WHEN MUON BECOMES GRADIENT DESCENT: FINITE-TIME FLOWS AND NOISE SMOOTHING NEAR THE OPTIMUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon is a recently proposed optimizer whose matrix update applies the polar factor (matrix sign) of a gradient or momentum matrix. We study an idealized no-momentum, no-weight-decay Muon update on matrix parameters $W \in \mathbb{R}^{m\times n}$ in deterministic and stochastic regimes. In the deterministic continuous-time limit, Muon is steepest descent with respect to the spectral norm and obeys the dissipation identity $\frac{d}{dt}F(W_t) = -\|\nabla F(W_t)\|_*$; under a nuclear Łojasiewicz-type error bound this yields *finite-time* convergence. With minibatch noise, moment matching gives an Itô SDE whose drift is the noise-smoothed polar map $\Psi(W) = \mathbb{E}[\mathrm{msgn}(\nabla F(W) + \xi)]$ rather than $\mathrm{msgn}(\nabla F(W))$. For Gaussian noise $\xi = sZ$, the drift linearizes near stationarity as $\Psi_s(Y) = (c_{m,n}/s)\,Y + o(\|Y\|_F/s)$ with $c_{m,n} = \mathbb{E}\|Z\|_*/(mn)$, and $c_{N,N} \sim 8/(3\pi\sqrt{N})$ for square $N \times N$ matrices. The resulting local dynamics have an Ornstein–Uhlenbeck noise neighborhood. On a matrix quadratic with isotropic Gaussian noise, SGD has no higher stationary noise floor when the two methods have matched local contraction rates. This motivates switching from Muon to SGD as the gradient enters the noise scale.

1 INTRODUCTION

Neural-network training often uses updates that differ substantially from the Euclidean gradient. Adam and AdamW use coordinatewise adaptive scaling (Kingma & Ba, 2015; Loshchilov & Hutter, 2019); Shampoo uses matrix preconditioning (Gupta et al., 2018); Lion uses signed momentum (Chen et al., 2023). Muon orthogonalizes gradient momentum through a polar-factor transformation (Jordan et al., 2024; Bernstein & Newhouse, 2024) and has been used in large language model training (DeepSeek-AI, 2026; Liu et al., 2025). Their convergence behavior, and when each should be preferred, is less well understood than that of gradient descent.

This paper studies Muon in a simplified form that is amenable to analysis. We strip away momentum, weight decay, and Newton–Schulz polar-factor approximations and study

$$W_{k+1} = W_k - \eta\ \mathrm{msgn}(G_k), \qquad G_k = g(W_k) \quad \text{or} \quad G_k = g(W_k) + \xi_k, \tag{1}$$

where $W_k \in \mathbb{R}^{m\times n}$, $g(W) := \nabla F(W)$, ξ_k is the centered minibatch noise (Section 5), and $\mathrm{msgn}(G)$ is the polar factor of G. This idealization isolates the effect of replacing the gradient by a spectrally normalized matrix direction.

Deterministic flow and finite-time convergence. In the continuous-time limit $\eta \to 0$ with $t = k\eta$, the deterministic update yields the idealized Muon flow

$$\dot{W}_t = -\,\mathrm{msgn}(\nabla F(W_t)). \tag{2}$$

Because $\mathrm{msgn}(G) \in \partial\|G\|_*$, this flow has the exact *nuclear* dissipation identity

$$\frac{d}{dt}F(W_t) = -\langle \nabla F(W_t), \mathrm{msgn}(\nabla F(W_t))\rangle = -\|\nabla F(W_t)\|_*, \tag{3}$$

qualitatively different from gradient flow's $-\|\nabla F(W_t)\|_F^2$ dissipation. The nuclear-norm rate is first order in the gradient magnitude rather than quadratic, so it dominates when the gradient is small. Under a nuclear Łojasiewicz-type error bound this yields *finite-time* convergence, in contrast to the merely exponential rate of gradient flow under the usual Polyak–Łojasiewicz condition.

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.

