{alg}} > 0$, $\lambda \geq 0$

Initialize $m_0 = 0$ and $v_0 = 0$
for $t = 1, \dots, T$ **do**
  Compute $g_t = B^{-1} \sum_{b=1}^B g_t^b$ at x_t
  $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_t \odot g_t)$
  $\widehat{m}_t = m_t / (1-\beta_1^t)$ and $\widehat{v}_t = v_t / (1-\beta_2^t)$
  $x_{t+1} = (1-\eta_t\lambda) x_t - \eta_t \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon_{\mathrm{alg}})$
end for
Return x_{T+1}

B.3 AdaSGD and Adam-Norm

The one-block partition $\mathcal{P} = \{[d]\}$ in Algorithm 2 gives Algorithm 4, with $\epsilon_{\mathrm{alg}} = \epsilon_{\mathrm{bc}}$. It uses the global second moment of AdaSGD (Wang & Wiens, 2020). Setting $\lambda = 0$ removes weight decay.

Algorithm 4 AdaSGD as the one-block specialization of Algorithm 2

Require: $x_1 \in \mathbb{R}^d, T, B \in \mathbb{N}_+, \{\eta_t\}_{t=1}^T, \beta_1, \beta_2 \in [0,1), \epsilon_{\text{alg}} > 0, \lambda \geq 0$
1: Initialize $m_0 = 0 \in \mathbb{R}^d$ and $v_0 = 0 \in \mathbb{R}$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $g_t = B^{-1} \sum_{b=1}^B g_t^b$ at x_t
4: $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
5: $v_t = \beta_2 v_{t-1} + (1-\beta_2)\|g_t\|_2^2/d$
6: $\widehat{m}_t = m_t/(1-\beta_1^t)$ and $\widehat{v}_t = v_t/(1-\beta_2^t)$
7: $x_{t+1} = (1-\eta_t\lambda)x_t - \eta_t \widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon_{\text{alg}})$
8: **end for**
9: Return x_{T+1}

The original AdaSGD notation uses $u_t = m_t/(1-\beta_1)$ and $s_t = dv_t$. Its unnormalized momentum update is recovered by taking $\eta_t = \alpha(1-\beta_1^t)/(1-\beta_1)$ and $\lambda = \epsilon_{\text{alg}} = 0$. For Adam-Norm (Wang et al., 2024a), set $\nu_t = dv_t$, $\eta_t^{\text{Norm}} = \sqrt{d}\,\eta_t$, $\epsilon^{\text{Norm}} = \sqrt{d}\,\epsilon_{\text{alg}}$, and $\lambda^{\text{Norm}} = \lambda/\sqrt{d}$. These substitutions preserve the update. We choose the more popular name Adam-Norm following Chezhegov et al. (2025).

B.4 MUON

Algorithm 5 implements the momentum recursion in Eq. (20) with a numerical matrix-sign step. Algorithm 6 gives the quintic Newton–Schulz iteration used by Jordan et al. (2024). The exact-polar analysis replaces $\text{NS}_K(M_t, \epsilon_{\text{NS}})$ by $\text{msign}\,(M_t)$.

Algorithm 5 Muon

Require: $X_1 \in \mathbb{R}^{m\times n}, T, B, K \in \mathbb{N}_+, \{\eta_t\}_{t=1}^T, \beta \in [0,1), \epsilon_{\text{NS}} > 0$
1: Initialize $M_0 = 0$
2: **for** $t = 1, \ldots, T$ **do**
3: Compute $G_t = B^{-1} \sum_{b=1}^B G_t^b$ at X_t
4: $M_t = \beta M_{t-1} + (1-\beta) G_t$
5: $O_t = \text{NS}_K(M_t, \epsilon_{\text{NS}})$ using Algorithm 6
6: $X_{t+1} = X_t - \eta_t O_t$
7: **end for**
8: Return X_{T+1}

Algorithm 6 Newton–Schulz matrix orthogonalization

Require: $M \in \mathbb{R}^{m\times n}$, iteration count K, $\epsilon_{\text{NS}} > 0$
1: Set $a = 3.4445$, $b = -4.7750$, $c = 2.0315$
2: $Z = M/(\|M\|_{\text{F}} + \epsilon_{\text{NS}})$
3: **if** $m > n$ **then**
4: $Z = Z^\top$
5: **end if**
6: **for** $k = 1, \ldots, K$ **do**
7: $A = ZZ^\top$
8: $Z = aZ + (bA + cA^2)Z$
9: **end for**
10: **if** $m > n$ **then**
11: $Z = Z^\top$
12: **end if**
13: Return Z

C Proofs of the Upper Bounds in Section 3.1

We prove Theorems 1 and 2 for Algorithm 1 with $\eta_t \equiv \eta$ and $\lambda = 0$. Eq. (3) and $d_{\min} \le d$ imply $1 - \beta_2 \le 1/8 < 1$. Let

$$\nabla_t := \nabla F(x_t), \qquad \xi_t := g_t - \nabla_t, \qquad D_t^{(j)} := \sqrt{v_t^{(j)}} + \epsilon_{\text{alg}}.$$

The theorem conditions give

$$\frac{L\eta}{1-\beta_1} = \frac{\sigma_B}{\sqrt{d}} \frac{(1-\beta_2)^{3/2-1/p}}{1-\beta_1} \le \frac{\sigma_B}{\sqrt{d}} (1-\beta_2)^{1/2-1/p}, \tag{22}$$

because $1 - \beta_1 \ge 1 - \beta_2$. Thus the statistical floor in Eq. (2) satisfies every optimization-stability inequality involving $L\eta/(1-\beta_1)$ used below.

We work with the mini-batch gradient $g_t = B^{-1} \sum_{b=1}^{B} g_t^b$ through its effective noise level. By Lemma C.1,

$$\mathbb{E}_{t-1} \left[\|g_t - \nabla F(x_t)\|^p \right] \le 2\sigma^p B^{1-p} =: \sigma_B^p, \qquad \sigma_B := 2^{1/p} \sigma B^{-(p-1)/p}. \tag{23}$$

Thus the proof can treat g_t as a *single stochastic gradient* with noise level σ_B.

C.1 Mini-Batch Moment Bound

Lemma C.1 (Hilbert-space von Bahr–Esseen bound)**.** *Let $\mathcal{H}$ be an outer-history σ-field, and let $X_1, \dots, X_B$ be random vectors in a Hilbert space with finite pth moments for $p \in (1, 2]$. Suppose they are conditionally independent given $\mathcal{H}$ and satisfy $\mathbb{E}[X_b \mid \mathcal{H}] = 0$ for every b. Then*

$$\mathbb{E} \left\| \sum_{b=1}^{B} X_b \right\|^p \le 2 \sum_{b=1}^{B} \mathbb{E} \|X_b\|^p . \tag{24}$$

The same bound holds conditionally on $\mathcal{H}$. Applying it with $\mathcal{H} = \mathcal{F}_{t-1}$ and $X_b = g_t^b - \nabla F(x_t)$ gives Eq. (23) *under Assumption 3.*

Proof. We first prove the deterministic inequality

$$\|u + v\|^p \le \|u\|^p + p \|u\|^{p-2} \langle u, v \rangle + 2 \|v\|^p , \tag{25}$$

where the middle term is defined as zero when $u = 0$. For $p = 2$, this follows by expanding the squared norm. Suppose $1 < p < 2$ and $u, v \ne 0$. Divide by $\|v\|^p$, and put

$$a := \frac{\|u\|}{\|v\|}, \qquad c := \frac{\langle u, v \rangle}{\|u\| \, \|v\|} \in [-1, 1].$$

The remainder after the first two terms on the right of Eq. (25) becomes

$$R_a(c) := (a^2 + 1 + 2ac)^{p/2} - a^p - pa^{p-1}c.$$

For fixed a, this is concave in c. At $c = 1$, concavity of $t \mapsto t^{p-1}$ gives $R_a(1) \le 1$. The derivative of $R_a(1)$ with respect to a is nonpositive, and its value at $a = 0$ is one. If $c = -1$ and $0 \le a \le 1$, then

$$R_a(-1) = (1-a)^p + a^{p-1}(p-a) \le 2,$$

because the second term is maximized at $a = p - 1$ and is at most one. If $c = -1$ and $a \ge 1$, convexity of $t \mapsto t^p$ gives

$$R_a(-1) \le p \left\{ a^{p-1} - (a-1)^{p-1} \right\} \le p \le 2.$$

The only possible interior maximizer is $c = -1/(2a)$, which requires $a \ge 1/2$. At that point, $R_a(c) = (p/2)a^{p-2} \le p2^{1-p} \le 2$. This proves Eq. (25). The cases $u = 0$ or $v = 0$ are immediate.

Now condition on the outer history and reveal $X_1, \dots, X_B$ successively. Conditional independence and centering give $\mathbb{E}[X_b \mid X_1, \dots, X_{b-1}] = 0$. Apply Eq. (25) with $u = \sum_{a<b} X_a$ and $v = X_b$, and take the conditional expectation. The inner product term vanishes, so iteration yields

$$\mathbb{E} \left\| \sum_{b=1}^{B} X_b \right\|^p \le 2 \sum_{b=1}^{B} \mathbb{E} \|X_b\|^p .$$

This proves Eq. (24). The scalar form is the classical von Bahr–Esseen inequality (von Bahr & Esseen, 1965). Sharper constants for inequalities of this type are studied by Pinelis (2015). Taking $X_b = g_t^b - \nabla_t$ and using $g_t - \nabla_t = B^{-1}\sum_{b=1}^{B} X_b$ yields

$$\mathbb{E}_{t-1}\left\|g_t - \nabla_t\right\|^p \le 2B^{-p}\sum_{b=1}^{B}\mathbb{E}_{t-1}\left\|g_t^b - \nabla_t\right\|^p \le 2\sigma^p B^{1-p}.$$

□

Lemma C.1 is the only step that aggregates the B individual oracle samples. It allows the rest of the upper bound proof to treat the mini-batch average as a single stochastic gradient whose conditional pth noise moment is bounded by σ_B^p.

C.2 Bounds on the Updates

Lemma C.2 (Bounded block update)**.** *If $\beta_2^2 \le \beta_1 \le \beta_2$, then for every t and block j,*

$$\left\|m_t^{(j)}\right\|^2 \le 4d_j v_t^{(j)}, \qquad \left\|\frac{m_t^{(j)}}{D_t^{(j)}}\right\| \le 2\sqrt{d_j}. \tag{26}$$

If u_t denotes the concatenated adaptive direction, then $\|u_t\| \le 2\sqrt{d}$. Consequently, for every $1 \le s < t \le T+1$,

$$\|x_t - x_s\| \le 2\eta\sqrt{d}\,(t-s). \tag{27}$$

Proof. If $\beta_2 = 0$, the memory window forces $\beta_1 = 0$. Then $m_t^{(j)} = g_t^{(j)}$ and $d_j v_t^{(j)} = \left\|g_t^{(j)}\right\|^2$, so the claim follows directly. We henceforth assume $\beta_2 > 0$. Unrolling the two recursions in Algorithm 1 gives

$$m_t^{(j)} = (1-\beta_1)\sum_{s=1}^{t}\beta_1^{t-s}g_s^{(j)}, \quad d_j v_t^{(j)} = (1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|g_s^{(j)}\right\|^2. \tag{28}$$

Insert the factors $\beta_2^{(t-s)/2}$ and $\beta_2^{-(t-s)/2}$ into Eq. (28). Weighted Cauchy–Schwarz then yields

$$\begin{aligned}\left\|m_t^{(j)}\right\|^2 &= (1-\beta_1)^2\left\|\sum_{s=1}^{t}\frac{\beta_1^{t-s}}{\beta_2^{(t-s)/2}}\left(\beta_2^{(t-s)/2}g_s^{(j)}\right)\right\|^2 \\ &\le (1-\beta_1)^2\left(\sum_{s=1}^{t}\left(\frac{\beta_1^2}{\beta_2}\right)^{t-s}\right)\left(\sum_{s=1}^{t}\beta_2^{t-s}\left\|g_s^{(j)}\right\|^2\right) \le \frac{(1-\beta_1)^2}{(1-\beta_2)(1-\beta_1^2/\beta_2)}d_j v_t^{(j)}.\end{aligned}$$

The memory window in Eq. (2) gives

$$1-\beta_1 \le 1-\beta_2^2 = (1-\beta_2)(1+\beta_2) \le 2(1-\beta_2), \quad \text{and } 1-\frac{\beta_1^2}{\beta_2} \ge 1-\beta_2.$$

Thus the fraction above is at most 4, which proves the first inequality. Moreover, $D_t^{(j)} = \sqrt{v_t^{(j)}} + \epsilon_{\text{alg}} \ge \sqrt{v_t^{(j)}}$, so

$$\left\|\frac{m_t^{(j)}}{D_t^{(j)}}\right\|^2 \le \frac{\left\|m_t^{(j)}\right\|^2}{v_t^{(j)}} \le 4d_j,$$

where the claim is immediate when $v_t^{(j)} = 0$ because then Eq. (28) also gives $m_t^{(j)} = 0$. Taking square roots proves the block bound. Finally, if $u_t^{(j)} := m_t^{(j)}/D_t^{(j)}$, then

$$\|u_t\|^2 = \sum_{j=1}^{J}\left\|u_t^{(j)}\right\|^2 \le 4\sum_{j=1}^{J}d_j = 4d.$$

□

C.3 Core and Tail Decomposition

The following split is used only in the proof. The algorithm continues to use the original g_t. For an event A, $\mathbb{I}_A$ denotes its indicator and A^c denotes its complement. Define

$$Z_t := \|\xi_t\|, \quad H_t := \|\nabla_t\|, \quad M := \sigma_B(1-\beta_2)^{-1/p}, \quad I_t := \{Z_t > M,\ Z_t > 2H_t\}. \tag{29}$$

Thus I_t is the event on which the noise is both above the statistical threshold M and larger than twice the current gradient. Define the tail and core parts of the stochastic gradient by

$$g_{t,\text{tail}} := g_t \mathbb{I}_{I_t}, \qquad g_{t,\text{core}} := g_t \mathbb{I}_{I_t^c}, \qquad g_t = g_{t,\text{core}} + g_{t,\text{tail}}. \tag{30}$$

For $a \in \{\text{core}, \text{tail}\}$, define the first- and second-moment buffers generated by the sequence $\{g_{s,a}\}_{s\le t}$ as

$$m_{t,a}^{(j)} := (1-\beta_1)\sum_{s=1}^{t} \beta_1^{t-s} g_{s,a}^{(j)}, \quad v_{t,a}^{(j)} := (1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s} \frac{\left\|g_{s,a}^{(j)}\right\|^2}{d_j}. \tag{31}$$

Because I_s and I_s^c are disjoint, we have the pathwise identities

$$m_t^{(j)} = m_{t,\text{core}}^{(j)} + m_{t,\text{tail}}^{(j)}, \qquad v_t^{(j)} = v_{t,\text{core}}^{(j)} + v_{t,\text{tail}}^{(j)}. \tag{32}$$

Note that the decomposition above is only used in the proof. Algorithm 1 still uses the original g_t.

Lemma C.3 (Core and tail moments)**.** *Conditionally on $\mathcal{F}_{t-1}$,*

$$\mathbb{P}(I_t \mid \mathcal{F}_{t-1}) \le 1-\beta_2 \quad \textit{a.s.}, \tag{33}$$

$$\mathbb{E}_{t-1}\|g_{t,\text{tail}}\| \le \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p}, \tag{34}$$

$$\|\nabla_t - \mathbb{E}_{t-1} g_{t,\text{core}}\| \le \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p}, \tag{35}$$

$$\mathbb{E}_{t-1}\|g_{t,\text{core}} - \nabla_t\|^2 \le \sigma_B^2(1-\beta_2)^{1-2/p} + \frac{5}{2}\sigma_B(1-\beta_2)^{1-1/p}\|\nabla_t\|. \tag{36}$$

Proof. To lighten notation, write $Z := Z_t$ and $H := H_t$ in this proof. All expectations below are conditional on $\mathcal{F}_{t-1}$. Since $I_t \subseteq \{Z > M\}$, Markov's inequality and Eq. (23) give

$$\mathbb{P}(I_t \mid \mathcal{F}_{t-1}) \le \mathbb{P}(Z > M \mid \mathcal{F}_{t-1}) \le \frac{\mathbb{E}_{t-1} Z^p}{M^p} \le 1-\beta_2. \tag{37}$$

This proves Eq. (33).

On I_t, the condition $Z > 2H$ implies

$$\|g_t\| \le \|\nabla_t\| + \|\xi_t\| = H + Z \le \frac{3}{2}Z.$$

For every nonnegative random variable Z and $M > 0$,

$$Z\mathbb{I}_{\{Z>M\}} \le M^{1-p} Z^p. \tag{38}$$

Therefore,

$$\mathbb{E}_{t-1}\|g_{t,\text{tail}}\| \le \frac{3}{2}\mathbb{E}_{t-1}[Z\mathbb{I}_{\{Z>M\}}] \le \frac{3}{2}M^{1-p}\mathbb{E}_{t-1}Z^p \le \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p},$$

which proves Eq. (34). Since $g_t = g_{t,\text{core}} + g_{t,\text{tail}}$ and $\mathbb{E}_{t-1} g_t = \nabla_t$, so we have $\nabla_t - \mathbb{E}_{t-1} g_{t,\text{core}} = \mathbb{E}_{t-1} g_{t,\text{tail}}$. Jensen's inequality and Eq. (34) prove Eq. (35).

It remains to control the core second moment. From Eq. (30),

$$g_{t,\text{core}} - \nabla_t = (\nabla_t + \xi_t)\mathbb{I}_{I_t^c} - \nabla_t(\mathbb{I}_{I_t} + \mathbb{I}_{I_t^c}) = \xi_t \mathbb{I}_{I_t^c} - \nabla_t \mathbb{I}_{I_t}.$$

The two terms have disjoint supports, so the cross terms are zero. Thus,

$$\mathbb{E}_{t-1}\|g_{t,\text{core}} - \nabla_t\|^2 = \mathbb{E}_{t-1}\left\|\xi_t\mathbb{I}_{I_t^c} - \nabla_t\mathbb{I}_{I_t}\right\|^2 = \mathbb{E}_{t-1}[Z^2\mathbb{I}_{I_t^c}] + H^2\mathbb{P}(I_t \mid \mathcal{F}_{t-1}). \tag{39}$$

The first term $\mathbb{E}_{t-1}[Z^2\mathbb{I}_{I_t^c}]$ in Eq. (39). On I_t^c, either $Z \le M$ or $M < Z \le 2H$. For the first region,

$$\mathbb{E}_{t-1}[Z^2\mathbb{I}_{\{Z\le M\}}] \le M^{2-p}\mathbb{E}_{t-1}Z^p \le \sigma_B^2(1-\beta_2)^{1-2/p}. \tag{40}$$

For the second region, Markov's inequality gives

$$\mathbb{E}_{t-1}[Z^2\mathbb{I}_{\{M<Z\le 2H\}}] \le 4H^2\mathbb{P}(Z>M) \overset{(37)}{\le} (2H)^2(1-\beta_2), \tag{41}$$

while $Z \le 2H$ gives

$$\mathbb{E}_{t-1}[Z^2\mathbb{I}_{\{M<Z\le 2H\}}] \le (2H)^{2-p}\mathbb{E}_{t-1}Z^p \le (2H)^{2-p}\sigma_B^p. \tag{42}$$

Using $\min\{a,b\} \le a^{(p-1)/p}b^{1/p}$ in Eqs. (41) and (42), we obtain

$$\mathbb{E}_{t-1}[Z^2\mathbb{I}_{\{M<Z\le 2H\}}] \le \left((2H)^2(1-\beta_2)\right)^{\frac{p-1}{p}}\left((2H)^{2-p}\sigma_B^p\right)^{\frac{1}{p}} = 2H\sigma_B(1-\beta_2)^{1-1/p}. \tag{43}$$

Combining Eqs. (40) and (43) gives

$$\mathbb{E}_{t-1}[Z^2\mathbb{I}_{I_t^c}] \le \sigma_B^2(1-\beta_2)^{1-2/p} + 2H\sigma_B(1-\beta_2)^{1-1/p}. \tag{44}$$

The second term $H^2\mathbb{P}(I_t)$ in Eq. (39). On I_t we have $Z > 2H$, and hence

$$H^2\mathbb{P}(I_t \mid \mathcal{F}_{t-1}) \le \frac{H}{2}\mathbb{E}_{t-1}[Z\mathbb{I}_{I_t}] \overset{(38)}{\le} \frac{H}{2}M^{1-p}\mathbb{E}_{t-1}Z^p \overset{(37)}{\le} \frac{H}{2}\sigma_B(1-\beta_2)^{1-1/p}. \tag{45}$$

Combining Eqs. (39), (44) and (45) gives

$$\mathbb{E}_{t-1}\left\|g_{t,\mathrm{core}} - \nabla_t\right\|^2 \le \sigma_B^2(1-\beta_2)^{1-2/p} + \frac{5}{2}\sigma_B(1-\beta_2)^{1-1/p}H,$$

which is Eq. (36). □

C.4 Controlling the Tail Contribution

Let $m_{t,\mathrm{tail}}^{(j)}$ and $v_{t,\mathrm{tail}}^{(j)}$ be the first- and second-moment buffers generated by $g_{t,\mathrm{tail}}$ alone, per Eq. (31).

Lemma C.4 (Sparse tails gain a square-root memory factor)**.** *Under the memory window in Eq.* (2)*,*

$$\mathbb{E}\left[\sum_{j=1}^{J}\frac{\left\|m_{t,\mathrm{tail}}^{(j)}\right\|^2}{\sqrt{v_{t,\mathrm{tail}}^{(j)}}}\right] \le 6\sqrt{d}\,\sigma_B(1-\beta_2)^{3/2-1/p}. \tag{46}$$

with the convention $0/0 = 0$*.*

Proof. Fix a block j and set $A_s := (1-\beta_1)\beta_1^{t-s}$ and $B_s := (1-\beta_2)\beta_2^{t-s}$. By Eq. (31),

$$m_{t,\mathrm{tail}}^{(j)} = \sum_{s=1}^{t} A_s g_{s,\mathrm{tail}}^{(j)}, \qquad d_j v_{t,\mathrm{tail}}^{(j)} = \sum_{s=1}^{t} B_s \left\|g_{s,\mathrm{tail}}^{(j)}\right\|^2. \tag{47}$$

We first derive a deterministic inequality. If $\sum_{s=1}^{t} B_s \left\|g_{s,\mathrm{tail}}^{(j)}\right\|^2 = 0$, then every tail vector in block j is zero and the claim is immediate under the convention $0/0 = 0$. Assume below that this denominator is positive. Expanding the squared numerator gives

$$\left\|\sum_{s=1}^{t} A_s g_{s,\mathrm{tail}}^{(j)}\right\|^2 = \sum_{s,k=1}^{t} A_s A_k \left\langle g_{s,\mathrm{tail}}^{(j)}, g_{k,\mathrm{tail}}^{(j)}\right\rangle \le \sum_{s,k=1}^{t} A_s A_k \left\|g_{s,\mathrm{tail}}^{(j)}\right\| \left\|g_{k,\mathrm{tail}}^{(j)}\right\|.$$

If I_s does not occur, then $g_{s,\mathrm{tail}} = 0$. If I_s occurs, the denominator satisfies

$$\sqrt{\sum_{\ell=1}^{t} B_\ell \left\|g_{\ell,\mathrm{tail}}^{(j)}\right\|^2} \ge \sqrt{B_s}\left\|g_{s,\mathrm{tail}}^{(j)}\right\|.$$

Dividing term by term therefore yields

$$\frac{\left\|m_{t,\text{tail}}^{(j)}\right\|^2}{\sqrt{v_{t,\text{tail}}^{(j)}}} = \sqrt{d_j}\frac{\left\|\sum_{s=1}^t A_s g_{s,\text{tail}}^{(j)}\right\|^2}{\sqrt{\sum_{s=1}^t B_s \left\|g_{s,\text{tail}}^{(j)}\right\|^2}} \le \sum_{s,k=1}^t \frac{A_s A_k}{\sqrt{B_s}} \mathbb{I}_{I_s} \sqrt{d_j} \left\|g_{k,\text{tail}}^{(j)}\right\|. \tag{48}$$

After summing Eq. (48) over the blocks, Cauchy–Schwarz gives

$$\sum_{j=1}^J \sqrt{d_j} \left\|g_{k,\text{tail}}^{(j)}\right\| \le \left(\sum_{j=1}^J d_j\right)^{1/2} \left(\sum_{j=1}^J \left\|g_{k,\text{tail}}^{(j)}\right\|^2\right)^{1/2} = \sqrt{d}\,\|g_{k,\text{tail}}\|. \tag{49}$$

Taking expectations after summing Eq. (48) over the blocks gives

$$\mathbb{E}\left[\sum_{j=1}^J \frac{\left\|m_{t,\text{tail}}^{(j)}\right\|^2}{\sqrt{v_{t,\text{tail}}^{(j)}}}\right] \le \sqrt{d} \sum_{s,k=1}^t \frac{A_s A_k}{\sqrt{B_s}} \mathbb{E}\left[\mathbb{I}_{I_s}\,\|g_{k,\text{tail}}\|\right]. \tag{50}$$

We next evaluate the deterministic weights. If $\beta_2 = 0$, then the memory window forces $\beta_1 = 0$, and all three sums below equal one. Suppose henceforth that $\beta_2 > 0$. The finite geometric sum formula gives

$$\sum_{s=1}^t \frac{A_s^2}{\sqrt{B_s}} = \frac{(1-\beta_1)^2}{\sqrt{1-\beta_2}} \frac{1-\left(\beta_1^2/\sqrt{\beta_2}\right)^t}{1-\beta_1^2/\sqrt{\beta_2}} \le \frac{(1-\beta_1)^2}{\sqrt{1-\beta_2}\left(1-\beta_1^2/\sqrt{\beta_2}\right)}, \tag{51}$$

$$\sum_{s=1}^t \frac{A_s}{\sqrt{B_s}} = \frac{1-\beta_1}{\sqrt{1-\beta_2}} \frac{1-\left(\beta_1/\sqrt{\beta_2}\right)^t}{1-\beta_1/\sqrt{\beta_2}} \le \frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\beta_1/\sqrt{\beta_2}\right)}, \tag{52}$$

$$\sum_{s=1}^t A_s = 1-\beta_1^t. \tag{53}$$

The ratios in Eqs. (51) and (52) lie in $[0,1)$ because $\beta_2^2 \le \beta_1 \le \beta_2$. To bound the infinite sums sharply, set $q := \sqrt{\beta_2}$ and $a := \beta_1$, so $q^4 \le a \le q^2 < q$. For the linear weight in Eq. (52),

$$\sqrt{1-\beta_2}\sum_{s=1}^t \frac{A_s}{\sqrt{B_s}} \le \frac{1-a}{1-a/q} = \frac{q(1-a)}{q-a} \le \frac{q(1-q^2)}{q-q^2} = 1+q \le 2.$$

The inequality before the endpoint value follows because $q(1-a)/(q-a)$ is increasing in a, with derivative $q(1-q)/(q-a)^2 \ge 0$. For the squared weight in Eq. (51),

$$\frac{1}{\sqrt{1-\beta_2}}\sum_{s=1}^t \frac{A_s^2}{\sqrt{B_s}} \le \frac{(1-a)^2}{(1-q^2)(1-a^2/q)} \le \frac{(1-q^4)^2}{(1-q^2)(1-q^7)} = \frac{(1-q^2)(1+q^2)^2}{1-q^7} \le 2.$$

The second inequality follows because the middle expression decreases with a. Indeed,

$$\frac{\mathrm{d}}{\mathrm{d}a}\log\left(\frac{(1-a)^2}{1-a^2/q}\right) = \frac{2(a-q)}{(1-a)(q-a^2)} < 0.$$

The final inequality follows from

$$2(1-q^7) - (1-q^2)(1+q^2)^2 = (1-q)(2q^6+q^5+q^4+q+1) \ge 0.$$

Consequently,

$$\sum_{s=1}^t \frac{A_s^2}{\sqrt{B_s}} \le 2\sqrt{1-\beta_2}, \qquad \sum_{s=1}^t \frac{A_s}{\sqrt{B_s}} \le \frac{2}{\sqrt{1-\beta_2}}, \qquad \sum_{s=1}^t A_s \le 1. \tag{54}$$

For the diagonal terms $s = k$, Eqs. (34) and (54) gives

$$\sqrt{d}\sum_{s=1}^t \frac{A_s^2}{\sqrt{B_s}} \mathbb{E}\left[\|g_{s,\text{tail}}\|\right] \le 3\sqrt{d}\,\sigma_B (1-\beta_2)^{3/2-1/p}.$$

For the off-diagonal terms $s \neq k$, we proceed with the tower property together with the conditional bounds in Eqs. (33) and (34). We distinguish the two time orders.

If $s < k$, then $I_s \in \mathcal{F}_s \subseteq \mathcal{F}_{k-1}$. Hence $\mathbb{I}_{I_s}$ is $\mathcal{F}_{k-1}$-measurable, and

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{I_s}\left\|g_{k,\text{tail}}\right\|\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{I_s}\left\|g_{k,\text{tail}}\right\| \mid \mathcal{F}_{k-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{I}_{I_s}\mathbb{E}_{k-1}\left[\left\|g_{k,\text{tail}}\right\|\right]\right] \\ &\overset{(34)}{\leq} \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p}\mathbb{E}\left[\mathbb{I}_{I_s}\right] \\ &= \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p}\mathbb{P}(I_s) \\ &= \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p}\mathbb{E}\left[\mathbb{P}(I_s \mid \mathcal{F}_{s-1})\right] \\ &\overset{(33)}{\leq} \frac{3}{2}\sigma_B(1-\beta_2)^{2-1/p}.\end{aligned}$$

If $k < s$, then $g_{k,\text{tail}} \in \mathcal{F}_k \subseteq \mathcal{F}_{s-1}$. Thus $\|g_{k,\text{tail}}\|$ is $\mathcal{F}_{s-1}$-measurable, and

$$\begin{aligned}\mathbb{E}\left[\mathbb{I}_{I_s}\left\|g_{k,\text{tail}}\right\|\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}_{I_s}\left\|g_{k,\text{tail}}\right\| \mid \mathcal{F}_{s-1}\right]\right] \\ &= \mathbb{E}\left[\left\|g_{k,\text{tail}}\right\|\mathbb{P}(I_s \mid \mathcal{F}_{s-1})\right] \\ &\overset{(33)}{\leq} (1-\beta_2)\mathbb{E}\left[\left\|g_{k,\text{tail}}\right\|\right] \\ &= (1-\beta_2)\mathbb{E}\left[\mathbb{E}_{k-1}\left[\left\|g_{k,\text{tail}}\right\|\right]\right] \\ &\overset{(34)}{\leq} \frac{3}{2}\sigma_B(1-\beta_2)^{2-1/p}.\end{aligned}$$

Therefore, for both time orders,

$$\mathbb{E}\left[\mathbb{I}_{I_s}\left\|g_{k,\text{tail}}\right\|\right] \leq \frac{3}{2}\sigma_B(1-\beta_2)^{2-1/p}, \qquad s \neq k. \tag{55}$$

Finally, we can compute the total off-diagonal quantity. In view of

$$\sum_{\substack{1\leq s,k\leq t\\ s\neq k}} \frac{A_s A_k}{\sqrt{B_s}} = \left(\sum_{s=1}^{t}\frac{A_s}{\sqrt{B_s}}\right)\left(\sum_{k=1}^{t}A_k\right) - \sum_{s=1}^{t}\frac{A_s^2}{\sqrt{B_s}} \leq \left(\sum_{s=1}^{t}\frac{A_s}{\sqrt{B_s}}\right)\left(\sum_{k=1}^{t}A_k\right) \overset{(54)}{\leq} \frac{2}{\sqrt{1-\beta_2}},$$

and Eq. (55), the off-diagonal contribution is at most $3\sqrt{d}\sigma_B(1-\beta_2)^{3/2-1/p}$. Adding the diagonal and off-diagonal contributions proves Eq. (46). □

C.5 Momentum Error and Descent

Let $m_{t,\text{core}}$ be the first-moment buffer generated by $g_{t,\text{core}}$. At the point where they first enter the tracking argument, define

$$e_t := m_{t,\text{core}} - \nabla_t, \qquad b_t := \nabla_t - \mathbb{E}_{t-1}\left[g_{t,\text{core}}\right], \qquad \zeta_t := g_{t,\text{core}} - \mathbb{E}_{t-1}\left[g_{t,\text{core}}\right]. \tag{56}$$

Lemma C.5 (Momentum tracking)**.** *With $\nabla_0 = \nabla_1$ and $e_0 = -\nabla_1$,*

$$\mathbb{E}_{t-1}[\|e_t\|^2] \leq \beta_1\|e_{t-1}\|^2 + \frac{2\beta_1^2}{1-\beta_1}\|\nabla_t - \nabla_{t-1}\|^2 + 2(1-\beta_1)\|b_t\|^2 + (1-\beta_1)^2\mathbb{E}_{t-1}[\|\zeta_t\|^2].$$

Proof. By Eq. (31), the core momentum obeys

$$m_{t,\text{core}} = \beta_1 m_{t-1,\text{core}} + (1-\beta_1)g_{t,\text{core}}.$$

The definitions of b_t and ζ_t give

$$g_{t,\text{core}} = \nabla_t - b_t + \zeta_t.$$

Since $m_{t-1,\text{core}} = e_{t-1} + \nabla_{t-1}$, subtracting ∇_t from the core-momentum recursion yields the exact identity

$$e_t = \beta_1 e_{t-1} - \beta_1(\nabla_t - \nabla_{t-1}) - (1-\beta_1)b_t + (1-\beta_1)\zeta_t. \tag{57}$$

Conditionally on $\mathcal{F}_{t-1}$, the first three terms on the right are deterministic and $\mathbb{E}_{t-1}[\zeta_t] = 0$. Therefore,

$$\mathbb{E}_{t-1}\left[\|e_t\|^2\right] = \|\beta_1 e_{t-1} - \beta_1(\nabla_t - \nabla_{t-1}) - (1-\beta_1)b_t\|^2 + (1-\beta_1)^2\mathbb{E}_{t-1}\left[\|\zeta_t\|^2\right].$$

For any vectors a, c and $\beta \in [0, 1)$,

$$\|\beta a + c\|^2 \le \beta \|a\|^2 + \frac{1}{1-\beta}\|c\|^2. \tag{58}$$

Apply Eq. (58) with $a = e_{t-1}$ and $c = -\beta_1(\nabla_t - \nabla_{t-1}) - (1-\beta_1)b_t$. Then use $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ to obtain

$$\mathbb{E}_{t-1}[\|e_t\|^2] \le \beta_1 \|e_{t-1}\|^2 + \frac{2\beta_1^2}{1-\beta_1}\|\nabla_t - \nabla_{t-1}\|^2 + 2(1-\beta_1)\|b_t\|^2 + (1-\beta_1)^2\mathbb{E}_{t-1}[\|\zeta_t\|^2],$$

which completes the proof. □

Define

$$A_t := \sum_j \frac{\left\|\nabla_t^{(j)}\right\|^2}{D_t^{(j)}}, \quad C_t := \sum_j \frac{\left\|m_{t,\text{core}}^{(j)}\right\|^2}{D_t^{(j)}}, \quad \mathcal{T}_t := \sum_j \frac{\left\|m_{t,\text{tail}}^{(j)}\right\|^2}{D_t^{(j)}}. \tag{59}$$

Lemma C.6 (Progress bound)**.** *Under the choices in Theorem 1,*

$$\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}[A_t] \le 42\sqrt{\frac{\Delta L}{T}} + \frac{5}{2}\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathsf{S}_{\mathcal{P}}(\nabla_t)]. \tag{60}$$

Proof. By Eq. (32), $m_t = m_{t,\text{core}} + m_{t,\text{tail}}$. For the core term, the identity $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ gives

$$\left\langle \nabla_t^{(j)}, m_{t,\text{core}}^{(j)} \right\rangle = \frac{1}{2}\left\|\nabla_t^{(j)}\right\|^2 + \frac{1}{2}\left\|m_{t,\text{core}}^{(j)}\right\|^2 - \frac{1}{2}\left\|e_t^{(j)}\right\|^2.$$

For the tail term, Young's inequality gives

$$\left\langle \nabla_t^{(j)}, m_{t,\text{tail}}^{(j)} \right\rangle \ge -\frac{1}{4}\left\|\nabla_t^{(j)}\right\|^2 - \left\|m_{t,\text{tail}}^{(j)}\right\|^2.$$

Adding the two displays proves

$$\left\langle \nabla_t^{(j)}, m_t^{(j)} \right\rangle \ge \frac{1}{4}\left\|\nabla_t^{(j)}\right\|^2 + \frac{1}{2}\left\|m_{t,\text{core}}^{(j)}\right\|^2 - \frac{1}{2}\left\|e_t^{(j)}\right\|^2 - \left\|m_{t,\text{tail}}^{(j)}\right\|^2. \tag{61}$$

Let $u_t^{(j)} := m_t^{(j)}/D_t^{(j)}$. Since F is L-smooth and $x_{t+1} = x_t - \eta u_t$,

$$F(x_{t+1}) \le F(x_t) - \eta\sum_{j=1}^{J}\frac{\left\langle \nabla_t^{(j)}, m_t^{(j)}\right\rangle}{D_t^{(j)}} + \frac{L\eta^2}{2}\|u_t\|^2.$$

Because $D_t^{(j)} \ge \epsilon_{\text{alg}}$ and $m_t^{(j)} = m_{t,\text{core}}^{(j)} + m_{t,\text{tail}}^{(j)}$,

$$\|u_t\|^2 \le 2\sum_{j=1}^{J}\frac{\left\|m_{t,\text{core}}^{(j)}\right\|^2 + \left\|m_{t,\text{tail}}^{(j)}\right\|^2}{(D_t^{(j)})^2} \le \frac{2}{\epsilon_{\text{alg}}}(C_t + \mathcal{T}_t). \tag{62}$$

Substituting Eq. (61) and using $D_t^{(j)} \ge \epsilon_{\text{alg}}$ for the tracking error gives the pathwise inequality

$$F(x_{t+1}) \le F(x_t) - \frac{\eta}{4}A_t - \frac{\eta}{2}C_t + \frac{\eta}{2\epsilon_{\text{alg}}}\|e_t\|^2 + \eta\mathcal{T}_t + \frac{L\eta^2}{\epsilon_{\text{alg}}}(C_t + \mathcal{T}_t). \tag{63}$$

Taking full expectations in Eq. (63) and summing from $t = 1$ to T gives a telescoping objective difference. Since $\sum_{t=1}^{T} \mathbb{E}\left[F(x_t) - F(x_{t+1})\right] \leq F(x_1) - \inf_x F(x) \leq \Delta$ by Assumption 1, we obtain

$$\frac{\eta}{4}\sum_{t=1}^{T}\mathbb{E}\left[A_t\right] + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[C_t\right] \leq \Delta + \frac{\eta}{2\epsilon_{\text{alg}}}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|^2\right] + \eta\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{T}_t\right] + \frac{L\eta^2}{\epsilon_{\text{alg}}}\sum_{t=1}^{T}\mathbb{E}\left[C_t + \mathcal{T}_t\right]. \tag{64}$$

Thus, the only term in Eq. (64) that is not yet expressed through C_t, $\mathcal{T}_t$, or lower-order quantities (those can be bounded later) is the accumulated tracking error. To control it, sum Lemma C.5 from $t = 1$ to T. The index shift satisfies

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|^2\right] - \beta_1\sum_{t=1}^{T}\mathbb{E}\left[\|e_{t-1}\|^2\right] &= (1-\beta_1)\sum_{t=1}^{T-1}\mathbb{E}\left[\|e_t\|^2\right] + \mathbb{E}\left[\|e_T\|^2\right] - \beta_1\|e_0\|^2 \\ &\geq (1-\beta_1)\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|^2\right] - \beta_1\|e_0\|^2.\end{aligned}$$

After rearranging the terms and dividing by $1 - \beta_1$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\|e_t\|^2\right] \leq \frac{\beta_1}{1-\beta_1}\|e_0\|^2 + \frac{2\beta_1^2}{(1-\beta_1)^2}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_t - \nabla_{t-1}\|^2\right] \\ + 2\sum_{t=1}^{T}\mathbb{E}\left[\|b_t\|^2\right] + (1-\beta_1)\sum_{t=1}^{T}\mathbb{E}\left[\|\zeta_t\|^2\right].\end{aligned} \tag{65}$$

Here we set $\nabla_0 = \nabla_1$, so the drift at $t = 1$ is zero. For $t \geq 2$, smoothness and the update give

$$\|\nabla_t - \nabla_{t-1}\|^2 \leq L^2\|x_t - x_{t-1}\|^2 \leq L^2\eta^2\|u_{t-1}\|^2 \leq \frac{2L^2\eta^2}{\epsilon_{\text{alg}}}(C_{t-1} + \mathcal{T}_{t-1}).$$

We set $C_0 = \mathcal{T}_0 = 0$. Combining the drift bound above with Eq. (65) and then inserting the result into Eq. (64) gives

$$\begin{aligned}&\frac{\eta}{4}\sum_{t=1}^{T}\mathbb{E}\left[A_t\right] + \left(\frac{\eta}{2} - \frac{L\eta^2}{\epsilon_{\text{alg}}} - 2\eta\beta_1^2\left(\frac{L\eta}{\epsilon_{\text{alg}}(1-\beta_1)}\right)^2\right)\sum_{t=1}^{T}\mathbb{E}\left[C_t\right] \\ &\qquad\leq \Delta + \frac{\eta\beta_1}{2\epsilon_{\text{alg}}(1-\beta_1)}\|e_0\|^2 + \frac{\eta}{\epsilon_{\text{alg}}}\sum_{t=1}^{T}\mathbb{E}\left[\|b_t\|^2\right] + \frac{\eta(1-\beta_1)}{2\epsilon_{\text{alg}}}\sum_{t=1}^{T}\mathbb{E}\left[\|\zeta_t\|^2\right] \\ &\qquad\quad + \left(\eta + \frac{L\eta^2}{\epsilon_{\text{alg}}} + 2\eta\beta_1^2\left(\frac{L\eta}{\epsilon_{\text{alg}}(1-\beta_1)}\right)^2\right)\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{T}_t\right].\end{aligned} \tag{66}$$

We verify the two absorptions in Eq. (66). By Eqs. (2) and (22),

$$\epsilon_{\text{alg}} = 4\frac{\sigma_B}{\sqrt{d}}(1-\beta_2)^{1/2-1/p} \geq \frac{4L\eta}{1-\beta_1}, \qquad \frac{L\eta}{\epsilon_{\text{alg}}(1-\beta_1)} \leq \frac{1}{4}. \tag{67}$$

Since $\beta_1 \leq 1$, the tracking-drift contribution in either coefficient is bounded by

$$2\eta\beta_1^2\left(\frac{L\eta}{\epsilon_{\text{alg}}(1-\beta_1)}\right)^2 \leq \frac{\eta}{8}, \text{ and that } \frac{L\eta^2}{\epsilon_{\text{alg}}} \leq \frac{\eta(1-\beta_1)}{4} \leq \frac{\eta}{4}.$$

Therefore the coefficient of $\sum_t \mathbb{E}\left[C_t\right]$ on the left of Eq. (66) is at least $\eta/2 - \eta/4 - \eta/8 = \eta/8 \geq 0$ and may be discarded. Similarly, the coefficient of $\sum_t \mathbb{E}\left[\mathcal{T}_t\right]$ on the right is at most $\eta + \eta/4 + \eta/8 = 11\eta/8$. Dividing the resulting inequality by $\eta/4$ and then by T yields

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[A_t\right] \leq \frac{4\Delta}{\eta T} + \frac{2\beta_1\|e_0\|^2}{\epsilon_{\text{alg}}(1-\beta_1)T} + \frac{11}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{T}_t\right] \\ + \frac{4}{\epsilon_{\text{alg}}T}\sum_{t=1}^{T}\mathbb{E}\left[\|b_t\|^2\right] + \frac{2(1-\beta_1)}{\epsilon_{\text{alg}}T}\sum_{t=1}^{T}\mathbb{E}\left[\|\zeta_t\|^2\right].\end{aligned} \tag{68}$$

We bound the five terms on the right-hand side of Eq. (68), after dividing each by $\sqrt{d}$. We repeatedly use the two consequences of the parameter choices

$$\sqrt{\frac{\Delta L}{T}} = \sigma_B(1-\beta_2)^{3/2-1/p}, \qquad \sqrt{d}\,\epsilon_{\text{alg}} = 4\sigma_B(1-\beta_2)^{1/2-1/p}, \tag{69}$$

where the first identity follows from the tuning in Eq. (2), and the second equality follows directly from the same display.

First, substituting $\eta = \sqrt{\Delta/(LdT)}$ from Eq. (2) gives

$$\frac{4\Delta}{\eta T\sqrt{d}} = 4\sqrt{\frac{\Delta L}{T}}. \tag{70}$$

For the initialization term, $e_0 = -\nabla_1$. The standard smoothness descent step $F(x_1 - \nabla_1/L) \le F(x_1) - \|\nabla_1\|^2/(2L)$, together with Assumption 1, yields $\|e_0\|^2 = \|\nabla_1\|^2 \le 2L\Delta$. Moreover, Eqs. (2) and (22) give $\epsilon_{\text{alg}}(1-\beta_1) \ge 4L\eta$. Hence

$$\frac{2\beta_1\|e_0\|^2}{\sqrt{d}\,\epsilon_{\text{alg}}(1-\beta_1)T} \le \frac{4L\Delta}{4\sqrt{d}\,L\eta T} = \frac{\Delta}{\sqrt{d}\,\eta T} = \sqrt{\frac{\Delta L}{T}}. \tag{71}$$

For the tail term, the denominator satisfies $D_t^{(j)} \ge \sqrt{v_t^{(j)}} \ge \sqrt{v_{t,\text{tail}}^{(j)}}$. By Eq. (59), $\mathcal{T}_t \le \sum_{j=1}^{J} \left\|m_{t,\text{tail}}^{(j)}\right\|^2 / \sqrt{v_{t,\text{tail}}^{(j)}}$. Applying Lemma C.4 at each t and then using Eq. (69),

$$\frac{11}{2\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{T}_t\right] \le \frac{11}{2\sqrt{d}} \cdot 6\sqrt{d}\,\sigma_B(1-\beta_2)^{3/2-1/p} = 33\sqrt{\frac{\Delta L}{T}}. \tag{72}$$

For the truncation bias, Eqs. (35) and (56) give $\|b_t\| \le \frac{3}{2}\sigma_B(1-\beta_2)^{1-1/p}$ almost surely. Consequently, by the second identity in Eq. (69),

$$\frac{4}{\sqrt{d}\,\epsilon_{\text{alg}}T}\sum_{t=1}^{T}\mathbb{E}\left[\|b_t\|^2\right] \le \frac{9\sigma_B^2(1-\beta_2)^{2-2/p}}{\sqrt{d}\,\epsilon_{\text{alg}}} \le \frac{9}{4}\sigma_B(1-\beta_2)^{3/2-1/p} = \frac{9}{4}\sqrt{\frac{\Delta L}{T}}. \tag{73}$$

Finally, since conditional expectation minimizes mean squared distance,

$$\mathbb{E}_{t-1}\left[\|\zeta_t\|^2\right] = \mathbb{E}_{t-1}\left[\|g_{t,\text{core}} - \mathbb{E}_{t-1}\left[g_{t,\text{core}}\right]\|^2\right] \le \mathbb{E}_{t-1}\left[\|g_{t,\text{core}} - \nabla_t\|^2\right].$$

Using Eq. (36), taking full expectations, and using $1-\beta_1 \le 2(1-\beta_2)$ together with Eq. (69), we obtain

$$\begin{aligned}\frac{2(1-\beta_1)}{\sqrt{d}\,\epsilon_{\text{alg}}T}\sum_{t=1}^{T}\mathbb{E}\left[\|\zeta_t\|^2\right] &\le \frac{4(1-\beta_2)}{\sqrt{d}\,\epsilon_{\text{alg}}}\left(\sigma_B^2(1-\beta_2)^{1-2/p} + \frac{5}{2}\sigma_B(1-\beta_2)^{1-1/p}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_t\|\right]\right)\\ &\le \sqrt{\frac{\Delta L}{T}} + \frac{5}{2}(1-\beta_2)^{3/2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_t\|\right].\end{aligned} \tag{74}$$

The bounds Eqs. (70) to (74) explain separately how every contribution is reduced to the common scale $\sqrt{\Delta L/T}$, apart from the final gradient-dependent term. Combining the displays and using $\|\nabla_t\| \le \sqrt{d/d_{\min}}\mathsf{S}_{\mathcal{P}}(\nabla_t)$, we obtain

$$\begin{aligned}\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[A_t\right] &\le \left(4+1+33+\frac{9}{4}+1\right)\sqrt{\frac{\Delta L}{T}} + \frac{5}{2}\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right]\\ &\le 42\sqrt{\frac{\Delta L}{T}} + \frac{5}{2}\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right],\end{aligned}$$

which proves Eq. (60). □

C.6 Bounding the Denominators

Lemma C.7 (Denominator mass in the natural geometry)**.** *Under the assumptions of Theorem 1,*

$$\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{J} d_j D_t^{(j)}\right] \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right] + 9\sigma_B(1-\beta_2)^{1/2-1/p}, \tag{75}$$

where $\mathsf{S}_{\mathcal{P}}(\nabla_t) := \frac{1}{\sqrt{d}}\sum_{j=1}^{J}\sqrt{d_j}\left\|\nabla_t^{(j)}\right\|$.

Proof. Recall that $g_s = \nabla_s + \xi_s$. For every $t \in [T]$ and $j \in [J]$, define the signal and noise buffers

$$v_{t,\mathrm{sig}}^{(j)} := (1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\frac{\left\|\nabla_s^{(j)}\right\|^2}{d_j}, \qquad v_{t,\mathrm{noise}}^{(j)} := (1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\frac{\left\|\xi_s^{(j)}\right\|^2}{d_j}. \tag{76}$$

Here "signal" refers to the true gradients ∇_s, whereas "noise" refers to the oracle errors $\xi_s = g_s - \nabla_s$. To see why these two parts can be separated, view $\left\{\sqrt{(1-\beta_2)\beta_2^{t-s}/d_j}\, g_s^{(j)}\right\}_{s=1}^{t}$ as one vector in the Hilbert direct sum $(\mathbb{R}^{d_j})^t$. Since $g_s^{(j)} = \nabla_s^{(j)} + \xi_s^{(j)}$, Minkowski's inequality gives

$$\sqrt{v_t^{(j)}} = \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\frac{\left\|\nabla_s^{(j)}+\xi_s^{(j)}\right\|^2}{d_j}} \le \sqrt{v_{t,\mathrm{sig}}^{(j)}} + \sqrt{v_{t,\mathrm{noise}}^{(j)}}.$$

Therefore, using $D_t^{(j)} = \sqrt{v_t^{(j)}} + \epsilon_{\mathrm{alg}}$,

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J} d_j D_t^{(j)} \le \underbrace{\frac{1}{\sqrt{d}}\sum_{j=1}^{J} d_j\sqrt{v_{t,\mathrm{sig}}^{(j)}}}_{\text{signal mass}} + \underbrace{\frac{1}{\sqrt{d}}\sum_{j=1}^{J} d_j\sqrt{v_{t,\mathrm{noise}}^{(j)}}}_{\text{noise mass}} + \underbrace{\sqrt{d}\,\epsilon_{\mathrm{alg}}}_{\text{floor mass}}. \tag{77}$$

We bound these three terms separately. We will repeatedly use the tuning identity $\sqrt{\frac{\Delta L}{T}} = \sigma_B(1-\beta_2)^{3/2-1/p}$ from Eq. (69).

Signal contribution. For each $s \in [t]$, write $\nabla_s^{(j)} = \nabla_t^{(j)} + (\nabla_s^{(j)} - \nabla_t^{(j)})$. A second application of Minkowski's inequality, now to the signal buffer, gives

$$\sqrt{v_{t,\mathrm{sig}}^{(j)}} \le \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\frac{\left\|\nabla_t^{(j)}\right\|^2}{d_j}} + \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\frac{\left\|\nabla_s^{(j)}-\nabla_t^{(j)}\right\|^2}{d_j}}.$$

Since $(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s} = 1-\beta_2^t \le 1$, multiplying by $d_j/\sqrt{d}$ and summing over $j \in [J]$ shows that the first term is at most $\mathsf{S}_{\mathcal{P}}(\nabla_t)$. For the drift term, Cauchy–Schwarz across the blocks gives

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}\sqrt{d_j}\sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|\nabla_s^{(j)}-\nabla_t^{(j)}\right\|^2} \le \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|\nabla_s-\nabla_t\right\|^2}.$$

Consequently,

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J} d_j\sqrt{v_{t,\mathrm{sig}}^{(j)}} \le \mathsf{S}_{\mathcal{P}}(\nabla_t) + \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|\nabla_s-\nabla_t\right\|^2}. \tag{78}$$

By Eq. (27) and Assumption 2, for every $1 \le s \le t \le T$,

$$\left\|\nabla_s - \nabla_t\right\| \le L\left\|x_s - x_t\right\| \le 2L\eta\sqrt{d}\,(t-s).$$

Using $\eta = \sqrt{\Delta/(LdT)}$ and changing variables $k = t - s$,

$$(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|\nabla_s - \nabla_t\right\|^2 \le 4L^2\eta^2 d(1-\beta_2)\sum_{k=0}^{t-1}k^2\beta_2^k \le \frac{8L^2\eta^2 d}{(1-\beta_2)^2} = \frac{8\Delta L}{T(1-\beta_2)^2},$$

where we used $\sum_{k=0}^{\infty}k^2\beta_2^k = \beta_2(1+\beta_2)/(1-\beta_2)^3 \le 2(1-\beta_2)^{-3}$. Taking square roots and using Eq. (69) gives,

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{v_{t,\mathrm{sig}}^{(j)}} \le \mathsf{S}_{\mathcal{P}}(\nabla_t) + 2\sqrt{2}\,\sigma_B(1-\beta_2)^{1/2-1/p}. \tag{79}$$

Noise contribution. From Eq. (76),

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{v_{t,\mathrm{noise}}^{(j)}} = \frac{1}{\sqrt{d}}\sum_{j=1}^{J}\sqrt{d_j}\sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|\xi_s^{(j)}\right\|^2} \le \underbrace{\sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\left\|\xi_s\right\|^2}}_{:=Z_t},$$

where the last step is Cauchy–Schwarz over $j = 1, \dots, J$. Since $p/2 \le 1$, the map $x \mapsto x^{p/2}$ is subadditive on $\mathbb{R}_+$. Thus

$$\mathbb{E}\left[Z_t^p\right] \le (1-\beta_2)^{p/2}\sum_{s=1}^{t}\beta_2^{(t-s)p/2}\mathbb{E}\left[\left\|\xi_s\right\|^p\right] \le \sigma_B^p(1-\beta_2)^{p/2}\frac{1}{1-\beta_2^{p/2}}.$$

For $p/2 \in (1/2, 1]$, concavity gives $1 - \beta_2^{p/2} \ge p(1-\beta_2)/2$. Therefore,

$$\mathbb{E}\left[\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{v_{t,\mathrm{noise}}^{(j)}}\right] \le \mathbb{E}\left[Z_t\right] \le \left(\mathbb{E}\left[Z_t^p\right]\right)^{1/p} \le 2\sigma_B(1-\beta_2)^{1/2-1/p}. \tag{80}$$

Stabilizer contribution. By Eq. (2),

$$\sqrt{d}\,\epsilon_{\mathrm{alg}} = 4\sigma_B(1-\beta_2)^{1/2-1/p}. \tag{81}$$

Finally, average Eq. (77) over $t = 1, \dots, T$, take expectations, and apply Eqs. (79) to (81). This gives

$$\begin{aligned}\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{J}d_j D_t^{(j)}\right] &\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right] + (2\sqrt{2}+2+4)\sigma_B(1-\beta_2)^{1/2-1/p}\\ &\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right] + 9\sigma_B(1-\beta_2)^{1/2-1/p},\end{aligned}$$

which is exactly Eq. (75). □

C.7 Proof of Theorem 1

Proof of Theorem 1. Recall from Eq. (59) that

$$A_t = \sum_{j=1}^{J}\frac{\left\|\nabla_t^{(j)}\right\|^2}{D_t^{(j)}}, \qquad D_t^{(j)} = \sqrt{v_t^{(j)}} + \epsilon_{\mathrm{alg}}.$$

Write $H_T := T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right]$. For each fixed t, Cauchy–Schwarz over the blocks gives

$$\begin{aligned}\mathsf{S}_{\mathcal{P}}(\nabla_t)^2 &= \frac{1}{d}\left(\sum_{j=1}^{J}\frac{\left\|\nabla_t^{(j)}\right\|}{\sqrt{D_t^{(j)}}}\sqrt{d_j D_t^{(j)}}\right)^2\\ &\le \left(\frac{1}{\sqrt{d}}\sum_{j=1}^{J}\frac{\left\|\nabla_t^{(j)}\right\|^2}{D_t^{(j)}}\right)\left(\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j D_t^{(j)}\right) = \left(\frac{A_t}{\sqrt{d}}\right)\left(\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j D_t^{(j)}\right).\end{aligned} \tag{82}$$

Taking square roots in Eq. (82), summing over $t = 1, \dots, T$, and applying Cauchy–Schwarz on the product space $[T] \times \Omega$ yields

$$H_T^2 \le \left(\frac{1}{\sqrt{d}T} \sum_{t=1}^{T} \mathbb{E}\left[A_t\right] \right) \left(\frac{1}{\sqrt{d}T} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{j=1}^{J} d_j D_t^{(j)} \right] \right). \tag{83}$$

Indeed, this is the usual inequality $\left(\int \sqrt{XY}\, \mathrm{d}\mu \right)^2 \le \left(\int X\, \mathrm{d}\mu \right) \left(\int Y\, \mathrm{d}\mu \right)$ for the probability measure that is uniform over $t \in [T]$ and uses the underlying algorithm–oracle probability on Ω.