Stochastic minibatch dynamics. The stochastic dynamics are not obtained by adding Brownian noise to (2): because ξ_k enters *inside* the nonlinear polar map, the drift itself is changed. The averaged drift is the *noise-smoothed polar map* $\Psi(W) = \mathbb{E}_\xi[\operatorname{msgn}(g(W) + \xi)]$, not $\operatorname{msgn}(g(W))$. Matching the first two moments of one Muon step to an Itô diffusion gives the stochastic modified equation

$$d\operatorname{vec}(W_t) = -\operatorname{vec}(\Psi(W_t))\,dt + \sqrt{\eta}\,C_M(W_t)^{1/2}dB_t, \tag{4}$$

where $C_M(W)$ is the covariance of $\operatorname{vec}\operatorname{msgn}(g(W) + \xi)$. Sending $\eta \to 0$ at fixed minibatch law eliminates the diffusion and recovers the averaged ODE $\dot{W}_t = -\Psi(W_t)$; (4) captures the small but finite learning-rate fluctuations.

Near-optimum smoothing and switching. Minibatch noise smooths the discontinuous matrix sign near stationarity. For Gaussian noise $\xi = sZ$, set $\phi_s(Y) = \mathbb{E}\|Y + sZ\|_*$ and $\Psi_s(Y) = \mathbb{E}\operatorname{msgn}(Y + sZ) = \nabla\phi_s(Y)$. Near $Y = 0$,

$$\Psi_s(Y) = \frac{c_{m,n}}{s}Y + o(\|Y\|_F/s), \qquad c_{m,n} := \frac{\mathbb{E}\|Z\|_*}{mn}, \tag{5}$$

with $c_{N,N} \sim 8/(3\pi\sqrt{N})$ for square $N \times N$ matrices. Thus stochastic Muon has a locally scaled gradient drift, and the deterministic finite-time advantage disappears in this Gaussian-smoothed layer. This motivates using Muon away from stationarity and switching to SGD once the gradient enters the noise scale.

Contributions.

1. We formulate the no-momentum, no-decay deterministic Muon flow (2) and prove finite-time convergence under a nuclear Łojasiewicz-type error bound, with explicit hitting-time bounds for $\frac{1}{2}\|W - W^\star\|_F^2$ and $\|AWB - C\|_F^2$.
2. We derive the stochastic modified SDE (4) for minibatch Muon and compare it with the corresponding SDE for SGD. A Lyapunov inequality gives an $O(1/T) + O(\eta)$ bound on the time average of $\mathbb{E}\langle g, \Psi\rangle$.
3. We prove a linearization theorem for the Gaussian-smoothed polar map and give explicit asymptotics for $c_{m,n}$ in the scalar, vector, square, and rectangular regimes.
4. We analyze the near-optimum SDE on $\|AWB - C\|_F^2$, identify its scaled-gradient drift, and obtain a curvature-independent comparison of stochastic Muon and SGD stationary variances and objective floors.
5. We evaluate four predictions on synthetic matrix problems and discuss when switching from Muon to SGD may be useful.

2 Related work

Muon, polar maps, and non-Euclidean views. Muon orthogonalizes a momentum matrix via its polar factor (Jordan et al., 2024). Subsequent work interprets Muon and related rules as steepest descent in non-Euclidean geometries induced by matrix norms and dual pairs (Bernstein & Newhouse, 2024; 2025; Chen et al., 2026). Empirical scaling studies show Muon can match or improve on AdamW-style training at scale and that its behavior is tightly coupled to the choice of normalization layer (Liu et al., 2025; Abouzeid, 2026); other work surveys or proposes Muon variants, including spectral steps, second-moment preconditioning, and trust-region or adaptive scalings (Crawshaw et al., 2026; Liu et al., 2026; Cheng et al., 2026). Polynomial approximations of the polar factor, such as Newton–Schulz iterations, have been analyzed and optimized as numerical primitives (Amsel et al., 2026). Du & Su (2026) interpret Muon as an implicit Newton-type method that omits preconditioning by the input second moment, which leads to the Newton-Muon variant. Other work analyzes the convergence of Muon under idealized or approximate matrix-sign oracles (Shen et al., 2025; Kim & Oh, 2026). Preconditioning interpretations connect spectral orthogonalization to improved conditioning in two-layer models (Ma et al., 2026). Our idealization removes momentum, weight decay, and Newton–Schulz error to isolate the continuous-time flow and the way minibatch noise enters the polar map.