Lemma C.6 bounds the first factor in Eq. (83), while Lemma C.7 bounds the second. Substituting their two conclusions gives

$$H_T^2 \le \left(42\sqrt{\frac{\Delta L}{T}} + \frac{5}{2}\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2} H_T \right) \left(H_T + 9\sigma_B (1-\beta_2)^{1/2-1/p} \right). \tag{84}$$

To solve this scalar inequality, put

$$s := \sqrt{1-\beta_2}, \qquad Y := \frac{H_T}{\sigma_B (1-\beta_2)^{1-1/p}}.$$

The tuning identity in Eq. (69) gives

$$\sqrt{\frac{\Delta L}{T}} = s\, \sigma_B (1-\beta_2)^{1-1/p},$$

and Eq. (3) gives

$$\frac{5}{2}\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2} \le \frac{5s}{16}.$$

After dividing Eq. (84) by $\sigma_B^2 (1-\beta_2)^{2-2/p}$, we obtain

$$\left(1 - \frac{5s}{16}\right) Y^2 - \left(42s + \frac{45}{16}\right) Y - 378 \le 0. \tag{85}$$

At $Y = 42s + 25$, multiplying the left-hand side of Eq. (85) by 16 gives

$$2827 + 11785s - 10500s^2 - 8820s^3.$$

Since $0 < s \le 1/\sqrt{8}$ and the quadratic $11785 - 10500s - 8820s^2$ is decreasing and positive at $s = 1/\sqrt{8}$, the displayed expression is positive. The quadratic in Eq. (85) opens upward and has a negative constant term, hence it has exactly one positive root. Since $Y \ge 0$, we conclude

$$H_T \le \sigma_B (1-\beta_2)^{1-1/p} (42s + 25) = 42\sqrt{\frac{\Delta L}{T}} + 25\sigma_B (1-\beta_2)^{1-1/p}.$$

Finally, the tuning in Eq. (2) gives

$$\sigma_B (1-\beta_2)^{1-1/p} = (\Delta L)^{\frac{p-1}{3p-2}}\, \sigma_B^{\frac{p}{3p-2}}\, T^{-\frac{p-1}{3p-2}},$$

which proves Eq. (4). □

C.8 Proof of Theorem 2

Proof of Theorem 2. We verify the three requirements in Eq. (5) directly. The deterministic term in Eq. (4) is at most $\varepsilon/2$ because

$$T \ge 84^2 \frac{\Delta L}{\varepsilon^2} \quad \Longrightarrow \quad 42\sqrt{\frac{\Delta L}{T}} \le \frac{\varepsilon}{2}.$$

The stochastic term is at most $\varepsilon/2$ exactly when

$$25(\Delta L)^{\frac{p-1}{3p-2}} \sigma_B^{\frac{p}{3p-2}} T^{-\frac{p-1}{3p-2}} \le \frac{\varepsilon}{2} \quad \Longleftrightarrow \quad T \ge 50^{\frac{3p-2}{p-1}} \frac{\Delta L}{\varepsilon^2} \left(\frac{\sigma_B}{\varepsilon} \right)^{\frac{p}{p-1}}.$$

Finally, substituting the tuning in Eq. (2) shows that

$$T \geq \frac{\Delta L}{\sigma_B^2}\left(8\sqrt{\frac{d}{d_{\min}}}\right)^{\frac{3p-2}{p}} \quad \Longrightarrow \quad 1-\beta_2 \leq \frac{1}{8}\sqrt{\frac{d_{\min}}{d}}.$$

This is Eq. (3). It also implies $1-\beta_2 < 1/4$. Thus all assumptions of Theorem 1 hold, and the uniform random output satisfies $\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(\widehat{x}))\right] \leq \varepsilon$.

We next show why Eq. (6) reduces the maximum of three budgets to one term. Let T_{det}, T_{stoch}, and T_{tr} denote the three displayed quantities in Eq. (5). The first term in the minimum in Eq. (6) is equivalent to

$$50^{\frac{3p-2}{p-1}}\left(\frac{\sigma_B}{\varepsilon}\right)^{\frac{p}{p-1}} \geq 84^2,$$

which gives $T_{\mathrm{stoch}} \geq T_{\mathrm{det}}$. Direct cancellation of the common factor ΔL gives

$$\frac{T_{\mathrm{stoch}}}{T_{\mathrm{tr}}} = \frac{50^{\frac{3p-2}{p-1}}\left(\frac{\sigma_B}{\varepsilon}\right)^{\frac{3p-2}{p-1}}}{\left(8\sqrt{d/d_{\min}}\right)^{\frac{3p-2}{p}}} \geq 1.$$

The second term in the minimum in Eq. (6) ensures that the displayed ratio is at least one. Hence $T_{\mathrm{stoch}} = \max\{T_{\mathrm{det}}, T_{\mathrm{stoch}}, T_{\mathrm{tr}}\}$. Its ceiling is exactly the sufficient choice T_ε in Eq. (7). Multiplying by B and using $\lceil x \rceil \leq x+1$ gives the second inequality in Eq. (7). □

D Proofs of the Upper Bounds in Section 3.2

We prove Theorems 3 and 4 for Algorithm 2 in this section. Eq. (3) and $d_{\min} \leq d$ imply $1-\beta_2 \leq 1/8 < 1$. First, we record the normalized geometric-weight bounds used throughout the proof. The restriction $\beta_2^2 \leq \beta_1 \leq \beta_2$ in Eq. (9) gives

$$1-\beta_2 \leq 1-\beta_1 \leq 2(1-\beta_2), \qquad 1-\beta_2^t \leq 1-\beta_1^t \leq 2(1-\beta_2^t).$$

For $1 \leq s \leq t$, the normalized first- and second-moment weights are denoted by

$$\alpha_{t,s} := \frac{(1-\beta_1)\beta_1^{t-s}}{1-\beta_1^t}, \qquad \gamma_{t,s} := \frac{(1-\beta_2)\beta_2^{t-s}}{1-\beta_2^t}, \qquad \rho_t := \frac{1-\beta_2}{1-\beta_2^t}.$$

Both $\{\alpha_{t,s}\}_{s\leq t}$ and $\{\gamma_{t,s}\}_{s\leq t}$ sum to one. We also use the pathwise radius

$$\omega_T := \max\left\{\lambda R_1, (1-\lambda\eta)^T\lambda R_1 + \sqrt{2}\left\{1-(1-\lambda\eta)^T\right\}\right\}. \tag{86}$$

By Eqs. (3) and (9),

$$0 \leq \lambda\eta < 1, \qquad \omega_T \leq \lambda(R_1 + \sqrt{2}\eta T) \leq \frac{\sqrt{1-\beta_2}}{8}. \tag{87}$$

D.1 Normalized Weights and Update Bounds

Lemma D.1 (Normalized geometric weights)**.** *Under the tuning and condition in Eqs.* (3) *and* (9)*, the following bounds hold:*

$$\max\left\{(1-\beta_2), \frac{1}{t}\right\} \leq \rho_t \leq (1-\beta_2) + \frac{1}{t}, \tag{88}$$

$$\sum_{t=1}^{T}\rho_t \leq T(1-\beta_2) + 2 + \log\min\left\{T, (1-\beta_2)^{-1}\right\}, \tag{89}$$

$$\sum_{t=1}^{T}\sqrt{\rho_t} \leq T\sqrt{1-\beta_2} + \frac{4}{\sqrt{1-\beta_2}}, \tag{90}$$

$$\sum_{s=1}^{t}\alpha_{t,s}^2 \leq \frac{16}{15}\rho_t, \qquad \sum_{s=1}^{t}\alpha_{t,s}^2(t-s) \leq 1, \tag{91}$$

$$\sum_{s=1}^{t}\gamma_{t,s}(t-s)^2 \leq 9\rho_t^{-2}, \qquad \sum_{s=1}^{t}\gamma_{t,s}^{p/2} \leq 2\rho_t^{p/2-1}. \tag{92}$$

Proof. The inequality $1-\beta_2^t \le t(1-\beta_2)$ gives $\rho_t \ge 1/t$, while $1-\beta_2^t \le 1$ gives $\rho_t \ge 1-\beta_2$. For the upper bound, Bernoulli's inequality gives

$$\beta_2^{-t} = \left(1 + \frac{1-\beta_2}{\beta_2}\right)^t \ge 1 + t\frac{1-\beta_2}{\beta_2} \ge 1 + t(1-\beta_2).$$

Consequently,

$$1-\beta_2^t \ge \frac{t(1-\beta_2)}{1+t(1-\beta_2)}, \qquad \rho_t \le (1-\beta_2) + \frac{1}{t},$$

which proves Eq. (88).

We next prove Eq. (89). If $T \le (1-\beta_2)^{-1}$, then

$$\sum_{t=1}^{T} [\rho_t - (1-\beta_2)] \le \sum_{t=1}^{T} \frac{1}{t} \le 1 + \log T.$$

If $T > (1-\beta_2)^{-1}$, put $m := \lfloor (1-\beta_2)^{-1} \rfloor$. Using $\rho_t - (1-\beta_2) \le 1/t$ from Eq. (88) and $\sum_{t=1}^{m} 1/t \le 1 + \log m$ gives

$$\sum_{t=1}^{m} [\rho_t - (1-\beta_2)] \le 1 + \log(1-\beta_2)^{-1}.$$

For $t \ge m+1$, we have $t(1-\beta_2) > 1$ and hence $\beta_2^t \le e^{-t(1-\beta_2)} < e^{-1}$. Therefore

$$\sum_{t=m+1}^{\infty} [\rho_t - (1-\beta_2)] = \sum_{t=m+1}^{\infty} \frac{(1-\beta_2)\beta_2^t}{1-\beta_2^t} \le \frac{1-\beta_2}{1-e^{-1}} \sum_{t=m+1}^{\infty} \beta_2^t = \frac{\beta_2^{m+1}}{1-e^{-1}} < \frac{1}{e-1} < 1.$$

The two cases prove Eq. (89).

For Eq. (90), use the same integer $m = \lfloor (1-\beta_2)^{-1} \rfloor$. For $t \le m$, Eq. (88) and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ give

$$\sqrt{\rho_t} - \sqrt{1-\beta_2} \le t^{-1/2}, \qquad \sum_{t=1}^{m} t^{-1/2} \le 2\sqrt{m} \le \frac{2}{\sqrt{1-\beta_2}}.$$

For $t \ge m+1$, put $z := \beta_2^t < e^{-1}$. On $0 \le z \le e^{-1}$, the inequality $(1-z)^{-1/2} - 1 \le z$ follows after squaring from $z(1-z-z^2) \ge 0$, because $e^{-1} < (\sqrt{5}-1)/2$. Hence

$$\sum_{t=m+1}^{\infty} \left(\sqrt{\rho_t} - \sqrt{1-\beta_2}\right) \le \sqrt{1-\beta_2} \sum_{t=m+1}^{\infty} \beta_2^t = \frac{\beta_2^{m+1}}{\sqrt{1-\beta_2}} \le \frac{1}{\sqrt{1-\beta_2}}.$$

Thus the excess over $T\sqrt{1-\beta_2}$ is at most $3/\sqrt{1-\beta_2}$, which implies Eq. (90).

For fixed t, define $f_t(a) := \frac{1-a}{1+a}\frac{1+a^t}{1-a^t}$. A direct summation gives $\sum_{s=1}^{t} \alpha_{t,s}^2 = f_t(\beta_1)$. Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}a} \log f_t(a) = -\frac{2}{1-a^2} + \frac{2ta^{t-1}}{1-a^{2t}} \le 0, \text{ since } \frac{1-a^{2t}}{1-a^2} = \sum_{k=0}^{t-1} a^{2k} \overset{\text{AM-GM inequality}}{\ge} ta^{t-1}.$$

Thus f_t is decreasing and its maximum over $\beta_2^2 \le \beta_1 \le \beta_2$ occurs at $\beta_1 = \beta_2^2$. It follows that

$$\begin{aligned} \frac{\sum_{s=1}^{t} \alpha_{t,s}^2}{\rho_t} &= \frac{1-\beta_2^t}{1-\beta_2} f_t(\beta_1) \le \frac{1-\beta_2^t}{1-\beta_2} f_t(\beta_2^2) \\ &= \frac{1+\beta_2}{1+\beta_2^2} \frac{1+\beta_2^{2t}}{1+\beta_2^t} \le \frac{1+\beta_2}{1+\beta_2^2}. \end{aligned}$$

By Eq. (3), $\beta_2 \ge 7/8$. Since $x \mapsto (1+x)/(1+x^2)$ is decreasing on $[7/8, 1]$,

$$\frac{1+\beta_2}{1+\beta_2^2} \le \frac{1+7/8}{1+(7/8)^2} = \frac{120}{113} < \frac{16}{15}.$$

This proves the first bound in Eq. (91). For the second,

$$\sum_{s=1}^{t} \alpha_{t,s}^2 (t-s) = \frac{\sum_{k=0}^{t-1} k\beta_1^{2k}}{\left(\sum_{k=0}^{t-1} \beta_1^k\right)^2} \le \frac{\sum_{k=0}^{t-1} k\beta_1^{k}}{\left(\sum_{k=0}^{t-1} \beta_1^k\right)^2} \le \frac{\sum_{k=0}^{t-1} k\beta_1^{k}}{\sum_{k=0}^{t-1}(k+1)\beta_1^k} \le 1.$$

The second inequality comes from unrolling the Cauchy convolution of $\left(\sum_{k=0}^{t-1} \beta_1^k\right)^2$ while discarding high-order ($\ge t$) terms of β_1.

For the first bound in Eq. (92), suppose first that $t(1-\beta_2) \le 1$. Then Eq. (88) gives $\rho_t \le 2/t$, and hence

$$\sum_{s=1}^{t} \gamma_{t,s}(t-s)^2 \le \sum_{s=1}^{t} \gamma_{t,s} t^2 = t^2 \le 4\rho_t^{-2}.$$

If $t(1-\beta_2) > 1$, then $1-\beta_2^t \ge 1-e^{-1}$ and

$$\sum_{s=1}^{t} \gamma_{t,s}(t-s)^2 = \frac{1-\beta_2}{1-\beta_2^t} \sum_{k=0}^{t-1} k^2 \beta_2^k \le \frac{1-\beta_2}{1-e^{-1}} \cdot \frac{\beta_2(1+\beta_2)}{(1-\beta_2)^3} \le \frac{2}{(1-e^{-1})(1-\beta_2)^2}.$$

Because $\frac{2}{1-e^{-1}} < 9(1-e^{-1})^2$, thus the preceding display is at most $9(1-\beta_2^t)^2/(1-\beta_2)^2 = 9\rho_t^{-2}$.

Finally, geometric summation gives

$$\sum_{s=1}^{t} \gamma_{t,s}^{p/2} = \rho_t^{p/2} \sum_{k=0}^{t-1} \beta_2^{pk/2} = \rho_t^{p/2} \frac{1-\beta_2^{pt/2}}{1-\beta_2^{p/2}} \le \rho_t^{p/2} \frac{1-\beta_2^t}{(1-\beta_2)/2} = 2\rho_t^{p/2-1},$$

where the first inequality follows from $1-\beta_2^{pt/2} \le 1-\beta_2^t$ and $1-\beta_2^{p/2} \ge p(1-\beta_2)/2 \ge (1-\beta_2)/2$. □

Lemma D.2 (Corrected bounded update and decay radius)**.** *For every t and block j,*

$$\left\|\widehat{m}_t^{(j)}\right\|^2 \le 2d_j \widehat{v}_t^{(j)}, \qquad \left\| \frac{\widehat{m}_t^{(j)}}{\sqrt{\widehat{v}_t^{(j)}} + \epsilon_{\mathrm{bc}}} \right\| \le \sqrt{2d_j}. \tag{93}$$

Consequently, if $\widehat{u}_t$ denotes the corrected adaptive direction, then $\|\widehat{u}_t\| \le \sqrt{2d}$. Under $0 \le \lambda\eta \le 1$, the radius in Eq. (86) *satisfies, pathwise,*

$$\lambda \frac{\left\|x_t^{(j)}\right\|}{\sqrt{d_j}} \le \omega_T, \qquad \lambda \|x_t\| \le \omega_T \sqrt{d}, \qquad t \le T+1. \tag{94}$$

Proof. The ratio of the normalized weights is

$$\frac{\alpha_{t,s}}{\gamma_{t,s}} = \frac{1-\beta_1}{1-\beta_2} \frac{1-\beta_2^t}{1-\beta_1^t} \left(\frac{\beta_1}{\beta_2}\right)^{t-s}.$$

The three factors are at most 2, 1, and 1, respectively. The first follows from $1-\beta_1 \le 2(1-\beta_2)$, and the other two follow from $\beta_1 \le \beta_2$. Hence $\alpha_{t,s}/\gamma_{t,s} \le 2$ and

$$\sum_{s=1}^{t} \frac{\alpha_{t,s}^2}{\gamma_{t,s}} = \sum_{s=1}^{t} \alpha_{t,s} \frac{\alpha_{t,s}}{\gamma_{t,s}} \le 2 \sum_{s=1}^{t} \alpha_{t,s} = 2.$$

Weighted Cauchy–Schwarz now gives

$$\left\|\widehat{m}_t^{(j)}\right\|^2 = \left\|\sum_{s=1}^{t} \alpha_{t,s} g_s^{(j)}\right\|^2 \le \left(\sum_{s=1}^{t} \frac{\alpha_{t,s}^2}{\gamma_{t,s}}\right) \left(\sum_{s=1}^{t} \gamma_{t,s} \left\|g_s^{(j)}\right\|^2\right) \le 2d_j \widehat{v}_t^{(j)},$$

which proves Eq. (93).

The update in Eq. (8) and the first bound in Eq. (93) imply

$$\lambda \frac{\left\|x_{t+1}^{(j)}\right\|}{\sqrt{d_j}} \le (1-\lambda\eta)\lambda \frac{\left\|x_t^{(j)}\right\|}{\sqrt{d_j}} + \sqrt{2}\lambda\eta.$$

Iterating this recursion yields

$$\lambda \frac{\left\|x_{t+1}^{(j)}\right\|}{\sqrt{d_j}} \le (1-\lambda\eta)^t \lambda R_1 + \sqrt{2}\left\{1-(1-\lambda\eta)^t\right\}.$$

The right-hand side is affine in $(1-\lambda\eta)^t$, which for $0 \le t \le T$ lies between its endpoint values at $t=0$ and $t=T$. The definition in Eq. (86) therefore gives the blockwise bound in Eq. (94). Summing its square over the blocks gives the Euclidean bound. □

D.2 Tail and Momentum Bounds with Bias Correction

Use the same core–tail split as in Eqs. (29) and (30). Its threshold is $\sigma_B(1-\beta_2)^{-1/p}$. For the remainder of this proof, put

$$\nu := \sigma_B(1-\beta_2)^{1-1/p}, \qquad \mathcal{G}_T := \sqrt{\frac{\Delta L}{T}}. \tag{95}$$

The tuning in Eq. (9) gives the exact identities

$$\mathcal{G}_T = \sqrt{1-\beta_2}\,\nu, \qquad L\eta\sqrt{d} = \mathcal{G}_T. \tag{96}$$

Since $(1-\beta_1) \ge (1-\beta_2)$, the corrected statistical floor also dominates the optimization scale:

$$\frac{L\eta}{(1-\beta_1)} = \frac{\mathcal{G}_T}{\sqrt{d}\,(1-\beta_1)} \le \frac{\nu}{\sqrt{d}\sqrt{1-\beta_2}} = \frac{\sigma_B}{\sqrt{d}}(1-\beta_2)^{1/2-1/p}, \quad \epsilon_{\mathrm{bc}} = 8\frac{\nu}{\sqrt{d}\sqrt{1-\beta_2}}. \tag{97}$$

By Lemma C.3, conditionally on $\mathcal{F}_{t-1}$,

$$\begin{aligned} &\mathbb{P}(I_t \mid \mathcal{F}_{t-1}) \le 1-\beta_2, \quad \mathbb{E}_{t-1}\left\|g_{t,\mathrm{tail}}\right\| \le \frac{3}{2}\nu, \\ &\left\|b_t^{\mathrm{tr}}\right\| \le \frac{3}{2}\nu, \qquad \mathbb{E}_{t-1}\left\|\zeta_t\right\|^2 \le \frac{\nu^2}{1-\beta_2} + \frac{5}{2}\nu\left\|\nabla_t\right\|. \end{aligned} \tag{98}$$

where $b_t^{\mathrm{tr}} := \nabla_t - \mathbb{E}_{t-1}g_{t,\mathrm{core}}$ and $\zeta_t := g_{t,\mathrm{core}} - \mathbb{E}_{t-1}g_{t,\mathrm{core}}$.

For $a \in \{\mathrm{core}, \mathrm{tail}\}$, define the normalized split buffers by

$$\widehat{m}_{t,a} := \sum_{s=1}^{t} \alpha_{t,s} g_{s,a}, \qquad \widehat{v}_{t,a}^{(j)} := \sum_{s=1}^{t} \gamma_{t,s} \frac{\left\|g_{s,a}^{(j)}\right\|^2}{d_j}.$$

The core and tail samples have disjoint supports, so pathwise

$$\widehat{m}_t = \widehat{m}_{t,\mathrm{core}} + \widehat{m}_{t,\mathrm{tail}}, \qquad \widehat{v}_t^{(j)} = \widehat{v}_{t,\mathrm{core}}^{(j)} + \widehat{v}_{t,\mathrm{tail}}^{(j)}.$$

Define

$$\widehat{D}_t^{(j)} := \sqrt{\widehat{v}_t^{(j)}} + \epsilon_{\mathrm{bc}}, \quad \widehat{A}_t := \sum_j \frac{\left\|\nabla_t^{(j)}\right\|^2}{\widehat{D}_t^{(j)}}, \quad \widehat{C}_t := \sum_j \frac{\left\|\widehat{m}_{t,\mathrm{core}}^{(j)}\right\|^2}{\widehat{D}_t^{(j)}}, \quad \widehat{\mathcal{T}}_t := \sum_j \frac{\left\|\widehat{m}_{t,\mathrm{tail}}^{(j)}\right\|^2}{\widehat{D}_t^{(j)}}. \tag{99}$$

Lemma D.3 (Corrected sparse-tail bound)**.** *For every $t \in [T]$,*

$$\mathbb{E}\left[\widehat{\mathcal{T}}_t\right] \le 6\sqrt{d}\,\nu\sqrt{\rho_t}, \qquad \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\widehat{\mathcal{T}}_t\right] \le 6\sqrt{d}\,\mathcal{G}_T\left(1+\frac{4}{T(1-\beta_2)}\right). \tag{100}$$

Proof. Fix t and a block j. If $\widehat{v}_{t,\mathrm{tail}}^{(j)} = 0$, then every tail sample in that block is zero and $\widehat{m}_{t,\mathrm{tail}}^{(j)} = 0$. Otherwise, the definitions give

$$\widehat{m}_{t,\mathrm{tail}}^{(j)} = \sum_{s=1}^{t} \alpha_{t,s} g_{s,\mathrm{tail}}^{(j)}, \qquad d_j \widehat{v}_{t,\mathrm{tail}}^{(j)} = \sum_{s=1}^{t} \gamma_{t,s} \left\| g_{s,\mathrm{tail}}^{(j)} \right\|^2.$$

Set

$$Q_t^{(j)} := \left(\sum_{\ell=1}^{t} \gamma_{t,\ell} \left\| g_{\ell,\mathrm{tail}}^{(j)} \right\|^2 \right)^{1/2}.$$

Since $\widehat{D}_t^{(j)} \geq \sqrt{\widehat{v}_{t,\mathrm{tail}}^{(j)}} = Q_t^{(j)} / \sqrt{d_j}$, we have, block by block,

$$\begin{aligned} \frac{\left\| \widehat{m}_{t,\mathrm{tail}}^{(j)} \right\|^2}{\widehat{D}_t^{(j)}} &\leq \sqrt{d_j} \, \frac{\left\| \sum_{s=1}^{t} \alpha_{t,s} g_{s,\mathrm{tail}}^{(j)} \right\|^2}{Q_t^{(j)}} \\ &= \sqrt{d_j} \sum_{s,k \leq t} \alpha_{t,s} \alpha_{t,k} \frac{\left\langle g_{s,\mathrm{tail}}^{(j)}, g_{k,\mathrm{tail}}^{(j)} \right\rangle}{Q_t^{(j)}} \leq \sqrt{d_j} \sum_{s,k \leq t} \alpha_{t,s} \alpha_{t,k} \frac{\left\| g_{s,\mathrm{tail}}^{(j)} \right\| \left\| g_{k,\mathrm{tail}}^{(j)} \right\|}{Q_t^{(j)}}. \end{aligned}$$

If I_s occurs, then $Q_t^{(j)} \geq \sqrt{\gamma_{t,s}} \left\| g_{s,\mathrm{tail}}^{(j)} \right\|$. If it does not occur, the corresponding numerator is zero. Hence $\frac{\left\| g_{s,\mathrm{tail}}^{(j)} \right\|}{Q_t^{(j)}} \leq \frac{\mathbb{I}_{I_s}}{\sqrt{\gamma_{t,s}}}$, with the $0/0$ case interpreted as zero. Substituting this blockwise bound and then summing over j gives

$$\widehat{\mathcal{T}}_t \leq \sum_{s,k \leq t} \frac{\alpha_{t,s} \alpha_{t,k}}{\sqrt{\gamma_{t,s}}} \mathbb{I}_{I_s} \sum_{j=1}^{J} \sqrt{d_j} \left\| g_{k,\mathrm{tail}}^{(j)} \right\| \leq \sqrt{d} \sum_{s,k \leq t} \frac{\alpha_{t,s} \alpha_{t,k}}{\sqrt{\gamma_{t,s}}} \mathbb{I}_{I_s} \left\| g_{k,\mathrm{tail}} \right\|, \tag{101}$$

where the last line follows from Cauchy–Schwarz:

$$\sum_{j=1}^{J} \sqrt{d_j} \left\| g_{k,\mathrm{tail}}^{(j)} \right\| \leq \left(\sum_{j=1}^{J} d_j \right)^{1/2} \left(\sum_{j=1}^{J} \left\| g_{k,\mathrm{tail}}^{(j)} \right\|^2 \right)^{1/2} = \sqrt{d} \left\| g_{k,\mathrm{tail}} \right\|.$$

We next evaluate the two deterministic weight sums. Since $\beta_1^2 / \sqrt{\beta_2} \leq \beta_1$,

$$\frac{1}{\sqrt{\rho_t}} \sum_{s=1}^{t} \frac{\alpha_{t,s}^2}{\sqrt{\gamma_{t,s}}} = \frac{(1-\beta_1)^2}{(1-\beta_1^t)^2 \rho_t} \sum_{k=0}^{t-1} \left(\frac{\beta_1^2}{\sqrt{\beta_2}} \right)^k \leq \frac{(1-\beta_1)^2}{(1-\beta_1^t)^2 \rho_t} \frac{1-\beta_1^t}{1-\beta_1} = \frac{1-\beta_1}{1-\beta_2} \frac{1-\beta_2^t}{1-\beta_1^t} \leq 2.$$

Similarly, $\beta_1 / \sqrt{\beta_2} \leq \sqrt{\beta_1}$ gives

$$\sqrt{\rho_t} \sum_{s=1}^{t} \frac{\alpha_{t,s}}{\sqrt{\gamma_{t,s}}} = \frac{1-\beta_1}{1-\beta_1^t} \sum_{k=0}^{t-1} \left(\frac{\beta_1}{\sqrt{\beta_2}} \right)^k \leq \frac{1-\beta_1}{1-\beta_1^t} \frac{1-\beta_1^{t/2}}{1-\sqrt{\beta_1}} = \frac{1+\sqrt{\beta_1}}{1+\beta_1^{t/2}} \leq 2.$$

Thus

$$\sum_s \frac{\alpha_{t,s}^2}{\sqrt{\gamma_{t,s}}} \leq 2\sqrt{\rho_t}, \qquad \sum_s \frac{\alpha_{t,s}}{\sqrt{\gamma_{t,s}}} \leq \frac{2}{\sqrt{\rho_t}}.$$

We split Eq. (101) into diagonal and off-diagonal parts, following the same two-time-order argument as in Lemma C.4. For the diagonal part, the first mixed-weight bound and Eq. (98) give

$$\sqrt{d} \sum_{s=1}^{t} \frac{\alpha_{t,s}^2}{\sqrt{\gamma_{t,s}}} \mathbb{E} \left\| g_{s,\mathrm{tail}} \right\| \leq \sqrt{d} \, (2\sqrt{\rho_t}) \left(\frac{3}{2} \nu \right) = 3\sqrt{d} \, \nu \sqrt{\rho_t}.$$

For the off-diagonal terms, consider both time orders. If $s < k$, then I_s is $\mathcal{F}_{k-1}$-measurable, so

$$\mathbb{E} \left[\mathbb{I}_{I_s} \left\| g_{k,\mathrm{tail}} \right\| \right] = \mathbb{E} \left[\mathbb{I}_{I_s} \mathbb{E}_{k-1} \left\| g_{k,\mathrm{tail}} \right\| \right] \leq \frac{3}{2} \nu \mathbb{E} \mathbb{I}_{I_s} \leq \frac{3}{2} \nu (1-\beta_2).$$

If $k < s$, then $\|g_{k,\mathrm{tail}}\|$ is $\mathcal{F}_{s-1}$-measurable, and hence

$$\mathbb{E}\left[\mathbb{I}_{I_s}\|g_{k,\mathrm{tail}}\|\right] = \mathbb{E}\left[\|g_{k,\mathrm{tail}}\|\,\mathbb{P}(I_s \mid \mathcal{F}_{s-1})\right] \le (1-\beta_2)\mathbb{E}\|g_{k,\mathrm{tail}}\| \le \frac{3}{2}\nu(1-\beta_2).$$

Since $\sum_k \alpha_{t,k} = 1$, the total off-diagonal contribution is at most

$$\frac{3}{2}\sqrt{d}\,\nu(1-\beta_2)\left(\sum_s \frac{\alpha_{t,s}}{\sqrt{\gamma_{t,s}}}\right)\left(\sum_k \alpha_{t,k}\right) \le 3\sqrt{d}\,\nu\frac{1-\beta_2}{\sqrt{\rho_t}} \le 3\sqrt{d}\,\nu\sqrt{\rho_t}.$$

Adding the diagonal and off-diagonal parts proves the pointwise bound. Finally, Eq. (90) and $\mathcal{G}_T = \sqrt{1-\beta_2}\nu$ imply

$$\frac{6\sqrt{d}\,\nu}{T}\sum_{t=1}^{T}\sqrt{\rho_t} \le 6\sqrt{d}\,\mathcal{G}_T\left(1 + \frac{4}{T(1-\beta_2)}\right),$$

which proves the averaged bound. □

Define the corrected core tracking error

$$\widehat{e}_t := \widehat{m}_{t,\mathrm{core}} - \nabla_t = \mathcal{L}_t - \sum_{s=1}^{t}\alpha_{t,s}b_s^{\mathrm{tr}} + \mathcal{M}_t, \tag{102}$$

where

$$\mathcal{L}_t := \sum_{s=1}^{t}\alpha_{t,s}(\nabla_s - \nabla_t), \qquad \mathcal{M}_t := \sum_{s=1}^{t}\alpha_{t,s}\zeta_s,$$

and ζ_s is defined in Eq. (56). The first term $\mathcal{L}_t$ is a deterministic signal tracking error, and the second term $\mathcal{M}_t$ is a martingale tracking error.

Lemma D.4 (Corrected momentum tracking)**.** *Under Eq.* (3)*,*

$$\mathbb{E}\|\mathcal{M}_t\|^2 \le \frac{16}{15}\nu^2/(1-\beta_2)\rho_t + \frac{8}{3}\nu\rho_t\mathbb{E}\|\nabla_t\| + \frac{5}{2}\nu L\eta(\sqrt{2}+\omega_T)\sqrt{d}, \tag{103}$$

$$\sum_{t=1}^{T}\mathbb{E}\|\mathcal{L}_t\|^2 \le \frac{L^2\eta^2}{(1-\beta_1)^2}\sum_{t=1}^{T-1}\mathbb{E}\|\widehat{u}_t + \lambda x_t\|^2. \tag{104}$$

Moreover,

$$\|\widehat{u}_t + \lambda x_t\|^2 \le \frac{4}{\epsilon_{\mathrm{bc}}}(\widehat{C}_t + \widehat{\mathcal{T}}_t) + 2\omega_T^2 d. \tag{105}$$

Proof. For $r < s$, the vector ζ_r is $\mathcal{F}_{s-1}$-measurable and $\mathbb{E}_{s-1}\zeta_s = 0$. Therefore

$$\mathbb{E}\langle\zeta_r, \zeta_s\rangle = \mathbb{E}\langle\zeta_r, \mathbb{E}_{s-1}\zeta_s\rangle = 0.$$

Expanding the squared norm of $\mathcal{M}_t$ and using Eq. (98) consequently gives

$$\mathbb{E}\|\mathcal{M}_t\|^2 = \sum_{s=1}^{t}\alpha_{t,s}^2\mathbb{E}\|\zeta_s\|^2 \le \frac{\nu^2}{1-\beta_2}\sum_s\alpha_{t,s}^2 + \frac{5}{2}\nu\sum_s\alpha_{t,s}^2\mathbb{E}\|\nabla_s\|.$$

By Lemma D.2 and Eq. (94), $\|\widehat{u}_r\| \le \sqrt{2d}$ and $\lambda\|x_r\| \le \omega_T\sqrt{d}$. Hence $\|\widehat{u}_r + \lambda x_r\| \le (\sqrt{2} + \omega_T)\sqrt{d}$ and, for $s \le t$,

$$\|\nabla_s - \nabla_t\| \le L\sum_{r=s}^{t-1}\|x_{r+1} - x_r\| \le L\eta(\sqrt{2}+\omega_T)\sqrt{d}\,(t-s). \tag{106}$$

and therefore

$$\|\nabla_s\| \le \|\nabla_t\| + L\eta(\sqrt{2}+\omega_T)\sqrt{d}\,(t-s).$$

Applying both bounds in Eq. (91) proves Eq. (103).

Expanding the gradient differences in $\mathcal{L}_t$ gives

$$\mathcal{L}_t = -\sum_{k=1}^{t-1} \frac{\beta_1^{t-k}(1-\beta_1^k)}{1-\beta_1^t}(\nabla_{k+1} - \nabla_k).$$

The coefficients are nonnegative and bounded by β_1^{t-k}. Weighted Cauchy–Schwarz therefore gives

$$\|\mathcal{L}_t\|^2 \le \frac{1}{1-\beta_1} \sum_{k=1}^{t-1} \beta_1^{t-k} \|\nabla_{k+1} - \nabla_k\|^2 .$$

Taking expectations, summing over t, and interchanging the summation order,

$$\sum_{t=1}^{T} \mathbb{E}\|\mathcal{L}_t\|^2 \le \frac{1}{1-\beta_1} \sum_{k=1}^{T-1} \mathbb{E}\|\nabla_{k+1} - \nabla_k\|^2 \sum_{t=k+1}^{T} \beta_1^{t-k} \le \frac{1}{(1-\beta_1)^2} \sum_{k=1}^{T-1} \mathbb{E}\|\nabla_{k+1} - \nabla_k\|^2 .$$

By Assumption 2 and the update of Algorithm 2,

$$\|\nabla_{k+1} - \nabla_k\|^2 \le L^2 \|x_{k+1} - x_k\|^2 = L^2\eta^2 \|\widehat{u}_k + \lambda x_k\|^2 .$$

Substituting this estimate into the preceding summed inequality proves Eq. (104).

Finally, the split identity for $\widehat{m}_t$ and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ imply $\|\widehat{u}_t\|^2 \le \frac{2}{\epsilon_{\mathrm{bc}}}(\widehat{C}_t + \widehat{\mathcal{T}}_t)$. Together with $\lambda^2\|x_t\|^2 \le \omega_T^2 d$ from Eq. (94), this yields

$$\|\widehat{u}_t + \lambda x_t\|^2 \le 2\|\widehat{u}_t\|^2 + 2\lambda^2\|x_t\|^2 \le \frac{4}{\epsilon_{\mathrm{bc}}}(\widehat{C}_t + \widehat{\mathcal{T}}_t) + 2\omega_T^2 d,$$

which is Eq. (105). □

D.3 Descent and Denominator Bounds with Bias Correction

Lemma D.5 (Corrected progress)**.** *Under the conditions of Theorem 3, let*

$$H_T := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla_t)\right].$$

Then

$$\frac{1}{\sqrt{d}T}\sum_{t=1}^{T} \mathbb{E}\left[\widehat{A}_t\right] \le 60\mathcal{G}_T(1+\mathfrak{r}_T) + \left(4\sqrt{d/d_{\min}}(1-\beta_2)^{3/2} + 4\omega_T\right) H_T. \tag{107}$$

Proof. For each block, polarization gives

$$\left\langle \nabla_t^{(j)}, \widehat{m}_{t,\mathrm{core}}^{(j)} \right\rangle = \frac{1}{2}\left\|\nabla_t^{(j)}\right\|^2 + \frac{1}{2}\left\|\widehat{m}_{t,\mathrm{core}}^{(j)}\right\|^2 - \frac{1}{2}\left\|\widehat{e}_t^{(j)}\right\|^2,$$

while Young's inequality gives

$$\left\langle \nabla_t^{(j)}, \widehat{m}_{t,\mathrm{tail}}^{(j)} \right\rangle \ge -\frac{1}{4}\left\|\nabla_t^{(j)}\right\|^2 - \left\|\widehat{m}_{t,\mathrm{tail}}^{(j)}\right\|^2 .$$

Adding, dividing by $\widehat{D}_t^{(j)}$, summing over blocks, and using $\widehat{D}_t^{(j)} \ge \epsilon_{\mathrm{bc}}$ gives the pathwise bound

$$\langle \nabla_t, \widehat{u}_t \rangle \ge \frac{1}{4}\widehat{A}_t + \frac{1}{2}\widehat{C}_t - \frac{1}{2\epsilon_{\mathrm{bc}}}\|\widehat{e}_t\|^2 - \widehat{\mathcal{T}}_t.$$

Moreover, the blockwise bound in Eq. (94) gives

$$\lambda\|x_t\|_{\mathcal{P}} = \sqrt{d}\max_j \frac{\lambda\left\|x_t^{(j)}\right\|}{\sqrt{d_j}} \le \omega_T\sqrt{d}.$$

Applying the exact duality inequality in Lemma 1 with $h = \nabla_t$ and $u = \lambda x_t$ then gives

$$\lambda\,|\langle \nabla_t, x_t\rangle| \le \mathsf{S}_{\mathcal{P}}(\nabla_t)\,\|\lambda x_t\|_{\mathcal{P}} = \lambda\mathsf{S}_{\mathcal{P}}(\nabla_t)\,\|x_t\|_{\mathcal{P}} \le \omega_T\sqrt{d}\,\mathsf{S}_{\mathcal{P}}(\nabla_t).$$

Using these two bounds and Eq. (105) in the smoothness inequality yields

$$F(x_{t+1}) \le F(x_t) - \frac{\eta}{4}\widehat{A}_t - \frac{\eta}{2}\widehat{C}_t + \frac{\eta}{2\epsilon_{\mathrm{bc}}}\left\|\widehat{e}_t\right\|^2 + \eta\widehat{\mathcal{T}}_t + \eta\omega_T\sqrt{d}\,\mathsf{S}_{\mathcal{P}}(\nabla_t) + \frac{2L\eta^2}{\epsilon_{\mathrm{bc}}}(\widehat{C}_t + \widehat{\mathcal{T}}_t) + L\eta^2\omega_T^2 d. \tag{108}$$

Since $\sum_{s=1}^{t}\alpha_{t,s} = 1$, the truncation-bias bound in Eq. (98) gives,

$$\left\|\sum_{s=1}^{t}\alpha_{t,s}b_s^{\mathrm{tr}}\right\| \le \sum_{s=1}^{t}\alpha_{t,s}\left\|b_s^{\mathrm{tr}}\right\| \le \frac{3}{2}\nu.$$

Let $\mathcal{B}_t := \sum_{s=1}^{t}\alpha_{t,s}b_s^{\mathrm{tr}}$. The preceding display gives $\|\mathcal{B}_t\| \le 3\nu/2$, and Eq. (102) reads $\widehat{e}_t = \mathcal{L}_t - \mathcal{B}_t + \mathcal{M}_t$. Using twice the elementary inequality $\|a+b+c\|^2 \le 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, we obtain

$$\left\|\widehat{e}_t\right\|^2 = \left\|\mathcal{L}_t - \mathcal{B}_t + \mathcal{M}_t\right\|^2 \le 3\left\|\mathcal{L}_t\right\|^2 + 3\left\|\mathcal{B}_t\right\|^2 + 3\left\|\mathcal{M}_t\right\|^2 \le 3\left\|\mathcal{L}_t\right\|^2 + \frac{27}{4}\nu^2 + 3\left\|\mathcal{M}_t\right\|^2.$$

Combining Lemma D.4 and Eq. (98) and summing gives

$$\sum_{t=1}^{T}\mathbb{E}\left\|\widehat{e}_t\right\|^2 \le \frac{12L^2\eta^2}{\epsilon_{\mathrm{bc}}(1-\beta_1)^2}\sum_t\mathbb{E}(\widehat{C}_t + \widehat{\mathcal{T}}_t) + \frac{6L^2\eta^2T\omega_T^2 d}{(1-\beta_1)^2} + \frac{27}{4}T\nu^2 + \frac{16\nu^2}{5(1-\beta_2)}\sum_t\rho_t + 8\nu\sum_t\rho_t\mathbb{E}\left\|\nabla_t\right\| + \frac{15}{2}T\nu L\eta(\sqrt{2}+\omega_T)\sqrt{d}. \tag{109}$$

By Eq. (97), $\epsilon_{\mathrm{bc}} \ge 8L\eta/(1-\beta_1)$. In addition, the parameter restriction $\beta_2^2 \le \beta_1 \le \beta_2$ in Eq. (9) gives $1-\beta_1 \le 1-\beta_2^2 = (1-\beta_2)(1+\beta_2) \le 2(1-\beta_2)$. Eq. (3) and $d_{\min} \le d$ give $1-\beta_2 \le 1/8$, so $1-\beta_1 \le 1/4$. Therefore

$$\frac{2L\eta^2}{\epsilon_{\mathrm{bc}}} \le \frac{\eta(1-\beta_1)}{4} \le \frac{\eta}{16}, \qquad \frac{6L^2\eta^3}{\epsilon_{\mathrm{bc}}^2(1-\beta_1)^2} \le \frac{3\eta}{32}.$$

Hence the total coefficient of $\sum_t\widehat{C}_t$ after inserting Eq. (109) into Eq. (108) is at most $-\eta/2 + \eta/16 + 3\eta/32 = -11\eta/32 < 0$. The corresponding positive coefficient of $\widehat{\mathcal{T}}_t$ is at most $\eta(1 + 1/16 + 3/32) = 37\eta/32$. Discarding the nonnegative core term and telescoping the function values gives the complete bookkeeping inequality

$$\frac{\eta}{4}\sum_{t=1}^{T}\mathbb{E}\widehat{A}_t \le \Delta + \frac{37\eta}{32}\sum_{t=1}^{T}\mathbb{E}\widehat{\mathcal{T}}_t + \frac{27\eta T\nu^2}{8\epsilon_{\mathrm{bc}}} + \frac{8\eta\nu^2}{5\epsilon_{\mathrm{bc}}(1-\beta_2)}\sum_{t=1}^{T}\rho_t + \frac{4\eta\nu}{\epsilon_{\mathrm{bc}}}\sum_{t=1}^{T}\rho_t\mathbb{E}\left\|\nabla_t\right\| + \frac{15\eta T\nu L\eta(\sqrt{2}+\omega_T)\sqrt{d}}{4\epsilon_{\mathrm{bc}}} + \frac{3\eta L^2\eta^2T\omega_T^2 d}{\epsilon_{\mathrm{bc}}(1-\beta_1)^2} + \eta\omega_T\sqrt{d}\sum_{t=1}^{T}\mathbb{E}\mathsf{S}_{\mathcal{P}}(\nabla_t) + L\eta^2T\omega_T^2 d. \tag{110}$$

We bound the terms on the right-hand side of Eq. (110) in their displayed order. Throughout this calculation,

$$\epsilon_{\mathrm{bc}}\sqrt{d} = 8\frac{\nu}{\sqrt{1-\beta_2}}, \qquad \frac{\nu}{\epsilon_{\mathrm{bc}}\sqrt{d}} = \frac{\sqrt{1-\beta_2}}{8}. \tag{111}$$

The first term is the objective gap. After division by $\eta T\sqrt{d}/4$, it is $\frac{4\Delta}{\eta T\sqrt{d}} = 4\mathcal{G}_T$. The second term is the corrected sparse-tail contribution. By Eq. (100),

$$\frac{4}{\eta T\sqrt{d}}\cdot\frac{37\eta}{32}\sum_{t=1}^{T}\mathbb{E}\widehat{\mathcal{T}}_t = \frac{37}{8\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\widehat{\mathcal{T}}_t \le \frac{111}{4}\mathcal{G}_T\left(1 + \frac{4}{T(1-\beta_2)}\right).$$

The third term comes from the truncation bias in Eq. (109). Its normalized value is

$$\frac{4}{\eta T\sqrt{d}}\cdot\frac{27\eta T\nu^2}{8\epsilon_{\mathrm{bc}}} = \frac{27\nu^2}{2\epsilon_{\mathrm{bc}}\sqrt{d}} = \frac{27}{16}\mathcal{G}_T,$$

where the last equality uses Eqs. (96) and (111).

The fourth term is the pure martingale-variance contribution. Using Eqs. (89) and (111),

$$\frac{4}{\eta T\sqrt{d}}\cdot\frac{8\eta\nu^2}{5\epsilon_{\rm bc}(1-\beta_2)}\sum_{t=1}^{T}\rho_t = \frac{32\nu^2}{5\epsilon_{\rm bc}T\sqrt{d}(1-\beta_2)}\sum_{t=1}^{T}\rho_t \le \frac{4}{5}\mathcal{G}_T\left(1+\frac{2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}}{T(1-\beta_2)}\right).$$

The fifth term is the gradient-dependent martingale variance

$$\frac{4}{\eta T\sqrt{d}}\cdot\frac{4\eta\nu}{\epsilon_{\rm bc}}\sum_{t=1}^{T}\rho_t\mathbb{E}\left\|\nabla_t\right\| = 2\sqrt{1-\beta_2}\,\frac{1}{T}\sum_{t=1}^{T}\rho_t\mathbb{E}\left\|\nabla_t\right\|.$$

We split this average only while bounding this fifth term. Put $n_0 := \lceil(1-\beta_2)^{-1}\rceil$ and $m := \min\{T, n_0-1\}$. For $t\le m$, Eq. (106) and $\|\nabla_1\|^2 \le 2L\Delta$ give

$$\|\nabla_t\| \le \sqrt{2L\Delta} + L\eta(\sqrt{2}+\omega_T)\sqrt{d}\,(t-1).$$

Moreover, Eq. (88) gives

$$\sum_{t=1}^{m}\rho_t \le 2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}, \qquad \sum_{t=1}^{m}t\rho_t \le \frac{2}{1-\beta_2}.$$

Therefore the early indices contribute at most

$$\frac{2\sqrt{1-\beta_2}}{T}\sum_{t<n_0}\rho_t\mathbb{E}\left\|\nabla_t\right\| \le 2\sqrt{2}\mathcal{G}_T\sqrt{\frac{1-\beta_2}{T}}\left(2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}\right) + \frac{4(\sqrt{2}+\omega_T)\mathcal{G}_T}{T\sqrt{1-\beta_2}}.$$

For $t\ge n_0$, we have $\rho_t \le 2(1-\beta_2)$. If $n_0 > T$, this sum is empty. Hence the late indices contribute at most

$$2\sqrt{1-\beta_2}\,\frac{1}{T}\sum_{t\ge n_0}\rho_t\mathbb{E}\left\|\nabla_t\right\| \le 4\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}H_T.$$

The sixth term is the lag contribution from Eq. (103). Its normalized value is exactly

$$\frac{4}{\eta T\sqrt{d}}\cdot\frac{15\eta T\nu L\eta(\sqrt{2}+\omega_T)\sqrt{d}}{4\epsilon_{\rm bc}} = \frac{15}{8}\mathcal{G}_T\sqrt{1-\beta_2}(\sqrt{2}+\omega_T).$$

The seventh term is the squared-decay contribution from signal tracking. Using Eq. (97) and $1-\beta_1 \ge 1-\beta_2$,

$$\frac{4}{\eta T\sqrt{d}}\cdot\frac{3\eta L^2\eta^2 T\omega_T^2 d}{\epsilon_{\rm bc}(1-\beta_1)^2} = \frac{12L^2\eta^2\omega_T^2\sqrt{d}}{\epsilon_{\rm bc}(1-\beta_1)^2} \le \frac{3}{2}\mathcal{G}_T\frac{\omega_T^2}{1-\beta_2}.$$

The eighth term is the linear weight-decay contribution, which becomes

$$\frac{4}{\eta T\sqrt{d}}\cdot\eta\omega_T\sqrt{d}\sum_{t=1}^{T}\mathbb{E}\mathsf{S}_{\mathcal{P}}(\nabla_t) = 4\omega_T H_T.$$

The ninth and final term is the direct squared-decay contribution:

$$\frac{4}{\eta T\sqrt{d}}\cdot L\eta^2 T\omega_T^2 d = 4\mathcal{G}_T\omega_T^2.$$

Combining the displayed estimates gives

$$\begin{aligned}\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\widehat{A}_t \le{}& \left\{4+\frac{111}{4}+\frac{27}{16}+\frac{4}{5}+\frac{15}{8}\sqrt{1-\beta_2}(\sqrt{2}+\omega_T)+\frac{3}{2}\frac{\omega_T^2}{1-\beta_2}+4\omega_T^2\right\}\mathcal{G}_T\\ &+\left\{\frac{111}{T(1-\beta_2)}+\frac{4}{5}\frac{2+\log\min\{T,(1-\beta_2)^{-1}\}}{T(1-\beta_2)}+\frac{4(\sqrt{2}+\omega_T)}{T\sqrt{1-\beta_2}}\right\}\mathcal{G}_T\\ &+2\sqrt{2}\mathcal{G}_T\sqrt{\frac{1-\beta_2}{T}}\left(2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}\right)+\left\{4\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}+4\omega_T\right\}H_T.\end{aligned}$$

By Eqs. (3) and (87),

$$\frac{15}{8}\sqrt{1-\beta_2}(\sqrt{2}+\omega_T)<1, \qquad \frac{3}{2}\frac{\omega_T^2}{1-\beta_2}+4\omega_T^2\le\frac{1}{32}.$$

Consequently,

$$4+\frac{111}{4}+\frac{27}{16}+\frac{4}{5}+1+\frac{1}{32}<36,$$

$$111+\frac{4}{5}\left(2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}\right)+4(\sqrt{2}+\omega_T)\le 60\left(2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}\right),$$

and

$$2\sqrt{2}\left(2+\log\min\left\{T,(1-\beta_2)^{-1}\right\}\right)\le 6\left(1+\log\min\left\{T,(1-\beta_2)^{-1}\right\}\right).$$

Substitution into the preceding complete bound and the definition of $\mathfrak{r}_T$ prove

$$\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[\widehat{A}_t\right]\le 60\mathcal{G}_T(1+\mathfrak{r}_T)+\left(4\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}+4\omega_T\right)H_T,$$

which is Eq. (107). □

Lemma D.6 (Corrected denominator mass)**.** *Under the conditions of Theorem 3,*

$$\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{J}d_j\widehat{D}_t^{(j)}\right]\le H_T+15\frac{\nu}{\sqrt{1-\beta_2}}. \tag{112}$$

Proof. We follow the signal–noise–floor decomposition in Lemma C.7, with the normalized weights $\gamma_{t,s}$. For every block define

$$\widehat{v}_{t,\mathrm{sig}}^{(j)}:=\sum_{s=1}^{t}\gamma_{t,s}\frac{\left\|\nabla_s^{(j)}\right\|^2}{d_j}, \qquad \widehat{v}_{t,\mathrm{noise}}^{(j)}:=\sum_{s=1}^{t}\gamma_{t,s}\frac{\left\|\xi_s^{(j)}\right\|^2}{d_j}.$$

Since $g_s^{(j)}=\nabla_s^{(j)}+\xi_s^{(j)}$, Minkowski's inequality in the weighted direct-sum space gives

$$\sqrt{\widehat{v}_t^{(j)}}\le\sqrt{\widehat{v}_{t,\mathrm{sig}}^{(j)}}+\sqrt{\widehat{v}_{t,\mathrm{noise}}^{(j)}}.$$

Consequently,

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\widehat{D}_t^{(j)}\le\underbrace{\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{\widehat{v}_{t,\mathrm{sig}}^{(j)}}}_{\text{signal mass}}+\underbrace{\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{\widehat{v}_{t,\mathrm{noise}}^{(j)}}}_{\text{noise mass}}+\underbrace{\sqrt{d}\,\epsilon_{\mathrm{bc}}}_{\text{floor mass}}. \tag{113}$$

We bound the three terms in this order.

Signal contribution. For each $s\le t$, write $\nabla_s^{(j)}=\nabla_t^{(j)}+(\nabla_s^{(j)}-\nabla_t^{(j)})$. Because $\sum_{s=1}^{t}\gamma_{t,s}=1$, Minkowski's inequality gives

$$\sqrt{\widehat{v}_{t,\mathrm{sig}}^{(j)}}\le\frac{\left\|\nabla_t^{(j)}\right\|}{\sqrt{d_j}}+\frac{1}{\sqrt{d_j}}\sqrt{\sum_{s=1}^{t}\gamma_{t,s}\left\|\nabla_s^{(j)}-\nabla_t^{(j)}\right\|^2}.$$

Multiplying by $d_j/\sqrt{d}$, summing over j, and applying Cauchy–Schwarz across the blocks yields

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{\widehat{v}_{t,\mathrm{sig}}^{(j)}}\le\mathsf{S}_{\mathcal{P}}(\nabla_t)+\sqrt{\sum_{s=1}^{t}\gamma_{t,s}\left\|\nabla_s-\nabla_t\right\|^2}. \tag{114}$$

The path estimate Eq. (106) gives

$$\|\nabla_s-\nabla_t\|\le L\eta(\sqrt{2}+\omega_T)\sqrt{d}\,(t-s).$$

Substituting this bound and then using the first inequality in Eq. (92) gives the complete drift estimate

$$\sqrt{\sum_{s=1}^{t}\gamma_{t,s}\left\|\nabla_s-\nabla_t\right\|^2}\le L\eta(\sqrt{2}+\omega_T)\sqrt{d}\sqrt{\sum_{s=1}^{t}\gamma_{t,s}(t-s)^2}$$
$$\le\frac{3L\eta(\sqrt{2}+\omega_T)\sqrt{d}}{\rho_t}\le\frac{5\mathcal{G}_T}{1-\beta_2}=\frac{5\nu}{\sqrt{1-\beta_2}}.$$

Here the second line uses $\sum_s\gamma_{t,s}(t-s)^2\le 9\rho_t^{-2}$, and the last line uses Eqs. (3) and (87), $\rho_t\ge 1-\beta_2$, and $L\eta\sqrt{d}=\mathcal{G}_T=\sqrt{1-\beta_2}\,\nu$. Combining with Eq. (114),

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{\widehat{v}_{t,\mathrm{sig}}^{(j)}}\le\mathsf{S}_{\mathcal{P}}(\nabla_t)+\frac{5\nu}{\sqrt{1-\beta_2}}.\tag{115}$$

Noise contribution. Cauchy–Schwarz over the blocks gives

$$\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\sqrt{\widehat{v}_{t,\mathrm{noise}}^{(j)}}=\frac{1}{\sqrt{d}}\sum_{j=1}^{J}\sqrt{d_j}\sqrt{\sum_{s=1}^{t}\gamma_{t,s}\left\|\xi_s^{(j)}\right\|^2}\le\sqrt{\sum_{s=1}^{t}\gamma_{t,s}\left\|\xi_s\right\|^2}.$$

Since $p/2\le 1$, Lyapunov's inequality and subadditivity imply

$$\mathbb{E}\sqrt{\sum_{s=1}^{t}\gamma_{t,s}\left\|\xi_s\right\|^2}\le\left(\mathbb{E}\left(\sum_{s=1}^{t}\gamma_{t,s}\left\|\xi_s\right\|^2\right)^{p/2}\right)^{1/p}\le\left(\sum_{s=1}^{t}\gamma_{t,s}^{p/2}\mathbb{E}\left\|\xi_s\right\|^p\right)^{1/p}$$
$$\le 2^{1/p}\sigma_B\rho_t^{1/2-1/p}\le\frac{2\nu}{\sqrt{1-\beta_2}},$$

where the last line uses $1/2-1/p\le 0$, $\rho_t\ge 1-\beta_2$, and $2^{1/p}\le 2$.