Sign methods, Lion, and implicit constraints. Sign-based and normalized-gradient methods trade raw gradient magnitude for scale invariance and saddle escape (Bernstein et al., 2018; Balles & Hennig, 2018; Levy, 2016; Murray et al., 2019). The Lion-K framework (Chen et al., 2024) explains sign-momentum with decoupled weight decay through convex conjugacy and Lyapunov structure; when the regularizer is the nuclear norm, its conjugate is an indicator on the spectral-norm ball, linking weight decay to spectral constraints. We set weight decay to zero and study unconstrained flows.

SDE and stochastic modified equations. Diffusion limits for stochastic approximation are classical (Kushner & Yin, 2003); stochastic modified equations for SGD clarify small-step scaling and stationary behavior (Li et al., 2017; Mandt et al., 2017). For Muon, the noise enters inside $\mathrm{msgn}(\cdot)$, so the drift becomes $\mathbb{E}[\mathrm{msgn}(g(W)+\xi)]$.

Łojasiewicz-type geometry and finite-time flows. Łojasiewicz and Kurdyka–Łojasiewicz inequalities govern asymptotics of gradient-like systems (Łojasiewicz, 1963; Bolte et al., 2007; Karimi et al., 2016). Deterministic finite-time convergence is less common for smooth objectives; our assumption uses nuclear-norm gradient magnitude to obtain a sublinear differential inequality compatible with the Muon dissipation rate.

3 Preliminaries

For $W \in \mathbb{R}^{m\times n}$, let $\|W\|_F$ denote the Frobenius norm, $\|W\|_{\mathrm{op}}$ the spectral norm, and $\|W\|_*$ the nuclear norm. Let $r = \min(m,n)$. We write $\mathrm{vec}(W) \in \mathbb{R}^{mn}$ for the column-major vectorization of W into a single vector.

Definition 3.1 (Matrix sign / polar factor)**.** Let $G = U\Sigma V^\top$ be a compact singular value decomposition with rank ρ. The matrix sign is $\mathrm{msgn}(G) = UV^\top$. Equivalently, when G has full row rank, $\mathrm{msgn}(G) = (GG^\top)^{-1/2}G$, and analogously for other rectangular cases using the Moore–Penrose inverse.

Lemma 3.2 (Nuclear-norm subgradient)**.** *For every $G \in \mathbb{R}^{m\times n}$,*

$$\mathrm{msgn}(G) \in \partial\|G\|_*, \qquad \textit{in particular} \qquad \langle G, \mathrm{msgn}(G)\rangle = \|G\|_*. \tag{6}$$

Proof: Appendix A.1; see Rockafellar (1970) for background on convex subdifferentials.

We study the expected-risk objective $\min_{W\in\mathbb{R}^{m\times n}} F(W)$ with $F(W) = \mathbb{E}_{z\sim\mathcal{D}}[F(W,z)]$. The no-momentum Muon update is

$$W_{k+1} = W_k - \eta\ \mathrm{msgn}(G_k), \tag{7}$$

where $G_k = \nabla F(W_k)$ in the deterministic case and $G_k = g(W_k) + \xi_k$ in the stochastic minibatch case (Section 5).

4 Deterministic idealized Muon flow

Taking the formal limit $\eta \to 0$ with $t = k\eta$ in the deterministic version of (7) gives the idealized Muon flow

$$\dot{W}_t = -\,\mathrm{msgn}(\nabla F(W_t)), \tag{8}$$

with the convention $\dot{W}_t = 0$ at stationary points.