Stabilizer contribution. By Eq. (97), $\sqrt{d}\,\epsilon_{\mathrm{bc}}=\frac{8\nu}{\sqrt{1-\beta_2}}$. Taking expectations in Eq. (113), averaging over t, and using the signal, noise, and floor bounds gives

$$\frac{1}{\sqrt{d}T}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{j=1}^{J}d_j\widehat{D}_t^{(j)}\right]\le H_T+(5+2+8)\frac{\nu}{\sqrt{1-\beta_2}},$$

which is Eq. (112). □

D.4 Proof of Theorem 3

Proof of Theorem 3. Pointwise Cauchy–Schwarz in the natural geometry gives

$$\mathsf{S}_{\mathcal{P}}(\nabla_t)^2\le\left(\frac{\widehat{A}_t}{\sqrt{d}}\right)\left(\frac{1}{\sqrt{d}}\sum_{j=1}^{J}d_j\widehat{D}_t^{(j)}\right).$$

Cauchy–Schwarz over time and probability, followed by Lemmas D.5 and D.6, yields

$$H_T^2\le\left(60\mathcal{G}_T(1+\mathfrak{r}_T)+\left(4\sqrt{d/d_{\min}}(1-\beta_2)^{3/2}+4\omega_T\right)H_T\right)\left(H_T+15\frac{\nu}{\sqrt{1-\beta_2}}\right).\tag{116}$$

Put

$$s:=\sqrt{1-\beta_2},\qquad Y:=\frac{H_T}{\nu}.$$

By Eqs. (3) and (87),

$$4\sqrt{\frac{d}{d_{\min}}}(1-\beta_2)^{3/2}+4\omega_T\le\frac{s}{2}+\frac{s}{2}=s.$$

Since $\mathcal{G}_T = s\nu$, dividing Eq. (116) by ν^2 and using this bound gives

$$Y^2 \le [60s(1+\mathfrak{r}_T)+sY]\left(Y+\frac{15}{s}\right) \implies (1-s)Y^2 \le [60s(1+\mathfrak{r}_T)+15]\,Y + 900(1+\mathfrak{r}_T).$$

Eq. (3), together with $d_{\min} \le d$, gives $s = \sqrt{1-\beta_2} \le 1/\sqrt{8} < 2/5$, and hence $(1-s)^{-1} < 5/3$. Therefore

$$Y^2 \le [100s(1+\mathfrak{r}_T)+25]\,Y + 1500(1+\mathfrak{r}_T).$$

For nonnegative Y, c, d, the inequality $Y^2 \le cY + d$ implies $Y \le \frac{c+\sqrt{c^2+4d}}{2} \le c + \sqrt{d}$. Consequently,

$$Y \le 100s(1+\mathfrak{r}_T) + 25 + \sqrt{1500}\sqrt{1+\mathfrak{r}_T}.$$

Since $25 \le 25\sqrt{1+\mathfrak{r}_T}$ and $\sqrt{1500} < 39$, we obtain $Y \le 100\left\{s(1+\mathfrak{r}_T) + \sqrt{1+\mathfrak{r}_T}\right\}$. Multiplying the last inequality by ν and using $\mathcal{G}_T = \sqrt{1-\beta_2}\,\nu$ from Eq. (96) gives

$$H_T \le 100\left\{(1+\mathfrak{r}_T)\mathcal{G}_T + \sqrt{1+\mathfrak{r}_T}\,\nu\right\}.$$

It remains to express ν in the parameters of the theorem. Recalling the tuning in Eq. (9), we obtain

$$\nu = \sigma_B(1-\beta_2)^{1-1/p} = \sigma_B\left(\frac{\Delta L}{\sigma_B^2 T}\right)^{\frac{p-1}{3p-2}} = (\Delta L)^{\frac{p-1}{3p-2}}\sigma_B^{\frac{p}{3p-2}}T^{-\frac{p-1}{3p-2}}.$$

Finally, Eq. (23) gives $\sigma_B^{\frac{p}{3p-2}} = 2^{\frac{1}{3p-2}}\sigma^{\frac{p}{3p-2}}B^{-\frac{p-1}{3p-2}}$. Substituting the last two displays and $\mathcal{G}_T = \sqrt{\Delta L/T}$ yields exactly Eq. (11).

$\mathfrak{r}_T \to 0$ follows directly from the exact tuning. Indeed,

$$\frac{1}{T(1-\beta_2)} = \left(\frac{\sigma_B^2}{\Delta L}\right)^{\frac{p}{3p-2}} T^{-\frac{2(p-1)}{3p-2}}, \quad \sqrt{\frac{1-\beta_2}{T}} = \left(\frac{\Delta L}{\sigma_B^2}\right)^{\frac{p}{2(3p-2)}} T^{-\frac{2p-1}{3p-2}}.$$

Both powers of T are negative for $p > 1$, and the logarithm in Eq. (10) does not change their convergence to zero. This proves $\mathfrak{r}_T = o(1)$ for fixed problem parameters. □

D.5 Proof of Theorem 4

Proof of Theorem 4. Let T_0 denote the quantity inside the ceiling in Eq. (13), and take $T := \lceil T_0 \rceil$. We use the mini-batch scale $\sigma_B = 2^{1/p}\sigma B^{-(p-1)/p}$ inside the proof. The first threshold in Eq. (12) gives $\sigma_B/\varepsilon \ge 1$. The second threshold gives $200\sqrt{\Delta L}/\varepsilon \ge 2p/(p-1) \ge 4$, so $\Delta L/\varepsilon^2 \ge 1/2500$. Since $(3p-2)/(p-1) = 2 + p/(p-1) \ge 4$, $T_0 \ge \frac{(200\sqrt{2})^4}{2500} > 1$. Hence $T_0 \le T \le 2T_0$. Write $\delta := 1-\beta_2$. The exact tuning in Eq. (9) gives

$$\delta = \left(\frac{\Delta L}{\sigma_B^2 T}\right)^{\frac{p}{3p-2}} \le \left(\frac{\Delta L}{\sigma_B^2 T_0}\right)^{\frac{p}{3p-2}} = \left(\frac{\varepsilon}{200\sqrt{2}\,\sigma_B}\right)^{\frac{p}{p-1}} \le \frac{1}{8}\sqrt{\frac{d_{\min}}{d}}. \tag{117}$$

The first threshold in Eq. (12) gives $(\varepsilon/\sigma_B)^{p/(p-1)} \le \sqrt{d_{\min}/d}$, while $(200\sqrt{2})^{-p/(p-1)} \le 1/8$. These two inequalities prove the last comparison and hence Eq. (3), the geometric condition used by Theorem 3.

We next verify the bias-correction burn-in. We first record an elementary implication. For $a > 0$ and $r \ge 2$, if $X \ge 2r\,[1 + \log(1+2ar)]$, then

$$X \ge 2 + \log 2 + r\log(aX). \tag{118}$$

Indeed, set $X_0 := 2r[1+\log(1+2ar)]$. The function $X \mapsto X - 2 - \log 2 - r\log(aX)$ is increasing for $X \ge X_0 \ge 2r$. Moreover,

$$\log(aX_0) = \log(2ar) + \log\left(1+\log(1+2ar)\right) \le 2\log(1+2ar),$$

so its value at X_0 is at least $2r - 2 - \log 2 > 0$. This proves Eq. (118). Apply this implication with

$$X = 200\frac{\sqrt{\Delta L}}{\varepsilon}, \qquad r = \frac{p}{p-1}, \qquad a = \frac{\sqrt{2}\,\sigma_B}{\sqrt{\Delta L}}.$$

The second threshold in Eq. (12) is precisely its hypothesis. Therefore

$$200\frac{\sqrt{\Delta L}}{\varepsilon} \geq 2 + \log 2 + \frac{p}{p-1}\log\left(\frac{200\sqrt{2}\,\sigma_B}{\varepsilon}\right). \tag{119}$$

On the other hand, $T \leq 2T_0$ and the exact tuning give

$$\log \delta^{-1} \leq \frac{p}{p-1}\log\left(\frac{200\sqrt{2}\,\sigma_B}{\varepsilon}\right) + \frac{p}{3p-2}\log 2 \leq \frac{p}{p-1}\log\left(\frac{200\sqrt{2}\,\sigma_B}{\varepsilon}\right) + \log 2.$$

Combining this with Eq. (119), and using that $T\delta$ is increasing in T under the exact tuning, yields

$$\sqrt{T\delta} \geq \sqrt{T_0\left(\frac{\Delta L}{\sigma_B^2 T_0}\right)^{\frac{p}{3p-2}}} = 200\sqrt{2}\frac{\sqrt{\Delta L}}{\varepsilon} \geq \sqrt{2}\left(2 + \log \delta^{-1}\right).$$

Thus $T\delta \geq 2\left(2 + \log \delta^{-1}\right)^2$. Since $\log \min\{T, \delta^{-1}\} \leq \log \delta^{-1}$, the factorized transient in Eq. (10) satisfies

$$\mathfrak{r}_T \leq \frac{2 + \log \delta^{-1}}{T\delta} + \left(2 + \log \delta^{-1}\right)\frac{\delta}{\sqrt{T\delta}} \leq \frac{1}{4} + \frac{\delta}{\sqrt{2}} \leq \frac{1}{4} + \frac{1}{8\sqrt{2}} < 1,$$

where Eq. (117) gives $\delta \leq 1/8$. Uniform independent sampling gives $\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(\widehat{x}))\right] = T^{-1}\sum_t \mathbb{E}\mathsf{S}_{\mathcal{P}}(\nabla F(x_t))$. The choice $T \geq T_0$ implies

$$\nu = \sigma_B \delta^{1-1/p} \leq \frac{\varepsilon}{200\sqrt{2}}, \qquad \mathcal{G}_T = \sqrt{\delta}\,\nu \leq \frac{\varepsilon}{800}.$$

Applying Eq. (11) with $\mathfrak{r}_T < 1$ gives

$$\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F(\widehat{x}))\right] \leq 200\mathcal{G}_T + 100\sqrt{2}\,\nu \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{2} < \varepsilon.$$

Finally, $\lceil a \rceil \leq a + 1$ proves the last inequality in Eq. (13). □

E Proofs of the Lower Bounds

Throughout this appendix, R_ε and Q_ε are the two scales defined in Theorem 5. The proof has three steps. First, we construct a Euclidean problem whose useful directions must be discovered in order and calibrate its oracle to the pth-moment bound. Second, a random rotation makes the undiscovered directions invisible to an arbitrary adaptive method. Third, a Euclidean subspace of the Adam-mini geometry transfers the lower bound to $\mathsf{S}_{\mathcal{P}}$. The last subsection records a simpler exact transfer for a restricted class of methods. It is not used in the proof of Theorem 5.

E.1 Constructing the Euclidean Hard Instance

We begin with two definitions. For $x \in \mathbb{R}^m$, let $\mathrm{supp}(x) := \{i : x_i \neq 0\}$. For every integer $\ell \geq 0$, define $[x_{\leq \ell}]_i := x_i \mathbb{I}_{\{i \leq \ell\}}$. Thus $x_{\leq 0} = 0$ and $x_{\leq \ell} = x$ when $\ell \geq m$. For $\alpha \geq 0$, define

$$\mathrm{prog}_\alpha(x) := \max\big(\{i \in [m] : |x_i| > \alpha\} \cup \{0\}\big).$$

Suppose a method queries x_t and receives the stochastic gradient g_t. It is *zero-respecting* if

$$\mathrm{supp}(x_t) \subseteq \bigcup_{s<t} \mathrm{supp}(g_s) \qquad \text{for every } t \geq 1. \tag{120}$$

As in the oracle-complexity definition, an unqueried output is appended as the last query. Thus Eq. (120) also applies to the returned point.

An oracle response is specified by a measurable map $g : \mathbb{R}^m \times \mathcal{Z} \to \mathbb{R}^m$. When the method queries x and the oracle draws a random seed $\zeta \in \mathcal{Z}$, it returns the vector $g(x, \zeta)$. Thus g describes one random response, before averaging over the seed. When it is an unbiased oracle for an objective f, it also satisfies $\mathbb{E}_\zeta[g(x, \zeta)] = \nabla f(x)$ for every x. Different calls use independent copies of ζ.

Write $g_\zeta(x) := g(x, \zeta)$ and $k(x) := \mathrm{prog}_{1/4}(x)$. Following Arjevani et al. (2023, Definition 2), we call g a probability-θ zero chain if

$$\begin{aligned} \mathbb{P}_\zeta\{\forall x \in \mathbb{R}^m : \mathrm{prog}_0(g_\zeta(x)) \le k(x), \quad g_\zeta(x) = g_\zeta(x_{\le k(x)})\} &\ge 1-\theta, \\ \mathbb{P}_\zeta\{\forall x \in \mathbb{R}^m : \mathrm{prog}_0(g_\zeta(x)) \le k(x)+1, \quad g_\zeta(x) = g_\zeta(x_{\le k(x)+1})\} &= 1. \end{aligned} \tag{121}$$

The first line says that one call fails to expose a new coordinate with probability at least $1-\theta$. The second says that one call can never expose more than the next coordinate. The same seed appears throughout either event because the asserted property must hold simultaneously for every possible query. The two equalities also prevent the unreached coordinates of a query from changing the response, a fact used by the random rotation argument.

The next lemma is a parameterized form of the pth-moment calibration in Hübler et al. (2026, Lemma E.6). Its hard function is the canonical chain of Arjevani et al. (2023). We include the details because the later rotation needs the stochastic gradient map itself, not only the final complexity rate. Besides keeping the chain length r free, we modify the frontier response so that it obeys the two truncation identities needed by the rotation. The proof below verifies these identities and accounts for the resulting two-coordinate noise explicitly.

Lemma E.1 (Finite-p calibration of the canonical chain)**.** *There are constants $c_p, C_p > 0$, depending only on p, such that the following holds. Let $0 < \varepsilon \le c_p \min\{\sigma, \sqrt{\Delta L}\}$ and let*

$$1 \le r \le c_p \frac{\Delta L}{\varepsilon^2}. \tag{122}$$

There is an L-smooth objective $f_r : \mathbb{R}^{m_r} \to \mathbb{R}$ and a stochastic gradient oracle O_r, obtained from a probability-θ zero chain by a deterministic change of scale, such that

$$m_r = 8r, \quad f_r(0) - \inf f_r \le C_p \frac{r\varepsilon^2}{L} \le \Delta, \qquad \theta = \left(\frac{184\sqrt{2}\,\varepsilon}{\sigma}\right)^{\frac{p}{p-1}}, \quad \theta^{-1} = (184\sqrt{2})^{-\frac{p}{p-1}} Q_\varepsilon \tag{123}$$

and the oracle satisfies

$$\mathbb{E}_\zeta\left[\mathsf{O}_r(x,\zeta)\right] = \nabla f_r(x), \qquad \mathbb{E}_\zeta\left[\|\mathsf{O}_r(x,\zeta) - \nabla f_r(x)\|^p\right] \le \sigma^p \quad \textit{for every } x.$$

Every randomized zero-respecting method needs at least

$$c_p r Q_\varepsilon = c_p r \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}} \tag{124}$$

oracle calls before it can return a point with expected gradient norm at most ε.

Proof. Let $m_r := 8r$. Define

$$\Psi(t) := \begin{cases} 0, & t \le 1/2, \\ \exp\left(1 - \dfrac{1}{(2t-1)^2}\right), & t > 1/2, \end{cases} \qquad \Phi(t) := \sqrt{e}\int_{-\infty}^{t} \exp(-u^2/2)\,du,$$

and

$$F_{m_r}(x) := -\Psi(1)\Phi(x_1) + \sum_{i=2}^{m_r} \left\{\Psi(-x_{i-1})\Phi(-x_i) - \Psi(x_{i-1})\Phi(x_i)\right\}. \tag{125}$$

This is the canonical chain in Eqs. (16)–(17) of Arjevani et al. (2023). Parts 1–6 of their Lemma 2 give

$$F_{m_r}(0) - \inf F_{m_r} \le 12 m_r, \qquad \mathrm{Lip}(\nabla F_{m_r}) \le 152, \qquad \|\nabla F_{m_r}(x)\|_\infty \le 23.$$

The same lemma also shows that

$$\mathrm{prog}_1(x) < m_r \quad \Longrightarrow \quad \|\nabla F_{m_r}(x)\| \ge \left|\nabla_{\mathrm{prog}_1(x)+1} F_{m_r}(x)\right| > 1. \tag{126}$$

We next record the local structure of the gradient, which will be needed for the two truncation identities. Direct differentiation of Eq. (125) gives

$$\begin{aligned} \nabla_1 F_{m_r}(x) &= -\Psi(1)\Phi'(x_1) - \Psi'(-x_1)\Phi(-x_2) - \Psi'(x_1)\Phi(x_2), \\ \nabla_i F_{m_r}(x) &= -\Psi(-x_{i-1})\Phi'(-x_i) - \Psi(x_{i-1})\Phi'(x_i) \\ &\quad - \Psi'(-x_i)\Phi(-x_{i+1}) - \Psi'(x_i)\Phi(x_{i+1}), \qquad 2 \le i < m_r, \\ \nabla_{m_r} F_{m_r}(x) &= -\Psi(-x_{m_r-1})\Phi'(-x_{m_r}) - \Psi(x_{m_r-1})\Phi'(x_{m_r}). \end{aligned} \tag{127}$$

In particular, each gradient coordinate depends only on that coordinate and its immediate neighbors. Moreover, $\Psi(\pm u) = \Psi'(\pm u) = 0$ whenever $|u| \le 1/4$. If $k = k(x) := \mathrm{prog}_{1/4}(x)$, these facts imply

$$\begin{aligned} \mathrm{prog}_0(\nabla F_{m_r}(x)) &\le k+1, \\ \nabla_i F_{m_r}(x) &= \nabla_i F_{m_r}(x_{\le k}) \qquad \text{for every } i < k, \\ \nabla F_{m_r}(x) &= \nabla F_{m_r}(x_{\le k+1}). \end{aligned} \tag{128}$$

For the last identity, the only possible dependence of coordinate $k+1$ on x_{k+2} is multiplied by $\Psi'(\pm x_{k+1})$, which is zero because $|x_{k+1}| \le 1/4$. The conventions above cover $k = 0$ and $k = m_r$.

For $\zeta \sim \mathrm{Bernoulli}(\theta)$, define

$$[g_{m_r}(x,\zeta)]_i := \nabla_i F_{m_r}(x) \left[1 + \mathbb{I}_{\{i \ge \mathrm{prog}_{1/4}(x)\}} \left(\frac{\zeta}{\theta} - 1\right)\right]. \tag{129}$$

This is a variant of Arjevani et al. (2023, Eq. (18)) that randomizes the current frontier coordinate together with the next coordinate. We verify the required properties directly. Since $\mathbb{E}\zeta = \theta$, the map is unbiased coordinatewise. Also, $k(x_{\le k}) = k(x_{\le k+1}) = k$, with the first equality interpreted at $k = 0$ using $x_{\le 0} = 0$. When $\zeta = 0$, every coordinate $i \ge k$ is suppressed. The second line of Eq. (128) therefore gives $g_{m_r}(x,0) = g_{m_r}(x_{\le k},0)$ and $\mathrm{prog}_0(g_{m_r}(x,0)) \le k$. For either value of ζ, the first and third lines of Eq. (128) give $g_{m_r}(x,\zeta) = g_{m_r}(x_{\le k+1},\zeta)$ and $\mathrm{prog}_0(g_{m_r}(x,\zeta)) \le k+1$. These statements hold simultaneously for every x, so Eq. (121) follows.

The centered noise can affect at most two coordinates. More precisely,

$$\begin{aligned} g_{m_r}(x,\zeta) - \nabla F_{m_r}(x) &= \left(\frac{\zeta}{\theta} - 1\right) \sum_{\substack{i \in [m_r] \\ i \ge k(x)}} \nabla_i F_{m_r}(x) e_i, \\ \|g_{m_r}(x,\zeta) - \nabla F_{m_r}(x)\| &\le 23\sqrt{2} \left|\frac{\zeta}{\theta} - 1\right|. \end{aligned} \tag{130}$$

The following rescaling also applies to the soft-compressed chain below. Let

$$s := 4\varepsilon, \qquad \bar{\ell} := 155, \qquad b := \frac{\bar{\ell} s}{L}, \qquad a := sb,$$

and define $f_r(x) := aF_{m_r}(x/b)$ and $\mathsf{O}_r(x,\zeta) := s g_{m_r}(x/b,\zeta)$. Since $a/b = s$ and $152a/b^2 \le L$, the objective is L-smooth. Its gap is at most

$$12 a m_r \le C \frac{r\varepsilon^2}{L},$$

which is at most Δ after reducing c_p. At $y = x/b$, the centered oracle noise is $s(\zeta/\theta - 1)\sum_{i \ge k(y)} \nabla_i F_{m_r}(y) e_i$. For $0 < \theta \le 1$,

$$\mathbb{E}\,|\zeta/\theta - 1|^p = (1-\theta)^p \theta^{1-p} + 1 - \theta \le 2\theta^{1-p}. \tag{131}$$

After reducing c_p if necessary, the accuracy condition makes the value of θ in Eq. (123) at most one. Since $s = 4\varepsilon$, Eq. (130) gives

$$\mathbb{E}\,\|\mathsf{O}_r(x,\zeta) - \nabla f_r(x)\|^p \le 2(23\sqrt{2}s)^p \theta^{1-p} = 2^{1-p}\sigma^p \le \sigma^p.$$

Finally, scalar rescaling does not change coordinate supports, so the base-coordinate queries x_t/b are zero-respecting. Lemma 1 of Arjevani et al. (2023), applied with failure probability $1/2$, gives

$$\mathrm{prog}_0(x_t/b) < m_r \quad \text{for every} \quad t \le \frac{m_r - \log 2}{2\theta} \tag{132}$$

with probability at least $1/2$. The statement remains true after conditioning on the method's internal seed. On this event, $\mathrm{prog}_1(x_t/b) \le \mathrm{prog}_0(x_t/b) < m_r$. Eq. (126) and the scaling then give $\|\nabla f_r(x_t)\| > s = 4\varepsilon$. Hence the expected gradient norm is greater than 2ε. Since $m_r = 8r$ and θ^{-1} is given exactly in Eq. (123), the horizon in Eq. (132) is at least $c_p r Q_\varepsilon$. This proves Eq. (124). □

E.2 Random Rotation for General Algorithms

The preceding lemma controls algorithms whose queries can use only coordinates already seen in an oracle response. A random rotation hides the coordinate system and removes this restriction. For integers $D \geq m$, the Stiefel manifold $\mathrm{St}(D, m)$ is the set of ordered orthonormal m-frames in $\mathbb{R}^D$:

$$\mathrm{St}(D, m) := \{V \in \mathbb{R}^{D \times m} : V^\top V = I_m\}. \tag{133}$$

The abbreviation "St" stands for Stiefel. Every matrix in this set is a linear isometric embedding from $\mathbb{R}^m$ to $\mathbb{R}^D$, since $\|Vz\| = \|z\|$. We draw V uniformly by taking the first m columns of a Haar-uniform orthogonal matrix. Equivalently, the distribution is unchanged when all columns are rotated by any fixed orthogonal transformation of $\mathbb{R}^D$.

Random rotation for bounded queries. We use the following single-query specialization of Arjevani et al. (2023, Lemma 5, Eq. (27)) from the published *Mathematical Programming* version. Its proof in Appendix B.1 uses the truncated-history argument of Lemma 12 to handle adaptive queries. Let $0 < \theta \leq 1$, let g be a probability-θ zero chain on $\mathbb{R}^m$, and draw $V \sim \mathrm{Unif}(\mathrm{St}(D, m))$, and let an adaptive method query points y_t with $\|y_t\| \leq R$ through the rotated responses $Vg(V^\top y_t, \zeta_t)$. For every $\delta \in (0, 1)$,

$$\begin{aligned} D \geq \left\lceil m + 32R^2 \log\left(\frac{2m^2}{\theta\delta}\right)\right\rceil \\ \implies \quad \mathbb{P}\left\{\mathrm{prog}_{1/4}(V^\top y_t) < m \ \forall t \leq \frac{m - \log(2/\delta)}{2\theta}\right\} \geq 1 - \delta. \end{aligned} \tag{134}$$

The probability is over V and the fresh oracle seeds. A randomized method is covered by first conditioning on its internal seed. We refer to Eq. (134) as the random rotation lemma.

Lemma E.2 (Random rotation under a finite pth moment)**.** *Use the integer $m_r = 8r$ and the reveal probability θ from Lemma E.1. For a sufficiently large universal constant C_{rot}, set*

$$D_r := \left\lceil C_{\mathrm{rot}} m_r \log\left(\frac{8m_r^2}{\theta}\right)\right\rceil. \tag{135}$$

There is a family of rotated and softly compressed instances in $\mathbb{R}^{D_r}$ such that every randomized stochastic gradient method making at most $c_p r Q_\varepsilon$ calls has expected gradient norm at least $3\varepsilon/2$ on some member of the family. Every instance has initial gap at most Δ, is L-smooth, and satisfies Eq. (14)*. Moreover,*

$$D_r \leq Cr \log(2 + rQ_\varepsilon) \tag{136}$$

for a universal constant C.

Proof. Bounded-query simulation. We first rotate and compress the unscaled chain. Draw V uniformly from $\mathrm{St}(D_r, m_r)$. Set

$$\rho_R(x) := \frac{x}{\sqrt{1 + \|x\|^2 / R^2}}, \qquad R := 230\sqrt{m_r}, \qquad \eta_0 := \frac{1}{5}.$$

The Jacobian of ρ_R is

$$J_R(x) = \frac{I - \rho_R(x)\rho_R(x)^\top / R^2}{\sqrt{1 + \|x\|^2 / R^2}}. \tag{137}$$

In particular, $\|\rho_R(x)\| \leq R$ and $\|J_R(x)\|_{\mathrm{op}} \leq 1$. Define

$$\begin{aligned} \widehat{F}_{m_r,V}(x) &:= F_{m_r}(V^\top \rho_R(x)) + \frac{\eta_0}{2}\|x\|^2, \\ \widehat{\mathsf{O}}_{m_r,V}(x, \zeta) &:= J_R(x)^\top V g_{m_r}(V^\top \rho_R(x), \zeta) + \eta_0 x. \end{aligned}$$

We next prove that this family is hard. Let A be any randomized method for the compressed oracle, and condition on its internal seed. If x_t is its tth query, set $y_t := \rho_R(x_t)$. The sequence $\{y_t\}$ can be generated by a method that queries the bounded rotated oracle

$$\mathsf{O}^{\mathrm{rot}}_{m_r,V}(y, \zeta) := V g_{m_r}(V^\top y, \zeta). \tag{138}$$

Indeed, after receiving Eq. (138) at y_t, that method can reconstruct the response to A as $J_R(x_t)^\top \mathsf{O}^{\mathrm{rot}}_{m_r,V}(y_t,\zeta_t)+\eta_0 x_t$. Moreover, $\|y_t\|\le R$ for every t.

Applying the random rotation lemma. Since $R^2=230^2 m_r$ and $\log(8m_r^2/\theta)>1$, choosing the universal constant C_{rot} sufficiently large makes the definition of D_r in Eq. (135) satisfy the dimension condition in Eq. (134). Apply Eq. (134) with $m=m_r$ and $\delta=1/4$. We obtain

$$\mathbb{P}\left\{\mathrm{prog}_{1/4}(V^\top y_t)<m_r\ \forall t\le \frac{m_r-\log 8}{2\theta}\right\}\ge\frac{3}{4}. \tag{139}$$

The hiding argument uses the truncated history in Arjevani et al. (2023, Appendix B.1, Lemma 12). After fixing the oracle seeds and the algorithm's internal randomness, a query generated from this history depends only on the columns revealed so far. The two truncation identities in Eq. (121) make the truncated and actual histories agree until a hidden coordinate crosses the threshold $1/4$. Conditional on the revealed columns, the remaining columns are uniform in their orthogonal complement, so sphere concentration applies to the truncated query. A union bound over queries and hidden directions, together with the reveal-count bound, yields Eq. (139).

Large gradients before the last direction is exposed. On the event in Eq. (139), fix a query x, write $y:=\rho_R(x)$ and $q:=V^\top y$, and set

$$j:=\mathrm{prog}_1(q)+1,\qquad u_j:=Ve_j,\qquad w:=V\nabla F_{m_r}(q).$$

The progress event implies $j\le m_r$. Part 6 of Lemma 2 in Arjevani et al. (2023) gives

$$\left|u_j^\top w\right|=|\nabla_j F_{m_r}(q)|>1,\qquad \left|u_j^\top y\right|=|q_j|\le 1. \tag{140}$$

Part 3 of the same lemma gives $\|w\|=\|\nabla F_{m_r}(q)\|\le 23\sqrt{m_r}$.

First suppose $\|x\|\le R/2$ and put $h:=\sqrt{1+\|x\|^2/R^2}$. Then $h\le\sqrt{5}/2$, $x=hy$, and Eq. (137) yields

$$u_j^\top\nabla\widehat{F}_{m_r,V}(x)=\frac{u_j^\top w}{h}-\frac{(u_j^\top y)(y^\top w)}{R^2 h}+\eta_0 h\,u_j^\top y.$$

Using Eq. (140), $\|y\|\le R/2$, and $R=230\sqrt{m_r}$, we obtain

$$\left\|\nabla\widehat{F}_{m_r,V}(x)\right\|\ge\left|u_j^\top\nabla\widehat{F}_{m_r,V}(x)\right|>\frac{2}{\sqrt{5}}-\frac{23\sqrt{m_r}}{2R}-\frac{1}{2\sqrt{5}}=\frac{2}{\sqrt{5}}-\frac{1}{20}-\frac{1}{2\sqrt{5}}>\frac{1}{2}. \tag{141}$$

If $\|x\|>R/2$, then $\|J_R(x)\|_{\mathrm{op}}\le 2/\sqrt{5}$, and hence

$$\left\|\nabla\widehat{F}_{m_r,V}(x)\right\|\ge\eta_0\|x\|-\|J_R(x)\|_{\mathrm{op}}\|w\|>\frac{R}{10}-\frac{46}{\sqrt{5}}\sqrt{m_r}=\left(23-\frac{46}{\sqrt{5}}\right)\sqrt{m_r}>\frac{1}{2}. \tag{142}$$

Together, Eqs. (139), (141) and (142) show that, with probability at least $3/4$,

$$\left\|\nabla\widehat{F}_{m_r,V}(x_t)\right\|>\frac{1}{2}\quad\text{for every}\quad t\le\frac{m_r-\log 8}{2\theta}. \tag{143}$$

Scaling and oracle verification. Now use the same s,b,a as in the preceding proof and set

$$\widetilde{f}_{r,V}(x):=a\widehat{F}_{m_r,V}(x/b),\qquad \widetilde{\mathsf{O}}_{r,V}(x,\zeta):=s\widehat{\mathsf{O}}_{m_r,V}(x/b,\zeta).$$

To transfer the hard event through this rescaling, let B be any method for the scaled oracle. Construct an unscaled method A as follows. When B requests x_t, method A queries $y_t:=x_t/b$ and multiplies the returned stochastic gradient by s before passing it to B. The histories seen by B are therefore exactly those generated by $\widetilde{\mathsf{O}}_{r,V}$. The two methods use the same number of calls, and $\nabla\widetilde{f}_{r,V}(x_t)=s\nabla\widehat{F}_{m_r,V}(y_t)$. Arjevani et al. (2023, Lemma 7) shows that the unscaled compressed objective has gap at most $12m_r$ and 155-Lipschitz gradient. The same calculation as in Lemma E.1 therefore shows that $\widetilde{f}_{r,V}$ has gap at most Δ and is L-smooth. Since $m_r=8r$ and Eq. (123) gives the exact value of θ^{-1}, the horizon in Eq. (143) is at least $c_p r Q_\varepsilon$. On that event, every output within this budget has gradient norm greater than $s/2=2\varepsilon$. Consequently,

$$\mathbb{E}_{V,R_{\mathsf{B}},\zeta}\left\|\nabla\widetilde{f}_{r,V}(\widehat{x})\right\|\ge\frac{3}{4}\,2\varepsilon=\frac{3}{2}\varepsilon.$$

The final-query convention in Eq. (15) covers an output that was not among the method's earlier queries. For every fixed randomized method, averaging over V gives risk at least $3\varepsilon/2$. The supremum over fixed V is at least this average, so at least one fixed instance has risk above ε.

It remains to check the pth-moment property, since the final theorem of Arjevani et al. (2023) specializes the construction to a second moment. The deterministic quadratic terms cancel after centering, giving, at $y = x/b$ and $q = V^\top \rho_R(y)$,

$$\widetilde{\mathsf{O}}_{r,V}(x,\zeta) - \nabla \widetilde{f}_{r,V}(x) = s J_R(y)^\top V \left(g_{m_r}(q,\zeta) - \nabla F_{m_r}(q)\right).$$

The soft projection satisfies $\|J_R(y)\|_{\mathrm{op}} \le 1$, while $V^\top V = I_{m_r}$. Hence this centered noise has norm at most $s\,\|g_{m_r}(q,\zeta) - \nabla F_{m_r}(q)\|$, and Eqs. (123), (130) and (131) bound its pth moment by σ^p. The same identity before taking norms proves unbiasedness. Both properties hold for every fixed V. The matrix V indexes the hard instance and is not resampled at each oracle call.

Dimension count. Finally, $m_r = 8r$ and $\theta^{-1} = (184\sqrt{2})^{-p/(p-1)} Q_\varepsilon$. Substitution in Eq. (135) gives

$$D_r \le Cr \log(2 + rQ_\varepsilon),$$

which proves Eq. (136). □

Corollary 10 (Euclidean lower bound for arbitrary randomized methods)**.** *Under the accuracy condition in Eq.* (16)*, every randomized Euclidean stochastic gradient method needs*

$$\Omega\left(R_\varepsilon Q_\varepsilon\right) = \Omega\left(\frac{\Delta L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right) \tag{144}$$

queries in dimension

$$D_{\mathrm{hard}} \le C_p R_\varepsilon \log(2 + R_\varepsilon Q_\varepsilon). \tag{145}$$

For every integer $1 \le r \le c_p R_\varepsilon$, the same construction uses at most $C_p r \log(2 + rQ_\varepsilon)$ dimensions and gives a lower bound $c_p r Q_\varepsilon$.

Proof. Take r to be a sufficiently small constant multiple of R_ε in Lemmas E.1 and E.2. Leaving r unchanged gives the parameterized statement. □

E.3 A Euclidean Subspace for Every Partition

Proof of Lemma 2. The upper bound is deterministic. Weighted Cauchy–Schwarz gives

$$\mathsf{S}_{\mathcal{P}}(x) = \sum_{j=1}^{J} \sqrt{\frac{d_j}{d}} \left\|x^{(j)}\right\| \le \|x\|. \tag{146}$$

Let X be uniform on the Euclidean unit sphere in $\mathbb{R}^d$. The exact beta function formula is

$$\mathbb{E}\left[\left\|X^{(j)}\right\|\right] = \frac{\Gamma((d_j+1)/2)\Gamma(d/2)}{\Gamma(d_j/2)\Gamma((d+1)/2)}. \tag{147}$$

Wendel's gamma-ratio inequality (Wendel, 1948) implies

$$\frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \ge \frac{1}{2}\sqrt{k}, \qquad \frac{\Gamma(d/2)}{\Gamma((d+1)/2)} \ge \sqrt{\frac{2}{d}}.$$

Substituting into Eq. (147) gives

$$\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(X)\right] \ge \frac{1}{\sqrt{2}}. \tag{148}$$

Since $0 \le \mathsf{S}_{\mathcal{P}}(X) \le 1$, if $q := \mathbb{P}\{\mathsf{S}_{\mathcal{P}}(X) \ge 1/(2\sqrt{2})\}$, then

$$\frac{1}{\sqrt{2}} \le \mathbb{E}\mathsf{S}_{\mathcal{P}}(X) \le \frac{1-q}{2\sqrt{2}} + q.$$

This implies $q > 1/2$, so a median is at least $1/(2\sqrt{2})$. The map $x \mapsto \mathsf{S}_{\mathcal{P}}(x)$ is 1-Lipschitz by Eq. (146). Lévy's concentration lemma on the sphere (Milman & Schechtman, 1986) therefore gives, for a universal constant $c_{\mathrm{Lev}} > 0$,

$$\mathbb{P}\left\{\mathsf{S}_{\mathcal{P}}(X) < \frac{1}{4\sqrt{2}}\right\} \le 2\exp(-c_{\mathrm{Lev}} d). \tag{149}$$

To pass from one direction to all directions, we first discretize the unit sphere. A δ-net is a finite subset $\mathcal{N}$ of the sphere such that every unit vector z is within Euclidean distance δ of some $z_0 \in \mathcal{N}$. A standard volume argument gives a $1/(8\sqrt{2})$-net of the unit sphere in $\mathbb{R}^r$ with at most 24^r points.

Draw U uniformly from $\mathrm{St}(d, r)$ as defined in Eq. (133). For each fixed $z \in \mathcal{N}$, the vector Uz is uniform on the unit sphere in $\mathbb{R}^d$. Choose $c_{\dim} > 0$ sufficiently small that every positive integer $r \le c_{\dim} d$ has $d \ge 4$ and $2 \cdot 24^r \exp(-c_{\mathrm{Lev}} d) < 1$. A union bound in Eq. (149) then gives

$$\mathbb{P}\left\{\mathsf{S}_{\mathcal{P}}(Uz) \ge \frac{1}{4\sqrt{2}} \text{ for every } z \in \mathcal{N}\right\} > 0.$$

Fix a realization of U in this event. For an arbitrary unit vector z, choose $z_0 \in \mathcal{N}$ with $\|z - z_0\| \le 1/(8\sqrt{2})$. The upper bound in Eq. (146) and the isometry of U give

$$\mathsf{S}_{\mathcal{P}}(Uz) \ge \mathsf{S}_{\mathcal{P}}(Uz_0) - \mathsf{S}_{\mathcal{P}}(U(z - z_0)) \ge \frac{1}{4\sqrt{2}} - \|z - z_0\| \ge \frac{1}{8\sqrt{2}}.$$

For nonzero z, apply this inequality to $z/\|z\|$ and use $\mathsf{S}_{\mathcal{P}}(Uz) = \|z\|\, \mathsf{S}_{\mathcal{P}}(Uz/\|z\|)$. The case $z = 0$ is immediate. This proves the lower bound in Eq. (18) with $c_{\mathrm{sec}} = 1/(8\sqrt{2})$. □

E.4 Transferring the Lower Bound to Any Partition

Proof of Theorem 5. Set $\bar{\varepsilon} := \varepsilon / c_{\mathrm{sec}}$. The Euclidean hard-instance family in Corollary 10, at accuracy $\bar{\varepsilon}$, lives in a common space $\mathbb{R}^{D_{\mathrm{hard}}}$ (padding with inactive coordinates if necessary). By decreasing c_p if necessary, $\bar{\varepsilon}$ satisfies the accuracy condition of Corollary 10. Under the dimension condition in Eq. (16), its ambient dimension satisfies $D_{\mathrm{hard}} \le c_{\dim} d$ after adjusting universal constants. Hence Lemma 2 gives an isometry $U_{\mathcal{P}} : \mathbb{R}^{D_{\mathrm{hard}}} \to \mathbb{R}^d$.

Let A be an arbitrary randomized method in the ambient space. Define a Euclidean method B that interacts with its oracle O as follows. Run A with the same internal seed. When A queries x_t, method B queries $y_t := U_{\mathcal{P}}^\top x_t$ and passes the response $U_{\mathcal{P}} \mathsf{O}(y_t, \xi_t)$ back to A. It returns $U_{\mathcal{P}}^\top \widehat{x}$ when A returns $\widehat{x}$. If this output was not queried earlier, B makes the additional query counted in Eq. (15). Thus B uses no more calls than A.

Corollary 10 supplies a pair (f, O) that is hard for B. Lift this pair by defining

$$F_{U_{\mathcal{P}}}(x) := f(U_{\mathcal{P}}^\top x), \qquad \mathsf{O}_{U_{\mathcal{P}}}(x, \xi) := U_{\mathcal{P}} \mathsf{O}(U_{\mathcal{P}}^\top x, \xi). \tag{150}$$

Because $U_{\mathcal{P}}^\top U_{\mathcal{P}} = I_{D_{\mathrm{hard}}}$, the map $U_{\mathcal{P}}^\top$ is surjective. For every y, one may take $x = U_{\mathcal{P}} y$. Therefore the reference-point gap is preserved:

$$F_{U_{\mathcal{P}}}(0) - \inf_x F_{U_{\mathcal{P}}}(x) = f(0) - \inf_y f(y).$$

For all $x, y \in \mathbb{R}^d$,

$$\|\nabla F_{U_{\mathcal{P}}}(x) - \nabla F_{U_{\mathcal{P}}}(y)\| = \left\|U_{\mathcal{P}}\left(\nabla f(U_{\mathcal{P}}^\top x) - \nabla f(U_{\mathcal{P}}^\top y)\right)\right\| \le L\left\|U_{\mathcal{P}}^\top (x - y)\right\| \le L\|x - y\|,$$

so $F_{U_{\mathcal{P}}}$ is L-smooth. Moreover,

$$\mathbb{E}\left[\mathsf{O}_{U_{\mathcal{P}}}(x, \xi)\right] = U_{\mathcal{P}} \nabla f(U_{\mathcal{P}}^\top x) = \nabla F_{U_{\mathcal{P}}}(x),$$

$$\mathbb{E}\left[\|\mathsf{O}_{U_{\mathcal{P}}}(x, \xi) - \nabla F_{U_{\mathcal{P}}}(x)\|^p\right] = \mathbb{E}\left[\left\|\mathsf{O}(U_{\mathcal{P}}^\top x, \xi) - \nabla f(U_{\mathcal{P}}^\top x)\right\|^p\right] \le \sigma^p.$$

Here the equality in the second line uses that $U_{\mathcal{P}}$ is a Euclidean isometry. Finally, Lemma 2 gives

$$\mathsf{S}_{\mathcal{P}}(\nabla F_{U_{\mathcal{P}}}(x)) = \mathsf{S}_{\mathcal{P}}(U_{\mathcal{P}} \nabla f(U_{\mathcal{P}}^\top x)) \ge c_{\mathrm{sec}}\left\|\nabla f(U_{\mathcal{P}}^\top x)\right\|. \tag{151}$$

When B interacts with the selected base oracle, the history seen by A is exactly the history generated by the lifted oracle in Eq. (150). Components of x_t in $\ker U_{\mathcal{P}}^\top$ affect neither the objective gradient nor the oracle response. Thus an ambient guarantee $\mathbb{E}\left[\mathsf{S}_{\mathcal{P}}(\nabla F_{U_{\mathcal{P}}}(\widehat{x}))\right] \le \varepsilon$ would imply $\mathbb{E}\left[\left\|\nabla f(U_{\mathcal{P}}^\top \widehat{x})\right\|\right] \le \bar{\varepsilon}$, contradicting Corollary 10. This proves Eq. (17). □

E.5 Lower Bounds in Smaller Dimensions

When the dimension condition in Eq. (16) is unavailable, shortening the hard chain gives the following refinement. The logarithm reflects the dimension cost of the random rotation in Lemma E.2.

Proposition E.1 (Dimension-dependent lower bound)**.** *There are constants $c_p, C_p > 0$, depending only on p, such that the following holds for every partition. Retain the accuracy condition in Eq.* (16) *and replace its dimension requirement by*

$$d \geq C_p \log(2 + Q_\varepsilon). \tag{152}$$

Then

$$N^*(\varepsilon; \mathcal{P}) \geq c_p Q_\varepsilon \min\left\{R_\varepsilon, \frac{d}{\log(2 + dQ_\varepsilon)}\right\} = \widetilde{\Omega}\left(\min\left\{\frac{\Delta L}{\varepsilon^2}\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}, d\left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right\}\right). \tag{153}$$

Proof. Set $\bar{\varepsilon} := \varepsilon / c_{\mathrm{sec}}$ and use the lifting in the proof of Theorem 5. The corresponding values of $R_{\bar{\varepsilon}}$ and $Q_{\bar{\varepsilon}}$ differ from R_ε and Q_ε only by constants. Define

$$r_d := \left\lfloor c_p \min\left\{R_\varepsilon, \frac{d}{\log(2 + dQ_\varepsilon)}\right\}\right\rfloor. \tag{154}$$

After choosing the accuracy constant in Eq. (16) small enough and the constant in Eq. (152) large enough, we have $r_d \geq 1$. The first term in the minimum gives the gap condition $r_d \leq c_p R_{\bar{\varepsilon}}$. The second term and $\log(2 + r_d Q_\varepsilon) \leq \log(2 + dQ_\varepsilon)$ give

$$C_p r_d \log(2 + r_d Q_\varepsilon) \leq c_{\mathrm{dim}} d. \tag{155}$$

Thus the rotated instance fits into the Euclidean subspace supplied by Lemma 2. Applying Lemmas E.1 and E.2 with length r_d and then lifting as above gives at least $c_p r_d Q_\varepsilon$ calls. The floor in Eq. (154) changes only the constant, which proves Eq. (153). □

E.6 Exact Transfer for Wide Blocks

This subsection is not used in Theorems 5 and 6. For a zero-respecting method, selected blocks can carry coordinate-aligned copies of the same Euclidean instance. Given a nonempty set $I \subseteq [J]$, we call its members *active blocks*. Lemma E.3 computes the geometry of this embedding exactly. Proposition E.2 applies it under an explicit block-width condition.

Lemma E.3 (Exact diagonal geometry)**.** *Let $w_j := \sqrt{d_j / d}$ and choose a nonempty active set $I \subseteq [J]$. Put $W_I^2 := \sum_{j \in I} w_j^2$. For any integer $r \leq \min_{j \in I} d_j$, choose coordinate-aligned isometries $Q_j : \mathbb{R}^r \to \mathbb{R}^{d_j}$. Here coordinate-aligned means that each $Q_j e_i$ is a standard basis vector of $\mathbb{R}^{d_j}$. Define*

$$(Uz)^{(j)} = \frac{w_j}{W_I} Q_j z \quad (j \in I), \qquad (Uz)^{(j)} = 0 \quad (j \notin I). \tag{156}$$

Then $U^\top U = I_r$ and

$$\mathsf{S}_{\mathcal{P}}(Uz) = W_I \|z\|. \tag{157}$$

Proof. The isometry follows from $\sum_{j \in I} w_j^2 / W_I^2 = 1$. Moreover,

$$\mathsf{S}_{\mathcal{P}}(Uz) = \sum_{j \in I} w_j \frac{w_j}{W_I} \|Q_j z\| = W_I \|z\|.$$

□

By Eq. (157), an ambient target ε corresponds to the Euclidean target ε / W_I. The proposition writes d_{hard} for the dimension of that Euclidean instance and requires every active block to have at least d_{hard} coordinates.

Proposition E.2 (Exact diagonal lower bound transfer)**.** *Suppose*

$$0 < \varepsilon \le c_p W_I \min\{\sigma, \sqrt{\Delta L}\}, \qquad r_{\mathrm{hard}} := \left\lfloor c_p \frac{\Delta L W_I^2}{\varepsilon^2} \right\rfloor.$$

Apply Lemma E.1 at the base accuracy ε/W_I. The accuracy condition ensures $r_{\mathrm{hard}} \ge 1$ after reducing c_p if needed. Let $d_{\mathrm{hard}} \le C_p r_{\mathrm{hard}}$ be the actual ambient dimension of the hard instance supplied by Eq. (123)*. If*

$$\min_{j \in I} d_j \ge d_{\mathrm{hard}}, \tag{158}$$

then every randomized zero-respecting method requires

$$\Omega\left(\frac{\Delta L W_I^2}{\varepsilon^2}\left(\frac{\sigma W_I}{\varepsilon}\right)^{\frac{p}{p-1}}\right) \tag{159}$$

oracle calls.

Proof. Eq. (158) allows us to choose the isometries in Lemma E.3 with $r = d_{\mathrm{hard}}$. Lift the selected zero-chain instance through this map. The objective and oracle properties are preserved as in Eq. (150), while Eq. (157) gives exact stationarity transfer at the base accuracy ε/W_I.

It remains to verify that zero-respecting behavior is preserved. Suppose the base query is $y_t := U^\top x_t$ and the base oracle returns $q_t := \mathsf{O}(y_t, \xi_t)$. The corresponding ambient response is Uq_t. For every $i \in [d_{\mathrm{hard}}]$ and every active block j, the coordinate of Uq_t along $Q_j e_i$ is $(w_j/W_I)[q_t]_i$. Hence a base coordinate i is absent from all earlier q_s if and only if all its coordinate-aligned copies are absent from the corresponding ambient responses Uq_s. An ambient zero-respecting query x_t must then vanish on every such copy. From the formula for $U^\top$ it follows that $[y_t]_i = 0$. Therefore $\mathrm{supp}(y_t) \subseteq \bigcup_{s<t} \mathrm{supp}(q_s)$, so the simulated base method is zero-respecting. Applying Lemma E.1 at accuracy ε/W_I proves Eq. (159). □

This exact map requires wide active blocks and therefore does not cover singleton AdamW at high accuracy. The random-section construction does.

F Analysis of Muon

We prove Theorems 11 and 13 and Corollary 12 and record the scope of the comparison. All matrix inner products are Frobenius inner products.

F.1 Convergence and Parameter Choices

Let $d = mn$ and $r := \min\{m, n\}$.

Assumption 4 (Matrix objective and heavy-tailed oracle)**.** *The objective $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ is differentiable and bounded below. At iteration t, X_t is $\mathcal{F}_{t-1}$-measurable and the oracle supplies $B \in \mathbb{N}_+$ stochastic gradients $G_t^1, \dots, G_t^B \in \mathbb{R}^{m\times n}$, conditionally independent given $\mathcal{F}_{t-1}$. For some $\Delta, L > 0$, $\sigma \ge 0$, and $p \in (1, 2]$, the following hold for all $X, Y \in \mathbb{R}^{m\times n}$, $t \ge 1$, and $b \in [B]$:*

$$\begin{aligned} F(X_1) - \inf_X F(X) \le \Delta, \qquad & \|\nabla F(X) - \nabla F(Y)\|_{\mathrm{F}} \le L\|X - Y\|_{\mathrm{F}}, \\ \mathbb{E}_{t-1}[G_t^b] = \nabla F(X_t), \quad & \mathbb{E}_{t-1}\|G_t^b - \nabla F(X_t)\|_{\mathrm{F}}^p \le \sigma^p. \end{aligned} \tag{160}$$

Here $\mathcal{F}_0$ contains the initialization, hyperparameters, and internal random seed, and $\mathcal{F}_t := \sigma(\mathcal{F}_0, \{G_s^b : s \le t,\ b \in [B]\})$.

To see the equivalence with the vector setting, write $F(X) = f(\mathrm{vec}(X))$, $x_t = \mathrm{vec}(X_t)$, and $g_t^b = \mathrm{vec}(G_t^b)$. The identities $\|\mathrm{vec}(A)\|_2 = \|A\|_{\mathrm{F}}$ and $\nabla f(\mathrm{vec}(X)) = \mathrm{vec}(\nabla F(X))$ preserve the gap, smoothness, noise moments, and filtration in both directions. Thus, for $\sigma > 0$, Assumption 4 is exactly equivalent to Assumptions 1 to 3, with unchanged Δ, L, σ, p, B. Allowing $\sigma = 0$ also includes the noiseless case. For $G_t = B^{-1}\sum_{b=1}^B G_t^b$, Lemma C.1 gives the same effective noise level $\sigma_B = 2^{1/p}\sigma B^{-(p-1)/p}$ without a dimension factor.

We study the exact-sign recursion in Eq. (20). For a compact SVD $M = U_s \Sigma_s V_s^\top$ containing only the positive singular values, $\operatorname{msign}(M) = U_s V_s^\top$, with $\operatorname{msign}(0) = 0$. The proof uses this exact polar factor in place of the numerical orthogonalization in Algorithms 5 and 6.

Spectral–nuclear duality gives the dimension-normalized natural pair

$$\|D\|_{\mathrm{Muon}} := \sqrt{r}\,\|D\|_{\mathrm{op}}, \qquad \mathcal{N}(G) := \frac{\|G\|_*}{\sqrt{r}}. \tag{161}$$

This is the matrix analogue of AdamW's primal norm $\sqrt{d}\|u\|_\infty$ and dual criterion $\|g\|_1/\sqrt{d}$. In both cases the normalized dual norm is bounded by the Euclidean/Frobenius norm. They are nevertheless different stationarity criteria adapted to the respective updates.

We use the following two exponents throughout this appendix,

$$a_p := \frac{p-1}{3p-2}, \qquad b_p := \frac{p}{3p-2}. \tag{162}$$

Theorem 11 (Complete version of Theorem 8)**.** *Suppose $F\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ and its stochastic oracle satisfy Eq.* (160) *with $p \in (1,2]$, $\Delta, L > 0$, and $\sigma \geq 0$. Fix integers $B, T \geq 1$ and average B conditionally independent oracle samples at each iteration. Let $r = \min\{m,n\}$ and $\sigma_B = 2^{1/p}\sigma B^{-(p-1)/p}$. Run Eq.* (20) *from $M_0 = 0$ with constant hyperparameters*

$$\delta = \min\left\{1, \left(\frac{\Delta L}{\sigma_B^2 T}\right)^{\frac{p}{3p-2}}\right\}, \qquad \beta = 1-\delta, \qquad \eta = \sqrt{\frac{2\Delta\delta}{5rLT}}, \tag{163}$$

where $\delta = 1$ when $\sigma_B = 0$. With the exponents in Eq. (162)*, the iterates satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\mathcal{N}(\nabla F(X_t)) \leq 10\left[\sqrt{\frac{\Delta L}{T}} + (\Delta L)^{a_p}\sigma_B^{b_p}T^{-a_p} + (\Delta L)^{2a_p-1/2}\sigma_B^{2b_p}T^{-2a_p}\right]. \tag{164}$$

Proof overview. Writing $O_t = \operatorname{msign}(M_t)$, spectral–nuclear duality gives

$$\langle \nabla F(X_t), O_t\rangle \geq \|\nabla F(X_t)\|_* - 2\|M_t - \nabla F(X_t)\|_*.$$

Frobenius smoothness and $\|O_t\|_{\mathrm{F}}^2 \leq r$ then give descent. Decomposing the momentum error into initialization, a discounted martingale, and trajectory drift reduces its control to the Hilbert-space pth-moment inequality behind Lemma C.1. The conversion $\|A\|_* \leq \sqrt{r}\|A\|_{\mathrm{F}}$ is absorbed by the normalization in Eq. (161). The same argument extends to collections of matrix blocks, as shown in Appendix F.6.

The alternative momentum convention $B_t = \beta B_{t-1} + G_t$, $B_0 = 0$, gives $M_t = \delta B_t$ and hence the same update direction. Likewise, positive scalar bias correction of M_t preserves its matrix sign.

Lemma F.1 (A Hilbert-space martingale moment bound)**.** *Let $(Z_k)_{k=1}^t$ be a martingale-difference sequence in a Hilbert space with finite pth moments, $p \in (1,2]$. For deterministic scalars w_k,*

$$\mathbb{E}\left\|\sum_{k=1}^{t} w_k Z_k\right\|^p \leq 2\sum_{k=1}^{t}|w_k|^p\mathbb{E}\|Z_k\|^p. \tag{165}$$

Proof. Apply Eq. (25) to $u = \sum_{j<k} w_j Z_j$ and $v = w_k Z_k$ and condition on the history before Z_k is observed. The inner product term has zero conditional expectation. Iterating from the zero partial sum proves Eq. (165). Thus the proof of Lemma C.1 also applies across time under conditional centering. □

The descent argument in Step 1 follows the polar-alignment analysis of Shen et al. (2026); Yu et al. (2026d); Choudhury et al. (2026). The momentum analysis in Step 2 follows their separation of gradient drift from accumulated noise, together with the heavy-tailed martingale estimate of Hübler et al. (2026). We apply these arguments under the common Frobenius oracle and retain the initialization term explicitly.