Proposition 4.1 (Nuclear dissipation)**.** *For W_t solving (8) on an interval where $\nabla F(W_t)$ is continuous,*

$$\frac{d}{dt}F(W_t) = -\|\nabla F(W_t)\|_*. \tag{9}$$

Proof: Appendix A.2. Gradient flow $\dot{W}_t = -\nabla F(W_t)$, by contrast, dissipates F at the quadratic rate $-\|\nabla F(W_t)\|_F^2$. Muon's first-order dissipation is the geometric source of finite-time convergence.

4.1 FINITE-TIME CONVERGENCE

Assumption 4.2 (Nuclear Łojasiewicz error bound). *There exist $\kappa > 0$ and $\theta \in [0, 1)$ such that, on the sublevel set $\mathcal{S} = \{W : F(W) \le F(W_0)\}$,*

$$\|\nabla F(W)\|_* \ge \kappa \, (F(W) - F^\star)^\theta. \tag{10}$$

Theorem 4.3 (Finite-time convergence). *If F is C^1, bounded below, and Assumption 4.2 holds on the sublevel set reached by (8), then $F(W_t) = F^\star$ for all $t \ge T_\star$, where*

$$T_\star \le \frac{(F(W_0) - F^\star)^{1-\theta}}{\kappa(1-\theta)}. \tag{11}$$

Proof: Appendix A.3.

Remark 4.4 (Contrast with PL-driven gradient flow). Under the Polyak–Łojasiewicz inequality $\|\nabla F\|_F^2 \ge 2\mu(F - F^\star)$ (in particular, when F is μ-strongly convex), gradient flow gives the exponential bound $F(W_t) - F^\star \le e^{-2\mu t}(F(W_0) - F^\star)$, but never reaches $F^\star$ in finite time. Muon's finite-time behavior comes from replacing the squared-Frobenius dissipation by the nuclear-norm dissipation (9).

4.2 TWO ILLUSTRATIVE EXAMPLES

Identity quadratic. For $F(W) = \frac{1}{2}\|W - W^\star\|_F^2$, set $E_t = W_t - W^\star$. Along the Muon flow $\dot{E}_t = -\operatorname{msgn}(E_t)$, the singular vectors of E_t are fixed and each singular value decays at unit rate: $\sigma_i(t) = (\sigma_i(0) - t)_+$. Hence $E_t = 0$ for all $t \ge \|E_0\|_{\mathrm{op}}$. Assumption 4.2 holds with $\theta = 1/2$ and $\kappa = \sqrt{2}$.

Matrix quadratic. Consider $F(W) = \|AWB - C\|_F^2$, with $P := A^\top A$ and $Q := BB^\top$. This objective is analytically tractable and is also the local linearization of a single dense layer's contribution to a regression loss: freezing all other parameters in a deep network, the linearized residual seen by the layer with weight matrix W has the form $AWB - C$, where A and B collect the upstream/downstream activations (or Jacobians of subsequent layers) and C collects the target. Hence the dynamics on this objective serve as a local model for what Muon does to a single neural-network layer near a working point, with P and Q playing the role of the input- and output-side Gram matrices.

The gradient is $\nabla F(W) = 2A^\top(AWB - C)B^\top$. When $AW^\star B = C$ at an optimum, writing $E = W - W^\star$ gives

$$F(W) - F^\star = \operatorname{tr}(E^\top PEQ), \qquad \nabla F(W) = 2PEQ. \tag{12}$$

Corollary 4.5 (Finite-time hitting for full-rank matrix quadratics). *If $P \succeq p_{\min} I$ and $Q \succeq q_{\min} I$ with $p_{\min}, q_{\min} > 0$, then Assumption 4.2 holds with $\theta = 1/2$, $\kappa = 2\sqrt{p_{\min} q_{\min}}$, and deterministic Muon reaches the optimum by*

$$T_\star \le \frac{\sqrt{F(W_0) - F^\star}}{\sqrt{p_{\min} q_{\min}}}. \tag{13}$$

Proof: Appendix A.4.