Proof of Theorem 11. Write $\nabla_t := \nabla F(X_t)$ and $O_t := \operatorname{msign}(M_t)$. By the compact-SVD definition, for every matrix A,

$$\langle A, \operatorname{msign}(A)\rangle = \|A\|_*, \qquad \|\operatorname{msign}(A)\|_{\mathrm{op}} \leq 1, \qquad \|\operatorname{msign}(A)\|_{\mathrm{F}}^2 \leq r. \tag{166}$$

These identities also hold for rank-deficient A and for $A = 0$.

Step 1. Descent. Spectral–nuclear duality and the triangle inequality give

$$\langle \nabla_t, O_t \rangle = \langle M_t, O_t \rangle + \langle \nabla_t - M_t, O_t \rangle \geq \|M_t\|_* - \|\nabla_t - M_t\|_* \geq \|\nabla_t\|_* - 2\|M_t - \nabla_t\|_*.$$

Frobenius smoothness then yields

$$F(X_{t+1}) \leq F(X_t) - \eta \|\nabla_t\|_* + 2\eta \|M_t - \nabla_t\|_* + \frac{Lr\eta^2}{2}.$$

Sum over t, take expectations, and divide by $\eta T\sqrt{r}$ to obtain

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\mathcal{N}(\nabla_t) \leq \frac{\Delta}{\eta T\sqrt{r}} + \frac{L\eta\sqrt{r}}{2} + \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}\mathcal{N}(M_t - \nabla_t). \tag{167}$$

Step 2. Momentum tracking. Let $N_t := G_t - \nabla_t$ and $D_t := \nabla_{t-1} - \nabla_t$ for $t \geq 2$. Unrolling the recursion with $M_0 = 0$ gives

$$M_t - \nabla_t = -\beta^t \nabla_1 + \delta \sum_{k=1}^{t} \beta^{t-k} N_k + \sum_{k=2}^{t} \beta^{t-k+1} D_k. \tag{168}$$

We use $\mathcal{N}(A) \leq \|A\|_{\mathrm{F}}$ for all three terms. First, lower boundedness and L-smoothness imply $\|\nabla_1\|_{\mathrm{F}}^2 \leq 2L\Delta$, by evaluating the smoothness bound at $X_1 - \nabla_1/L$. Consequently,

$$\frac{1}{T}\sum_{t=1}^{T} \beta^t \mathcal{N}(\nabla_1) \leq \frac{\sqrt{2L\Delta}}{T\delta}.$$

Second, (N_k) is a martingale-difference sequence for the natural filtration, and Eq. (23) holds in Frobenius norm. Applying Lemma F.1 with $w_k = \delta\beta^{t-k}$ gives

$$\mathbb{E}\left\| \delta \sum_{k=1}^{t} \beta^{t-k} N_k \right\|_{\mathrm{F}}^{p} \leq 2\sigma_B^p \delta^p \sum_{k=1}^{t} \beta^{p(t-k)} \leq \frac{2\sigma_B^p \delta^p}{1-\beta^p} \leq 2\sigma_B^p \delta^{p-1}. \tag{169}$$

The last inequality uses $1 - \beta^p \geq 1 - \beta = \delta$. Jensen's inequality therefore bounds the expected normalized nuclear norm of this term by $2^{1/p}\sigma_B \delta^{(p-1)/p}$. Finally,

$$\|D_k\|_{\mathrm{F}} \leq L\|X_k - X_{k-1}\|_{\mathrm{F}} \leq L\eta\sqrt{r}, \qquad \mathcal{N}\left(\sum_{k=2}^{t} \beta^{t-k+1} D_k\right) \leq \frac{L\eta\sqrt{r}}{\delta}.$$

Combining the three terms in Eq. (168),

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\mathcal{N}(M_t - \nabla_t) \leq \frac{\sqrt{2L\Delta}}{T\delta} + 2^{1/p}\sigma_B \delta^{(p-1)/p} + \frac{L\eta\sqrt{r}}{\delta}. \tag{170}$$

Step 3. Parameter choices for the two momentum cases. Insert Eq. (170) into Eq. (167) and use $\delta \leq 1$ to obtain

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\mathcal{N}(\nabla_t) \leq \frac{\Delta}{\eta T\sqrt{r}} + \frac{5L\eta\sqrt{r}}{2\delta} + \frac{2\sqrt{2L\Delta}}{T\delta} + 2^{1+1/p}\sigma_B \delta^{(p-1)/p}. \tag{171}$$

The learning rate in Eq. (163) balances the first two terms, so

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\mathcal{N}(\nabla_t) \leq \sqrt{10}\sqrt{\frac{\Delta L}{T\delta}} + \frac{2\sqrt{2L\Delta}}{T\delta} + 2^{1+1/p}\sigma_B \delta^{(p-1)/p}. \tag{172}$$

If $\delta = 1$, then either the oracle is noiseless or $\sigma_B \leq \sqrt{\Delta L/T}$. Since $T \geq 1$, the entire right-hand side is $O(\sqrt{\Delta L/T})$. If $\delta < 1$, direct substitution with a_p, b_p from Eq. (162) gives

$$\sqrt{\frac{\Delta L}{T\delta}} = \sigma_B \delta^{(p-1)/p} = (\Delta L)^{a_p} \sigma_B^{b_p} T^{-a_p}, \quad \frac{\sqrt{\Delta L}}{T\delta} = (\Delta L)^{2a_p - 1/2} \sigma_B^{2b_p} T^{-2a_p}.$$

Since $\sqrt{10} + 2\sqrt{2} + 2^{1+1/p} \leq 10$, the two branches give Eq. (164). The bound Eq. (21) follows. □

F.2 Complexity

Corollary 12 (Complexity of Muon)**.** *Under the assumptions and tuning of Theorem 11, suppose* $0 < \varepsilon \le c_p \min\{\sigma_B, \sqrt{\Delta L}\}$*. A uniformly sampled iterate* X_τ*, with* τ *independent of the run, satisfies* $\mathbb{E}\mathcal{N}(\nabla F(X_\tau)) \le \varepsilon$ *for*

$$T = \left\lceil C_p \frac{\Delta L}{\varepsilon^2} \left(\frac{\sigma_B}{\varepsilon}\right)^{\frac{p}{p-1}} \right\rceil, \qquad BT = O\left(\frac{\Delta L}{\varepsilon^2} \left(\frac{\sigma}{\varepsilon}\right)^{\frac{p}{p-1}}\right). \tag{173}$$

Proof of Corollary 12. Set $s_p := p/(p-1)$ and choose $T = \lceil K_p(\Delta L/\varepsilon^2)(\sigma_B/\varepsilon)^{s_p} \rceil$. Recall a_p, b_p from Eq. (162). Since $2 + s_p = 1/a_p$ and $s_p a_p = b_p$, the three terms in Eq. (164) satisfy

$$\sqrt{\frac{\Delta L}{T}} \le K_p^{-1/2} \varepsilon \left(\frac{\varepsilon}{\sigma_B}\right)^{s_p/2},$$
$$(\Delta L)^{a_p} \sigma_B^{b_p} T^{-a_p} \le K_p^{-a_p} \varepsilon,$$
$$(\Delta L)^{2a_p - 1/2} \sigma_B^{2b_p} T^{-2a_p} \le K_p^{-2a_p} \frac{\varepsilon^2}{\sqrt{\Delta L}}.$$

Under the stated accuracy condition, a sufficiently large K_p makes their sum, including the factor 10 in Eq. (164), at most ε. Independent uniform selection of τ converts the average bound into the output guarantee. For the individual-sample cost, the ceiling gives

$$BT \le B + K_p 2^{1/(p-1)} R_\varepsilon Q_\varepsilon, \tag{174}$$

since $B\sigma_B^{s_p} = 2^{1/(p-1)}\sigma^{s_p}$. Moreover, $\varepsilon \le c_p\sigma_B$ implies $B \le 2^{1/(p-1)} c_p^{s_p} Q_\varepsilon$, while $\varepsilon \le c_p\sqrt{\Delta L}$ implies $R_\varepsilon \ge c_p^{-2}$. Thus the additive B is absorbed into $O(R_\varepsilon Q_\varepsilon)$. □

For a raw nuclear target $\varepsilon_{\mathrm{raw}}$, substituting $\varepsilon = \varepsilon_{\mathrm{raw}}/\sqrt{r}$ gives the leading upper complexity

$$O\left(r^{\frac{3p-2}{2(p-1)}} \frac{\Delta L}{\varepsilon_{\mathrm{raw}}^2} \left(\frac{\sigma}{\varepsilon_{\mathrm{raw}}}\right)^{\frac{p}{p-1}}\right)$$

in the corresponding stochastic regime. At $p = 2$ the factor is r^2, as in Shen et al. (2026).

F.3 Exact Lower Bound Transfer

Let $N^*_{\mathrm{nuc}}(\varepsilon; m, n)$ denote the minimax complexity in Eq. (15), with the corresponding matrix oracle class and $\mathcal{N}$ in place of $\mathsf{S}_\mathcal{P}$. The class still includes all randomized stochastic first-order methods, and the same output-query cost convention applies. Recall R_ε and Q_ε from Theorem 5.

Theorem 13 (Lower bound in Muon's normalized natural geometry)**.** *There are constants* $c_p, C_p > 0$ *such that, if*

$$0 < \varepsilon \le c_p \min\{\sigma, \sqrt{\Delta L}\}, \qquad r \ge C_p R_\varepsilon \log(2 + R_\varepsilon Q_\varepsilon), \tag{175}$$

then $N^*_{\mathrm{nuc}}(\varepsilon; m, n) \ge c_p R_\varepsilon Q_\varepsilon$*. More generally, retain the accuracy condition but require only* $r \ge C_p \log(2 + Q_\varepsilon)$*. Then*

$$N^*_{\mathrm{nuc}}(\varepsilon; m, n) \ge c_p Q_\varepsilon \min\left\{R_\varepsilon, \frac{r}{\log(2 + rQ_\varepsilon)}\right\}. \tag{176}$$

The transfer is exact. Let $D : \mathbb{R}^r \to \mathbb{R}^{m\times n}$ place a vector on the first r diagonal entries, and let $P = D^*$ extract them. Then $\|Dz\|_{\mathrm{F}} = \|z\|_2$ and $\|Dz\|_* = \|z\|_1$. For a singleton-block hard instance (f, O) of Theorem 5, set

$$\widetilde{F}(X) := f(PX), \qquad \widetilde{\mathsf{O}}(X, \xi) := D\mathsf{O}(PX, \xi). \tag{177}$$

The lifted pair preserves the gap, smoothness bound, unbiasedness, and pth noise moment bound. Its stationarity satisfies

$$\mathcal{N}(\nabla\widetilde{F}(X)) = \frac{\|\nabla f(PX)\|_1}{\sqrt{r}}.$$

Any matrix method can be simulated by querying the vector oracle at PX. The dimension requirement is on r.

The transfer follows earlier Muon lower bound constructions of Hübler et al. (2026); Tao et al. (2026). In particular, Tao et al. (2026) use a diagonal embedding to transfer vector ℓ_1 lower bounds to nuclear stationarity, while Hübler et al. (2026) embed heavy-tailed vector hard instances into matrix problems. Here the base instance is Theorem 5, and the simulation preserves the common Frobenius oracle for arbitrary randomized first-order methods.

Proof of Theorem 13. Let $D : \mathbb{R}^r \to \mathbb{R}^{m\times n}$ be the diagonal embedding in Eq. (177) and $P = D^*$. We have $PD = I_r$, $\|Dz\|_{\mathrm{F}} = \|z\|_2$, and $\|PX\|_2 \le \|X\|_{\mathrm{F}}$. Take an arbitrary randomized matrix method A with cost at most N and simulate it using a vector oracle in $\mathbb{R}^r$. Run the same internal seed, query at PX_t whenever A queries X_t, and return the embedded answer $D\mathsf{O}(PX_t, \xi_t)$ to A. When A outputs $\widehat{X}$, the vector method outputs $P\widehat{X}$. This simulation is measurable with respect to the vector history and uses independent oracle seeds exactly as in Section 4. Its cost is no larger. A previously queried matrix output projects to a previously queried vector, whereas an unqueried matrix output already allows the extra call counted in the cost convention. Projection can only reduce that extra cost.

Apply the singleton-block case of Theorem 5 in dimension r to this simulated method and let (f, O) be its hard instance. Define $(\widetilde{F}, \widetilde{\mathsf{O}})$ by Eq. (177). Surjectivity of P preserves the initial gap,

$$\widetilde{F}(0) - \inf_X \widetilde{F}(X) = f(0) - \inf_z f(z) \le \Delta.$$

Since $\nabla\widetilde{F}(X) = D\nabla f(PX)$, we have

$$\|\nabla\widetilde{F}(X) - \nabla\widetilde{F}(Y)\|_{\mathrm{F}} = \|\nabla f(PX) - \nabla f(PY)\|_2 \le L\|P(X-Y)\|_2 \le L\|X-Y\|_{\mathrm{F}}.$$

For every matrix query X, the lifted oracle obeys

$$\mathbb{E}_\xi\widetilde{\mathsf{O}}(X,\xi) = D\nabla f(PX) = \nabla\widetilde{F}(X),$$
$$\mathbb{E}_\xi\|\widetilde{\mathsf{O}}(X,\xi) - \nabla\widetilde{F}(X)\|_{\mathrm{F}}^p = \mathbb{E}_\xi\|\mathsf{O}(PX,\xi) - \nabla f(PX)\|_2^p \le \sigma^p.$$

Finally, the singular values of Dz are $|z_1|, \dots, |z_r|$, and hence

$$\mathcal{N}(\nabla\widetilde{F}(X)) = \frac{\|D\nabla f(PX)\|_*}{\sqrt{r}} = \frac{\|\nabla f(PX)\|_1}{\sqrt{r}}. \tag{178}$$

The matrix history is exactly the one produced by the lifted oracle. Off-diagonal components of queries affect neither the objective nor its oracle response. Thus a matrix guarantee better than the claimed lower bound would contradict the singleton-block vector lower bound. Applying Eq. (17) with $d = r$ proves the full-dimensional branch, while Eq. (153), with its admissibility requirement $r \ge C_p \log(2+Q_\varepsilon)$, proves the finite-dimensional branch. No zero-respecting restriction is imposed on A. □

F.4 Conditions for Comparing AdamW and Muon

The common oracle has identical Δ, L, σ, p, but the accuracy and dimension requirements of the two algorithmic guarantees need not coincide. For $B = 1$, the singleton-block specializations of Algorithms 1 and 2 use $d = mn$ and $d_{\min} = 1$ in Eqs. (6) and (12), respectively. The full Muon lower bound instead needs Eq. (175) with $r = \min\{m, n\}$. The latter dimension condition implies the AdamW lower bound dimension condition, since $d \ge r$, but does not imply AdamW's upper bound accuracy condition.

To see the distinction, both AdamW accuracy conditions contain, up to p-dependent constants, the restriction

$$\varepsilon \lesssim \sigma d^{-\frac{p-1}{2p}}, \qquad \text{or equivalently} \qquad d \lesssim Q_\varepsilon^2.$$

Because $d = mn \ge r^2$, a necessary compatibility condition for these sufficient upper and lower regimes is

$$R_\varepsilon \log(2 + R_\varepsilon Q_\varepsilon) \lesssim Q_\varepsilon. \tag{179}$$

The exact accuracy requirements, including the bias-correction transient, remain those in the main theorems. For square matrices and $1 < p < 2$, the ratio $Q_\varepsilon/R_\varepsilon$ grows polynomially as ε decreases with other problem parameters fixed, so compatible sizes exist at sufficiently small accuracy. At $p = 2$, however, this ratio is $\sigma^2/(\Delta L)$, so compatibility also constrains this ratio relative to the logarithm.

Under these joint conditions, Table 3 gives the same minimax complexity for AdamW and Muon.

Scope of the comparison. The matching bounds remove a distinction based solely on AdamW lacking the optimal heavy-tailed exponent in T under the common oracle. This is the scope of the refutation in Section 5, not a claim that the methods have identical full rates or performance on the same instance. Tail-dependent constants, transients, tuning, dimension, and interactions between noise and geometry can still change their relative performance, even after conversion to a common norm (Appendices F.5 and F.7). Heavy-tailed noise can therefore still contribute to practical differences. A fuller comparison requires theoretical analysis and controlled experiments that isolate these effects, which we leave to future work.

F.5 Relation to Existing Muon Theory

Three settings should be distinguished. Shen et al. (2026) analyze Muon under Frobenius smoothness and bounded-variance noise, recovering the $\sqrt{r}$ factor in raw nuclear stationarity. Yu et al. (2026d) allow heavy-tailed noise under generalized matrix smoothness and noise conditions, which include the Euclidean specialization considered here. Hübler et al. (2026) instead impose smoothness in a primal/dual norm pair and control noise directly in the nuclear norm for Muon, obtaining sharper native-geometry dimension dependence. These are different oracle classes when their constants are fixed. Complementary work treats Nesterov momentum and inexact polar decompositions (Choudhury et al., 2026). Section 5 uses matching bounds under a shared Euclidean oracle to interpret our AdamW result. These bounds in normalized natural geometries complement the matrix-adapted analyses above.

Recall the exponents a_p, b_p in Eq. (162). To specialize Yu et al. (2026d), transpose if needed so that $m = r \le n$ and put $E = [I_r\ 0]$. In their notation, take $\mathbf{L}_0 = LE$, $\mathbf{V}_0 = \sigma E/\sqrt{r}$, and $\mathbf{L}_1 = \mathbf{V}_1 = 0$. Their weighted smoothness and noise conditions then reduce to Eq. (160). Since $\|\mathbf{L}_0\|_* = rL$ and $\|\mathbf{V}_0\|_* = \sqrt{r}\sigma$, their leading stochastic factor at $B = 1$ is

$$(\Delta\|\mathbf{L}_0\|_*)^{a_p}\|\mathbf{V}_0\|_*^{b_p}T^{-a_p} = \sqrt{r}\,(\Delta L)^{a_p}\sigma^{b_p}T^{-a_p}.$$

Division by $\sqrt{r}$ recovers our leading normalized rate. Eq. (164) also retains the deterministic and initialization terms explicitly.

Hübler et al. (2026) control noise directly in nuclear norm and smoothness in the spectral/nuclear primal–dual pair. In contrast, our oracle controls Frobenius quantities. Let $\sigma_{*,p}$ and $\sigma_{\mathrm{F},p}$ be the minimal uniform nuclear and Frobenius noise moment bounds, respectively. Worst-case norm conversion only guarantees

$$\sigma_{*,p} \le \sqrt{r}\,\sigma_{\mathrm{F},p}, \qquad L_{\mathrm{nat}} \le rL,$$

where L_{nat} is the best constant in $\|\nabla F(X) - \nabla F(Y)\|_* \le L_{\mathrm{nat}}\|X - Y\|_{\mathrm{op}}$. Favorable native constants contain geometric information not supplied by the common Frobenius bounds. Indeed, Hübler et al. (2026) report nuclear-to- Frobenius noise ratios well below the worst-case $\sqrt{r}$ conversion. The common-oracle comparison matches the leading T-exponents without determining these potentially sharper native constants.

Mini-batching reduces iterations without changing the leading sample complexity in this regime. At $p = 2$ and $B = 1$, the leading raw nuclear term is $O(\sqrt{r}(\Delta L\sigma^2/T)^{1/4})$, agreeing with Corollary 4.2 of Shen et al. (2026).

F.6 Muon with Multiple Matrix Blocks

Let $X = (X^{(1)}, \dots, X^{(K)})$ with $X^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$, and equip the product space with $\|X\|_{\mathrm{F,prod}}^2 := \sum_\ell \|X^{(\ell)}\|_{\mathrm{F}}^2$. Assume L-smoothness and the pth noise moment bound in this product norm, which is exactly the Euclidean norm after concatenated vectorization. Put $r_\ell := \min\{m_\ell, n_\ell\}$ and $r_\Sigma :=$

$\sum_{\ell=1}^{K} r_\ell$. Apply Eq. (20) separately to every matrix with common η and β. The natural dual pair becomes

$$\|D\|_{\mathrm{Muon,prod}} := \sqrt{r_\Sigma} \max_\ell \|D^{(\ell)}\|_{\mathrm{op}}, \qquad \mathcal{N}^{\mathrm{prod}}(G) := \frac{1}{\sqrt{r_\Sigma}} \sum_{\ell=1}^{K} \|G^{(\ell)}\|_*. \tag{180}$$

Spectral–nuclear duality holds blockwise, and Cauchy–Schwarz gives

$$\mathcal{N}^{\mathrm{prod}}(G) \le \frac{1}{\sqrt{r_\Sigma}} \sum_\ell \sqrt{r_\ell} \|G^{(\ell)}\|_{\mathrm{F}} \le \|G\|_{\mathrm{F,prod}}.$$

Moreover, $\sum_\ell \| \operatorname{msign}\left(M_t^{(\ell)}\right) \|_{\mathrm{F}}^2 \le r_\Sigma$. These are exactly the two geometric bounds used in the upper bound proof. Thus Eqs. (164) and (173) hold with $\mathcal{N}^{\mathrm{prod}}$ and with r_Σ replacing r in the learning rate. The normalized rate is unchanged.

For the lower bound, embed a vector of dimension r_Σ along the collected block diagonals. This preserves the product Frobenius norm and maps the sum of block nuclear norms to the vector ℓ_1 norm. The proof of Theorem 13 therefore applies with r_Σ in place of r, under the corresponding dimension conditions.

F.7 Comparing the Natural Stationarity Measures

Takeaway. *For a common Frobenius target, the norm conversion favors Muon on diagonal gradients and AdamW on flat rank-one gradients.* Both entry concentration and the singular spectrum determine the comparison.

Table 3 gives the same minimax complexity for two different stationarity criteria. To compare their strength, evaluate both at the same matrix $G \in \mathbb{R}^{m\times n}$. Put $d = mn$, $r = \min\{m,n\}$, $q = \max\{m,n\}$, and write

$$\mathcal{A}(G) := \frac{\|\operatorname{vec}(G)\|_1}{\sqrt{d}}, \qquad \mathcal{N}(G) = \frac{\|G\|_*}{\sqrt{r}}. \tag{181}$$

Here $\mathcal{A}$ is AdamW's singleton-partition criterion. For $G \neq 0$, define the effective number of nonzero entries and the nuclear effective rank

$$s(G) := \frac{\|\operatorname{vec}(G)\|_1^2}{\|G\|_{\mathrm{F}}^2}, \qquad \rho(G) := \frac{\|G\|_*^2}{\|G\|_{\mathrm{F}}^2}. \tag{182}$$

They satisfy $1 \le s(G) \le |\operatorname{supp}(G)| \le d$ and $1 \le \rho(G) \le \operatorname{rank}(G) \le r$. Thus $s(G)/d$ measures how broadly the entry magnitudes are distributed, while $\rho(G)/r$ measures how broadly the singular values are distributed.

Proposition F.1 (Exact comparison of natural stationarity)**.** *For every nonzero $G \in \mathbb{R}^{m\times n}$,*

$$\frac{\mathcal{A}(G)}{\mathcal{N}(G)} = \sqrt{\frac{r\, s(G)}{d\, \rho(G)}}, \qquad \frac{1}{\sqrt{q}} \le \frac{\mathcal{A}(G)}{\mathcal{N}(G)} \le \sqrt{r}. \tag{183}$$

Both inequalities are sharp. In particular, $\mathcal{A}(G) \ge \mathcal{N}(G)$ if and only if $s(G)/d \ge \rho(G)/r$.

Proof. The equality follows from Eqs. (181) and (182). Decomposing G into coordinate matrices and using the triangle inequality for nuclear norm gives $\|G\|_* \le \|\operatorname{vec}(G)\|_1$. Cauchy–Schwarz gives $\|\operatorname{vec}(G)\|_1 \le \sqrt{d}\|G\|_{\mathrm{F}} \le \sqrt{d}\|G\|_*$. These inequalities prove the two bounds in Eq. (183). A matrix with one nonzero entry attains the lower bound. A rank-one matrix with equal absolute values in all entries attains the upper bound. □

Comparison at a fixed gradient. The smaller criterion reaches a numerical threshold more easily. AdamW's criterion is smaller when entry magnitudes are more concentrated relative to the singular values, as quantified by $s(G)/d < \rho(G)/r$. Muon's criterion is smaller when $s(G)/d > \rho(G)/r$. The nuclear norm is invariant under orthogonal changes of matrix bases, whereas the entrywise ℓ_1 norm depends on those bases. Entry orientation therefore matters in addition to rank.

Comparison at a common Frobenius target. At the same natural stationarity rate, a larger criterion gives a stronger bound on the Frobenius gradient norm. The exact conversion factors are

$$\mathcal{A}(G) = \sqrt{\frac{s(G)}{d}}\,\|G\|_{\mathrm{F}}, \qquad \mathcal{N}(G) = \sqrt{\frac{\rho(G)}{r}}\,\|G\|_{\mathrm{F}}. \tag{184}$$

For convergence guarantees, these factors must hold uniformly over the gradients encountered by the algorithms. Suppose those gradients belong to a class $\mathcal{H}$ containing a nonzero matrix, and define

$$c_A := \inf_{G\in\mathcal{H}\setminus\{0\}} \sqrt{s(G)/d}, \qquad c_N := \inf_{G\in\mathcal{H}\setminus\{0\}} \sqrt{\rho(G)/r}. \tag{185}$$

If the respective output gradients G_A, G_N satisfy $\mathbb{E}\mathcal{A}(G_A) \le u_T$ and $\mathbb{E}\mathcal{N}(G_N) \le u_T$, then

$$\mathbb{E}\|G_A\|_{\mathrm{F}} \le u_T/c_A, \qquad \mathbb{E}\|G_N\|_{\mathrm{F}} \le u_T/c_N. \tag{186}$$

Hence the conversion favors AdamW when $c_A > c_N$ and Muon when $c_N > c_A$. In the stochastic regime of Table 3, these conversions multiply the leading complexity upper bounds by $c_A^{-(3p-2)/(p-1)}$ and $c_N^{-(3p-2)/(p-1)}$, respectively, with the original theorem conditions applied to $c_A\varepsilon$ and $c_N\varepsilon$. This compares the strength of the existing guarantees after converting them to a common target.

Diagonal gradients favor Muon in the common target. For a rectangular diagonal matrix, nuclear and entrywise ℓ_1 norms coincide, so $\mathcal{N}(G) = \sqrt{q}\,\mathcal{A}(G)$. For example, if $k \le r$ diagonal entries have equal nonzero magnitude, then $s(G) = \rho(G) = k$ and

$$\mathcal{A}(G) = \sqrt{k/d}\,\|G\|_{\mathrm{F}}, \qquad \mathcal{N}(G) = \sqrt{k/r}\,\|G\|_{\mathrm{F}}.$$

On any class of diagonal gradients, $c_N = \sqrt{q}\,c_A$. The same native rate therefore gives Muon a Frobenius guarantee stronger by the factor $\sqrt{q}$ in the norm conversion.

Flat rank-one gradients favor AdamW in the common target. Let $G = \gamma uv^\top$, where $\gamma \neq 0$, $|u_i| = m^{-1/2}$, and $|v_j| = n^{-1/2}$. Then $s(G) = d$, $\rho(G) = 1$, and

$$\mathcal{A}(G) = \|G\|_{\mathrm{F}}, \qquad \mathcal{N}(G) = \|G\|_{\mathrm{F}}/\sqrt{r}.$$

For this class, $c_A = 1$ and $c_N = 1/\sqrt{r}$, giving AdamW the stronger Frobenius conversion by $\sqrt{r}$. A single-entry matrix also has rank one but attains the opposite extreme in Eq. (183). Together, these examples show why both entry concentration and the singular spectrum are needed to interpret the comparison. When $s(G)/d = \rho(G)/r$, the two stationarity measures coincide.

G Experimental Details for LLMs

This section describes the experimental details for producing Figures 1a and 1b in the main text. Codes can be found in the supplementary material.

G.1 Setup

We study gradient noise in Pythia-6.9B (Biderman et al., 2023), which has 6.86×10^9 parameters, at `step0`, `step72000`, and `step143000`. These checkpoints cover the initial, middle, and final stages of training. Model weights and per-sample gradients use BF16. We accumulate reference gradients in FP32 and scalar norms in FP64. The model runs in `eval()` mode with `use_cache=False`, without parameter updates or gradient clipping.

We use Pythia's standard non-deduplicated, pre-tokenized Pile, `pile-standard-pythia-preshuffled`. Each record contains 2,049 tokens. The first 2,048 are the causal-LM inputs and the last 2,048 are the shifted targets. For a record s_i,

$$g_i := \nabla_\theta \ell(\theta; s_i)$$

is the gradient of the mean token loss with respect to all participating model parameters.