Remark 4.6 (Anisotropy and curvature scale). Gradient flow $\dot{E} = -2PEQ$ contracts mode (i, j) at the rate $\exp(-2p_i q_j t)$, so the slowest mode is bottlenecked by $p_{\min} q_{\min}$. By contrast, in the diagonal case $P = \operatorname{diag}(p_i)$, $Q = \operatorname{diag}(q_i)$, $E = \operatorname{diag}(e_i)$, deterministic Muon gives $\dot{e}_i = -\operatorname{sign}(e_i)$ on each active mode, with hitting time $|e_i(0)|$ *independent* of the curvature scale. This is the deterministic mechanism by which Muon removes singular-value anisotropy in matrix problems.

5 STOCHASTIC MUON: SDE MODEL

For $z \sim \mathcal{D}$, write $g(W, z) = \nabla_W F(W, z)$ and $g(W) = \mathbb{E}_z g(W, z)$. Given a minibatch of size B, let $\xi_k := \frac{1}{B}\sum_b g(W_k, z_b) - g(W_k)$ be the centered minibatch noise, so $\mathbb{E}[\xi_k \mid W_k] = 0$. Stochastic Muon is

$$W_{k+1} = W_k - \eta \, \operatorname{msgn}(g(W_k) + \xi_k). \tag{14}$$

	SGD	Stochastic Muon
Discrete update	$W - \eta(g + \xi)$	$W - \eta\,\mathrm{msgn}(g + \xi)$
Drift	$-g(W)$	$-\mathbb{E}[\mathrm{msgn}(g(W) + \xi)]$
Diffusion (cov)	$\eta\Sigma_g(W)/B$	$\eta C_M(W)$
Noise in drift	unbiased	smoothed / biased
Near optimum	linear OU	smoothed polar OU

Table 1: SGD versus stochastic Muon at the SDE level.

Conditioning on $W_k = W$ and matching the first two moments to those of an Itô diffusion in $\mathbb{R}^{mn}$ with $t = k\eta$ yields the stochastic modified equation

$$d\,\mathrm{vec}(W_t) = -\,\mathrm{vec}(\Psi(W_t))\,dt + \sqrt{\eta}\,C_M(W_t)^{1/2}dB_t, \tag{15}$$

where $\Psi(W) := \mathbb{E}[\mathrm{msgn}(g(W) + \xi)]$ is the *smoothed polar drift* and $C_M(W) := \mathrm{Cov}(\mathrm{vec}\,\mathrm{msgn}(g(W) + \xi))$ is the matrix-sign covariance. Here (B_t) is a standard Brownian motion in $\mathbb{R}^{mn}$. In general $\Psi(W) \neq \mathrm{msgn}(g(W))$: minibatch noise enters *inside* the polar map and biases the drift. Sending $\eta \to 0$ at fixed minibatch law eliminates the diffusion term and gives the averaged ODE $\dot{W}_t = -\Psi(W_t)$. An analogous moment match for plain SGD yields

$$d\,\mathrm{vec}(W_t) = -\,\mathrm{vec}(g(W_t))\,dt + \sqrt{\eta/B}\,\Sigma_g(W_t)^{1/2}dB_t, \quad \Sigma_g(W) := \mathrm{Cov}\left(\mathrm{vec}\,g(W,z)\right), \tag{16}$$

with the unbiased gradient as drift. Table 1 summarizes the comparison; Appendix A.9 gives the full moment-matching derivation.