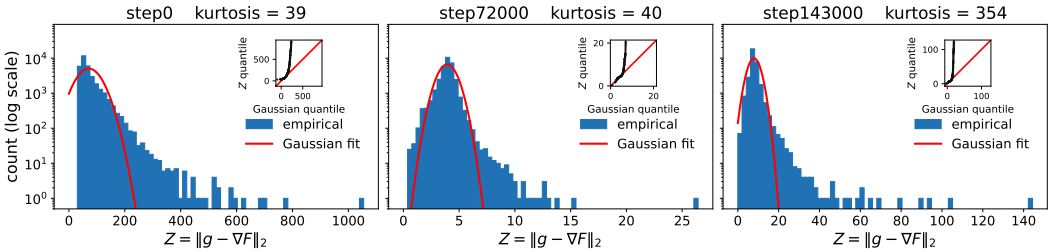


Figure 2: Gradient deviations $Z_i = \|g_i - \widehat{\nabla F}_R\|_2$ at the initial, middle, and final checkpoints of Pythia-6.9B pretraining, extending Figure 1a. Insets compare the data with fitted Gaussians. The centered sample kurtoses are 39, 40, and 354, compared with **3** for a Gaussian. Fitting details are in Appendix G.2.

At each checkpoint, we estimate the population gradient from an independent reference set of $R = 131{,}072$ records,

$$\widehat{\nabla F}_R := \frac{1}{R}\sum_{r=1}^{R} g_r.$$

We accumulate sums locally in FP32 and combine them with one final bucketed NCCL all-reduce. A disjoint set of $P = 32{,}768$ batch-size-one probes gives

$$\xi_i := g_i - \widehat{\nabla F}_R, \qquad Z_i := \|\xi_i\|_2 .$$

Computing the norms parameter by parameter in FP32, with FP64 sums, avoids storing full probe gradients. We reuse the same reference and probe indices at all three checkpoints. Halving the reference set has little effect on the estimated center. At every checkpoint, $\left\|\widehat{\nabla F}_{64\mathrm{k}} - \widehat{\nabla F}_{131\mathrm{k}}\right\| / \widehat{s}_Z < 0.004$, where $\widehat{s}_Z = (P^{-1}\sum_i Z_i^2)^{1/2}$.

G.2 Gradient Noise Distributions

Figures 2 and 3 extend Figures 1a and 1b to all three checkpoints. Both diagnostics use the same P radial deviations at each checkpoint.

We plot histogram counts on a logarithmic scale and estimate the mean and variance by

$$\overline{Z} := \frac{1}{P}\sum_{i=1}^{P} Z_i, \qquad \widehat{v}_Z := \frac{1}{P}\sum_{i=1}^{P}(Z_i - \overline{Z})^2.$$

Following Garg et al. (2021), we use the centered Pearson sample kurtosis as a descriptive tail diagnostic,

$$\widehat{\kappa}_Z := \frac{P^{-1}\sum_{i=1}^{P}(Z_i - \overline{Z})^4}{\left(P^{-1}\sum_{i=1}^{P}(Z_i - \overline{Z})^2\right)^2}. \tag{187}$$

The three checkpoints give kurtoses of 39, 40, and 354, compared with 3 for a scalar Gaussian. In each histogram, the red curve is the density of $\mathcal{N}(\overline{Z}, \widehat{v}_Z)$ scaled to histogram counts. Writing Φ_{N} for the standard normal distribution function, the Q–Q inset compares the ordered values $Z_{(i)}$ with

$$\overline{Z} + \sqrt{\widehat{v}_Z}\,\Phi_{\mathrm{N}}^{-1}\left(\frac{i - 1/2}{P}\right), \qquad i \in [P].$$

The empirical survival function in Figure 3 is

$$\widehat{S}(z) := \frac{1}{P}\sum_{i=1}^{P}\mathbb{I}_{\{Z_i > z\}}.$$

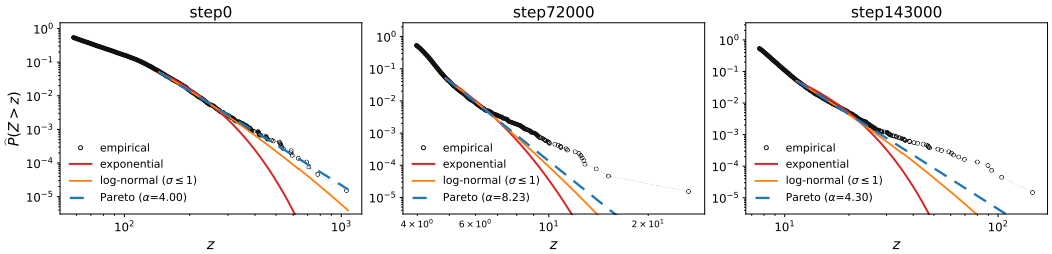


Figure 3: Empirical survival curves on log-log axes at all three checkpoints, extending Figure 1b. The exponential, lognormal, and Pareto references are fitted above the empirical 95th percentile and normalized to agree there. The empirical tails decay more slowly than all three fitted references. Fitting details are in Appendix G.2.

We fit each reference to observations above the empirical 95th percentile u, with $n_u := \sum_i \mathbb{I}_{\{Z_i > u\}}$. All curves are normalized to equal $\widehat{S}(u)$ at u. Maximum-likelihood fitting of the exponential exceedances gives

$$\widehat{\lambda} := \left(\frac{1}{n_u} \sum_{i:Z_i > u} (Z_i - u) \right)^{-1}, \qquad \widehat{S}_{\exp}(z) := \widehat{S}(u) e^{-\widehat{\lambda}(z-u)}.$$

The Pareto exponent and survival curve are

$$\widehat{\alpha} := \frac{n_u}{\sum_{i:Z_i > u} \log(Z_i/u)}, \qquad \widehat{S}_{\mathrm{par}}(z) := \widehat{S}(u) \left(\frac{z}{u} \right)^{-\widehat{\alpha}}.$$

The fitted exponents are 4.00, 8.23, and 4.30 in checkpoint order. For the lognormal reference, we maximize the conditional likelihood of the exceedances with the log-scale standard deviation constrained to at most one. With fitted parameters $(\widehat{\mu}, \widehat{\tau})$, the curve is

$$\widehat{S}_{\ln}(z) := \widehat{S}(u) \frac{1 - \Phi_{\mathrm{N}}((\log z - \widehat{\mu})/\widehat{\tau})}{1 - \Phi_{\mathrm{N}}((\log u - \widehat{\mu})/\widehat{\tau})}, \qquad 0 < \widehat{\tau} \le 1.$$

G.3 Mean Excess

The mean-excess function

$$e(u) := \mathbb{E}\left[Z - u \mid Z > u \right]$$

measures how far observations above u exceed that threshold on average. For an exponential distribution with rate λ, memorylessness gives the constant $e(u) = 1/\lambda$. An increasing curve is a standard diagnostic for slower tail decay (Embrechts & Schmidli, 1994). We estimate it by

$$\widehat{e}(u) := \frac{\sum_{i=1}^{P} (Z_i - u) \mathbb{I}_{\{Z_i > u\}}}{\sum_{i=1}^{P} \mathbb{I}_{\{Z_i > u\}}}.$$

We use the empirical quantiles u_τ at $\tau \in \{0.80, 0.85, 0.90, 0.925, 0.95, 0.975\}$ and report $\widehat{e}(u_\tau)/\widehat{s}_Z$ to compare checkpoints on the same scale. Confidence bands are percentile intervals from 2,000 bootstrap resamples of the exceedances at each threshold.

The increasing mean excess in Figure 4 agrees with the slow survival decay in Figure 3. The final checkpoint also has the largest kurtosis in Eq. (187).

G.4 Asymmetry of Large Pair Sums

The single-big-jump statistic of Asmussen & Lehtomaa (2017) measures whether one summand dominates a large pair sum, following the limiting behavior studied by Lehtomaa (2015). We randomly pair the standardized probe values $Z_i/\overline{Z}$ into $P/2$ pairs (U_i, V_i) and compute

$$J_i := \frac{|U_i - V_i|}{U_i + V_i}, \qquad T_{\mathrm{SBJ}}(d) := \frac{1}{N_d} \sum_{i:U_i+V_i > d} J_i,$$

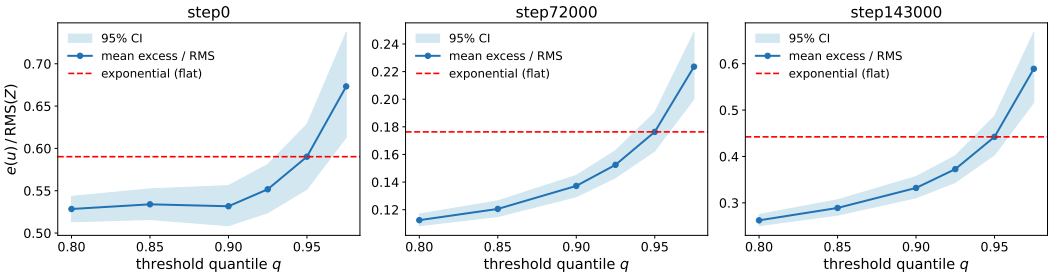


Figure 4: Normalized mean excess $\widehat{e}(u_\tau)/\widehat{s}_Z$ at threshold quantiles τ. The exponential reference is flat, whereas all three checkpoint curves increase. The strongest rise is at `step143000`.

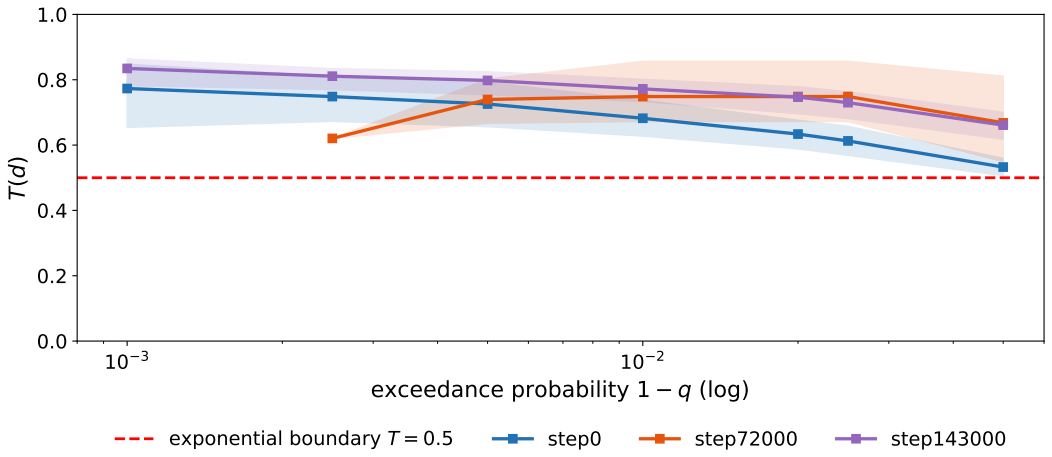


Figure 5: Exponential-calibrated single-big-jump statistic, denoted $T(d)$ in the plot. The red dashed line marks the exponential expectation $1/2$. All three checkpoint curves lie above it throughout the displayed threshold range.

where N_d is the number of retained pairs. For iid exponential variables U, V with the same rate, $U/(U+V)$ is uniform on $(0,1)$ and independent of the sum. Hence $|U-V|/(U+V)$ is also uniform on $(0,1)$, giving $\mathbb{E}[T_{\mathrm{SBJ}}(d)] = 1/2$ at every threshold. Larger values indicate more asymmetric pair sums than under this exponential reference.

For each level $\tau \in [0.95, 0.999]$, we choose the threshold d_τ from

$$e^{-d_\tau}(1+d_\tau) = 1-\tau.$$

This is the τ-quantile of the $\mathrm{Gamma}(2,1)$ distribution, the law of the sum of two independent $\mathrm{Exp}(1)$ variables. At each threshold, we average the statistic over $N_{\mathrm{perm}} = 100$ random pairings and shade the central 95% of their distribution. We check the finite-sample reference using 200 independent $\mathrm{Exp}(1)$ samples of size P, with ten pairings per sample.

In Figure 5, large pair sums are more asymmetric than under the exponential reference at every displayed threshold. Together with the mean-excess and survival curves, this supports the heavy-tail behavior seen at all three checkpoints.

H Experimental Details for Lower Bounds

We test the rate in Theorem 5 on the rotated finite-p zero-chain family of Lemma E.2. Each point in Figure 1c uses an instance scaled to its oracle budget, so the curve measures the rate across the hard family rather than along a single optimization trajectory. Codes can be found in the supplementary material.

H.1 Matching Hard Instances to Sample Budgets

We set $L = \sigma = \Delta_0 = 1$, where Δ_0 is the reference gap used to index the family, and write $q = p/(p-1)$. For each budget T, we choose

$$\begin{aligned} c_T &= \frac{4}{(184\sqrt{2})^q}, \quad \varepsilon_T = \left(\frac{c_T c_r \Delta_0 L \sigma^q}{T}\right)^{1/(2+q)}, \\ r &= \left\lfloor \frac{c_r \Delta_0 L}{\varepsilon_T^2} \right\rfloor, \quad m = 8r, \qquad \theta = \left(\frac{184\sqrt{2}\,\varepsilon_T}{\sigma}\right)^q. \end{aligned} \tag{188}$$

This gives $T \simeq 4r/\theta$ up to integer rounding, following the chain-progress budget in Eq. (132). The resulting scale $\varepsilon_T \propto T^{-(p-1)/(3p-2)}$ has the exponent predicted by Theorem 6. The chain F_m and rare-reveal oracle g_m are defined in Eqs. (125) and (129). If $k = \mathrm{prog}_{1/4}(y)$, the oracle multiplies gradient coordinates $i \geq k$ by ζ/θ, where $\zeta \sim \mathrm{Bernoulli}(\theta)$, and leaves the other coordinates unchanged. Information about the next link therefore arrives only on a reveal event. The oracle is unbiased, and Eqs. (130) and (131) bound its centered pth moment after scaling.

To keep the ambient dimension manageable, we use $d = D_r$ from Eq. (135) with $C_{\mathrm{rot}} = 1$. For each horizon, we obtain an orthonormal frame $V \in \mathbb{R}^{d\times m}$ by QR factorization of a Gaussian matrix and keep it fixed across seeds and optimizers. We then apply the soft compression in Lemma E.2,

$$\rho_R(u) = \frac{u}{\sqrt{1 + \|u\|^2/R^2}}, \qquad R = 230\sqrt{m}, \qquad \eta_0 = \frac{1}{5}, \qquad y(x) = V^\top \rho_R(x/b).$$

With the Jacobian J_R from Eq. (137), the scaled objective and oracle are

$$\begin{aligned} s = 4\varepsilon_T, \qquad b = 155 s/L, \qquad a = sb, \\ f_{\varepsilon_T}(x) = a\left[F_m(y(x)) + \frac{\eta_0}{2}\|x/b\|^2\right], \\ \mathsf{O}_{\varepsilon_T}(x,\zeta) = s\left[J_R(x/b)^\top V g_m(y(x),\zeta) + \eta_0 x/b\right]. \end{aligned} \tag{189}$$

The scaling in Lemmas E.1 and E.2 controls smoothness and the noise moment. Both optimizers start at the origin, with hyperparameters computed from the certified gap

$$\Delta_{\mathrm{eff}} = 12am \leq 238080\, c_r \Delta_0, \tag{190}$$

rather than the reference Δ_0. The small values of c_r in Table 4 are chosen so that the tested budgets lie before AdamW crosses the first gate. Rotation spreads the sparse chain gradient over the ambient coordinates and keeps the natural and Euclidean stationarity measures comparable across horizons. We report both measures below.

H.2 Optimizers and Parameter Choices

Both methods start with zero momentum, use one oracle call per update ($B = 1$), and apply no gradient clipping. For AdamW, we use the bias-corrected update in Eq. (8) with singleton blocks, $\lambda = 0$, and $\beta_1 = \beta_2$. With no weight decay, this is Adam, as labeled in the plots. Substituting the gap from Eq. (190) into Eq. (9) gives

$$\delta_A := 1 - \beta_2 = \left(\frac{\Delta_{\mathrm{eff}} L}{2^{2/p}\sigma^2 T}\right)^{p/(3p-2)}, \quad \eta_A = \sqrt{\frac{\Delta_{\mathrm{eff}}}{LdT}}, \quad \epsilon_{\mathrm{bc}} = \frac{2^{3+1/p}\sigma\delta_A^{1/2-1/p}}{\sqrt{d}}. \tag{191}$$

The condition $\sqrt{d}\,\delta_A \leq 1/8$ in Eq. (3) holds at every tested horizon.

We implement exact-polar Muon through the diagonal lift in Eq. (177), with $X = \mathrm{diag}(x)$ for $x \in \mathbb{R}^d$. The initialization and gradients are diagonal, so all iterates remain diagonal and $\mathrm{msign}\,(\mathrm{diag}(m_t)) = \mathrm{diag}(\mathrm{sign}(m_t))$. The update and tuning in Eqs. (20) and (163) then reduce to

$$\begin{aligned} \delta_M = \min\left\{1, \left(\frac{\Delta_{\mathrm{eff}} L}{2^{2/p}\sigma^2 T}\right)^{p/(3p-2)}\right\}, \quad \beta = 1 - \delta_M, \quad \eta_M = \sqrt{\frac{2\Delta_{\mathrm{eff}}\delta_M}{5dLT}}, \\ m_t = \beta m_{t-1} + (1-\beta) g_t, \quad x_{t+1} = x_t - \eta_M \,\mathrm{sign}(m_t). \end{aligned} \tag{192}$$

Table 4: Settings for the two sweeps. Ranges run from the smallest to the largest budget in Eq. (193). The exact value of c_T is given in Eq. (188).

Quantity	$p = 1.5$	$p = 2$
c_r	10^{-6}	3×10^{-8}
c_T (approx.)	2.27×10^{-7}	5.91×10^{-5}
Seeds	32	16
Master seed	20260921	20260922
ε_T	$3.726 \times 10^{-4} \rightarrow 1.483 \times 10^{-4}$	$8.65 \times 10^{-5} \rightarrow 2.74 \times 10^{-5}$
Chain parameter r	$7 \rightarrow 45$	$4 \rightarrow 40$
Chain length m	$56 \rightarrow 360$	$32 \rightarrow 320$
Expected reveals $T\theta$	$28.8 \rightarrow 181.8$	$16.0 \rightarrow 160.3$
Maximum $\sqrt{d}(1-\beta_2)$	0.0148	< 0.008

The matrix rank parameter in Eq. (163) is d, distinct from the chain parameter r in Eq. (188). The identity $\| \operatorname{diag}(z)\|_* = \|z\|_1$ makes Muon's normalized nuclear stationarity equal to AdamW's vector criterion $\|\nabla f(x)\|_1/\sqrt{d}$. We can therefore compare the methods on the same axis using vector updates, without a dense matrix SVD. All hyperparameters are fixed by these formulas before the sweeps.

H.3 Evaluation and Statistical Analysis

We use the same seven budgets for both values of p,

$$T \in \{31623, 68129, 146780, 316228, 681292, 1467799, 3162278\}. \tag{193}$$

The settings are summarized in Table 4. Every horizon satisfies $\theta < 1$ and Eq. (3), with more than 10 expected reveals. The dimension and chain length grow as ε_T decreases, following the construction.

We evaluate the analytic gradient before each update and record

$$\overline{S}_T = \frac{1}{T}\sum_{t=1}^{T} \frac{\|\nabla f_{\varepsilon_T}(x_t)\|_1}{\sqrt{d}}, \qquad \overline{S}_{2,T} = \frac{1}{T}\sum_{t=1}^{T} \|\nabla f_{\varepsilon_T}(x_t)\|_2. \tag{194}$$

The first average uses the singleton-block norm in Eq. (1). Both averages correspond to selecting an iterate uniformly from the queried points. Analytic gradients are used only to measure stationarity, not to update the optimizers.

Each marker is the mean across seeds. For each metric, we fit the log of its seed mean against $\log T$ by ordinary least squares at the five largest horizons, with this window fixed before running the sweeps. AdamW and Muon share the same oracle random stream at each horizon and seed. We therefore use 2,000 paired bootstrap resamples to obtain 95% percentile intervals for the slopes and their differences, conditional on the fixed rotation at each horizon. The dashed line illustrates the predicted exponent $-(p-1)/(3p-2)$. Its vertical offset uses the geometric-mean horizon and the median level of both optimizers for visual comparison, rather than the constant in the lower bound.

We use a NumPy reference implementation and a PyTorch implementation that batches seeds and compiles the update loop, both in FP64. Checks cover analytic gradients, oracle unbiasedness and pth moments, the diagonal sign identity, and agreement between the two implementations to 10^{-10} at each step. To understand how the methods behave on the chain, we also record reveal counts, hidden-chain progress, the ratio $\rho_{\mathrm{geom}} = \overline{S}_T/\overline{S}_{2,T}$, and the fraction of AdamW denominator coordinates for which $\sqrt{\widehat{v}_t} < \epsilon_{\mathrm{bc}}$.

H.4 Results and Interpretation

Figures 6 and 7 show the full results for $p = 1.5$ and $p = 2$, with natural metric slopes in Table 5. AdamW's slopes are -0.2050 and -0.2525, close to the predicted $-1/5$ and $-1/4$. Its Euclidean slopes are -0.2025 and -0.2532. Fitting all seven horizons at $p = 1.5$ gives -0.2010. The observed rates thus follow the predicted dependence on p over the tested finite-sample range.

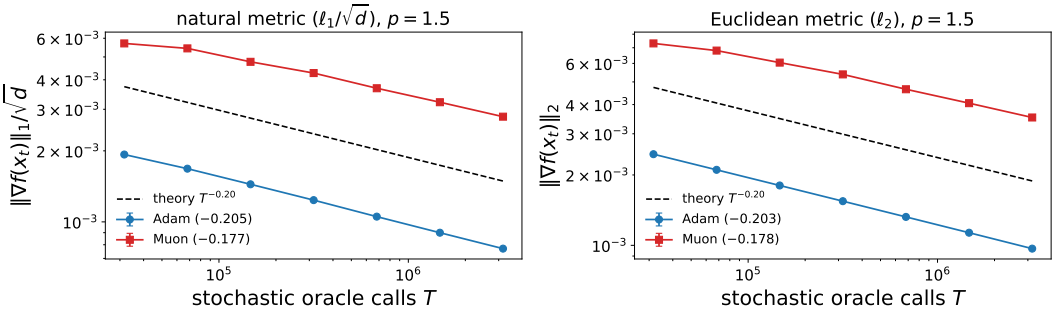


Figure 6: Full $p = 1.5$ results for Figure 1c. The left and right panels show average natural stationarity $\overline{S}_T$ and Euclidean stationarity $\overline{S}_{2,T}$ from Eq. (194). Each point uses an instance matched to its budget, averaged over 32 seeds with 95% bootstrap intervals. AdamW ($\lambda = 0$, labeled Adam) has slopes -0.205 and -0.203, close to the $T^{-1/5}$ reference, compared with -0.177 and -0.178 for exact-polar Muon. Both methods use the theorem tunings in Appendix H.2.

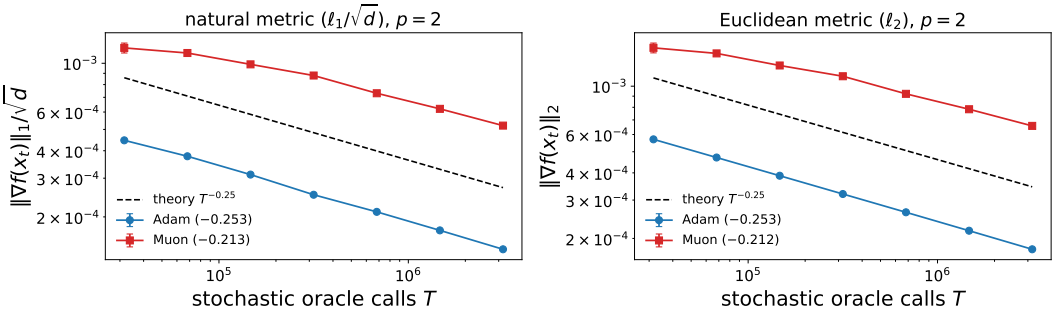


Figure 7: Results at $p = 2$ using 16 seeds, with 95% bootstrap intervals. AdamW's slope is -0.253 in both metrics, close to the predicted $-1/4$. Muon's slopes are -0.213 and -0.212. The construction and fitting window are the same as in Figure 6, with parameters given in Table 4.

Muon has a flatter natural metric slope over this range. Subtracting Muon's slope from AdamW's gives -0.0285 with paired interval $[-0.0314, -0.0254]$ at $p = 1.5$, and -0.0398 with interval $[-0.0429, -0.0369]$ at $p = 2$. At $p = 1.5$, the ratio of Muon's stationarity to AdamW's grows from 2.96 to 3.62. This separation concerns finite budgets under the stated tunings and does not determine Muon's asymptotic exponent.

AdamW remains at the first gate throughout both sweeps. Its denominator floor has a recorded active fraction of 100%, and the mean $\sqrt{\widehat{v}_t}/\epsilon_{\mathrm{bc}}$ is about 10^{-3} at $p = 1.5$. The update is therefore approximately $-(\eta_A/\epsilon_{\mathrm{bc}})\widehat{m}_t$, with zero hidden-chain progress at every tested horizon. The gradient at the origin in Eq. (127) is $\nabla F_m(0) = -\sqrt{e}\, e_1$, giving the initial natural stationarity

$$\frac{\|\nabla f_{\varepsilon_T}(0)\|_1}{\sqrt{d}} = 4\sqrt{e}\,\frac{\|Ve_1\|_1}{\sqrt{d}}\,\varepsilon_T \simeq 4\sqrt{\frac{2e}{\pi}}\,\varepsilon_T \simeq 5.26\,\varepsilon_T.$$

The observed averages stay close to this level. At $p = 1.5$, $\overline{S}_T/\varepsilon_T$ equals 5.179, 5.259, 5.265, 5.260, 5.220, 5.202, and 5.193 in increasing horizon order. At $p = 2$, the ratio ranges from 5.156 to 5.286. In both cases, stationarity tracks the family scale $\varepsilon_T \propto T^{-(p-1)/(3p-2)}$. Muon, in contrast, progresses through roughly half of the hidden chain, so the rotation does not prevent it from advancing.

Rotation is important for interpreting these rates. On the unrotated sparse chain, the natural metric can change relative to the Euclidean metric as the dimension grows. A canonical-chain pilot gave Euclidean slopes near -0.20 but natural metric slopes near -0.40, with ρ_{geom} contributing an extra slope of -0.20. After rotation, this ratio is nearly constant, with slope magnitude at most 0.005 in the sweeps and replication. An independent NumPy replication at $p = 1.5$ used 32 seeds and the first six horizons in Eq. (193). Fitting its five largest horizons gives an AdamW slope of -0.1969 (95% interval $[-0.1970, -0.1969]$) and a Muon slope of -0.1659. The shifted-window and full-grid fits

Table 5: Natural metric slopes over the five largest horizons. Brackets give 95% bootstrap intervals over seeds.

p	Theory	AdamW	Muon
1.5	-0.20	-0.2050 $[-0.2051, -0.2049]$	-0.1766 $[-0.1797, -0.1737]$
2	-0.25	-0.2525 $[-0.2527, -0.2524]$	-0.2127 $[-0.2156, -0.2095]$

again support the predicted exponent. Independent rotation draws also introduce small differences in the curve levels.

Fixed-instance control. We also test one fixed canonical instance with $p = 1.5$, $\varepsilon = 1.5 \times 10^{-3}$, $r = 44$, $m = 352$, and $\Delta_{\mathrm{eff}} \approx 23.8$. We replace Bernoulli reveals by a fixed amplitude law $A = 2^K$, with $\mathbb{P}(K = k) \propto 2^{-kp}/k^2$ for $k = 1, \dots, 60$, normalized over this range. The same gradient block as in Eq. (129) is multiplied by $A/\mathbb{E}A$. Only the horizon and its theorem tuning vary. We use 16 seeds with master seed 20260922 and $T \in \{2048, 8192, 32768, 131072, 524288, 2097152\}$. The measured frequency of $A \geq 2\mathbb{E}A$ ranges from 0.0153 to 0.0160, close to the law's probability 0.01581. AdamW stays near $\overline{S}_T = 5.2 \times 10^{-4}$ up to the matched horizon $T \approx 3.25 \times 10^4$, after which its trajectory average decreases roughly as T^{-1} with fitted slope -1.08. Muon stays near 5×10^{-3} with slope $+0.06$, about 300 times higher at the largest budget. A fixed instance need not sustain the minimax scaling beyond its matched budget, which is why the main rate experiment varies the instance with T.