Applying Itô's formula to (15) gives

$$\begin{aligned} dF(W_t) = &-\langle g(W_t), \Psi(W_t)\rangle\,dt + \sqrt{\eta}\,\langle g(W_t), C_M(W_t)^{1/2}dB_t\rangle \\ &+ \frac{\eta}{2}\,\mathrm{tr}\big(\nabla^2 F(W_t)\,C_M(W_t)\big)\,dt. \end{aligned} \tag{17}$$

Since $\|\,\mathrm{msgn}(\cdot)\|_F^2 \le r$ pointwise (the polar factor has at most r unit singular values), $\mathrm{tr}\,C_M(W) \le \mathbb{E}\|\,\mathrm{msgn}(g(W) + \xi)\|_F^2 \le r$. If F is L-smooth, the Hessian–covariance trace is therefore bounded by Lr, and taking expectations yields the Lyapunov inequality

$$\frac{d}{dt}\mathbb{E}F(W_t) \le -\mathbb{E}\langle g(W_t), \Psi(W_t)\rangle + \frac{\eta L r}{2}. \tag{18}$$

The first term is the stochastic analogue of the deterministic nuclear dissipation (9); the $O(\eta)$ correction reflects the finite step size. Integrating (18) over $[0, T]$ and dividing by T gives a non-asymptotic ergodic bound on the average dissipation rate:

$$\frac{1}{T}\int_0^T \mathbb{E}\langle g(W_t), \Psi(W_t)\rangle\,dt \le \frac{F(W_0) - F^\star}{T} + \frac{\eta L r}{2}. \tag{19}$$

Far from stationarity, when Ψ is close to $\mathrm{msgn}(g)$, the left-hand side approaches the time average of $\mathbb{E}\|g(W_t)\|_*$. In that regime, (19) parallels the deterministic dissipation identity, with an $O(1/T) + O(\eta)$ bound. Taking $T = \Theta(1/\eta)$ balances the two terms.

6 Gaussian-smoothed matrix sign

The drift in (15) is governed by

$$\Psi_s(Y) = \mathbb{E}\,\mathrm{msgn}(Y + sZ), \qquad \phi_s(Y) = \mathbb{E}\|Y + sZ\|_*, \qquad \nabla_Y \phi_s(Y) = \Psi_s(Y), \tag{20}$$

where $Z \in \mathbb{R}^{m\times n}$ has i.i.d. $N(0,1)$ entries and $s > 0$. Define the per-entry expected nuclear norm

$$c_{m,n} := \frac{\mathbb{E}\|Z\|_*}{mn}. \tag{21}$$

Theorem 6.1 (Linearization of the Gaussian-smoothed polar map)**.** *For i.i.d. Gaussian $Z \in \mathbb{R}^{m\times n}$,*

$$D\Psi_s(0)[H] = \frac{c_{m,n}}{s}H, \qquad D^2\phi_s(0)[H,H] = \frac{c_{m,n}}{s}\|H\|_F^2, \tag{22}$$

and consequently

$$\Psi_s(Y) = \frac{c_{m,n}}{s}\,Y + o(\|Y\|_F/s) \quad \textit{as } \|Y\|_F/s \to 0. \tag{23}$$

Proof: Appendix A.5.

Square matrices. For $m = n = N$, the singular values of $Z/\sqrt{N}$ converge to the quarter-circle law (Marchenko & Pastur, 1967; Edelman, 1988), giving $\mathbb{E}\|Z\|_* \sim \frac{8}{3\pi} N^{3/2}$ and hence

$$c_{N,N} \sim \frac{8}{3\pi\sqrt{N}}. \tag{24}$$

Combined with (23), this says the local Muon drift slope $c_{N,N}/s$ scales as $N^{-1/2}/s$: at fixed noise scale s, larger matrices linearize to a *smaller* effective gradient step. Closed forms for the scalar ($m = n = 1$), vector ($m = 1$), and rectangular ($m/n \to \gamma$) regimes are given in Appendix A.10.

7 Matrix quadratic: local SDE and modewise comparison with SGD

We specialize to $F(W) = \|AWB - C\|_F^2$ with $\nabla F(W^\star) = 0$, $E = W - W^\star$, $P = A^\top A$, $Q = BB^\top$, so that locally $g(W^\star + E) = 2PEQ$. Assume isotropic Gaussian noise $\xi = sZ$ with $Z_{ij} \sim N(0, 1)$.

Theorem 7.1 (Near-optimum stochastic Muon SDE)**.** *Near $W^\star$, the stochastic Muon SDE (15) reduces to*

$$d\,\mathrm{vec}(E_t) = -\frac{2c_{m,n}}{s}(Q \otimes P)\,\mathrm{vec}(E_t)\,dt + \sqrt{\eta\, r/(mn)}\,dB_t + \textit{(higher order)}, \tag{25}$$

where $r = \min(m, n)$. For SGD with the same noise scale,

$$d\,\mathrm{vec}(E_t) = -2(Q \otimes P)\,\mathrm{vec}(E_t)\,dt + \sqrt{\eta}\, s\, dB_t. \tag{26}$$

Proof and modewise derivation: Appendix A.7, A.11. Diagonalizing P and Q decouples (25)–(26) into scalar Ornstein–Uhlenbeck modes whose contraction-rate and stationary-variance ratios (35) are $c_{m,n}/s$ and $r/(s\mathbb{E}\|Z\|_*)$ respectively, both independent of the mode curvature $p_i q_j$. The variance ratio also governs the local objective floor $F(W^\star + E) - F^\star = \sum_{ij} p_i q_j \widetilde{e}_{ij}^2$. For square $N \times N$ matrices, $c_{N,N}/s \sim 8/(3\pi s\sqrt{N})$ by (24), so at the same learning rate η, stochastic Muon contracts faster than SGD when $s \lesssim 8/(3\pi\sqrt{N})$, and the variance ratio $\sim 3\pi/(8s\sqrt{N})$ shows that SGD has the lower stationary noise floor when $s \lesssim 3\pi/(8\sqrt{N})$. A comparison that does not depend on how the two learning rates are normalized is obtained by matching the local contraction rates: the variance ratio is then $r\,mn/(\mathbb{E}\|Z\|_*)^2 \geq 1$, which tends to $9\pi^2/64 \approx 1.39$ for square matrices (Appendix A.11). Near the optimum, stochastic Muon therefore never has a lower noise floor than SGD running at the same local speed.

Remark 7.2 (Near vs. far from the optimum). Far from $W^\star$, where $\sigma_{\min}(g(W)) \geq 4s(\sqrt{m} + \sqrt{n})$ (note $\mathbb{E}\|Z\|_{\mathrm{op}} \leq \sqrt{m} + \sqrt{n}$), a Taylor expansion of msgn at $Y = g(W)$ gives $\mathrm{msgn}(Y + sZ) = \mathrm{msgn}(Y) + s\,D\,\mathrm{msgn}_Y[Z] + O(s^2)$, and the centered first-order term averages to zero, so by Lemma A.2 (Appendix A.8)

$$\mathbb{E}\,\mathrm{msgn}(g(W) + sZ) = \mathrm{msgn}(g(W)) + O_{m,n}\left(\frac{s^2}{\sigma_{\min}(g(W))^2}\right)$$

in Frobenius norm. The expected stochastic Muon direction is then close to the deterministic Muon direction with a bias only second order in s. Near the optimum, $\sigma_{\min}(g(W))$ enters the noise scale, the Taylor expansion fails, and Theorem 6.1 takes over: when $\|g(W)\|_F \lesssim s$, $\mathbb{E}\,\mathrm{msgn}(g(W)+sZ) \approx (c_{m,n}/s)\,g(W)$, a scaled gradient direction. Muon's deterministic finite-time advantage is therefore a transient, away-from-optimum phenomenon that disappears in the Gaussian-smoothed layer, which motivates switching from Muon to SGD once the gradient enters the noise scale.

8 Experiments

We test four predictions on synthetic matrix problems using the no-momentum, no-decay update (7) and an exact msgn via SVD: (i) the deterministic finite-time hitting on the identity quadratic (Section 8.1); (ii) the dependence of the deterministic rate on $p_{\min}q_{\min}$ for the matrix quadratic (Section 8.2); (iii) the deterministic, stochastic, and hybrid behavior on the matrix quadratic $\|AWB - C\|_F^2$, including the modewise Muon/SGD variance ratio (Section 8.3); and (iv) the linearization slope $c_{m,n}/s$ of the Gaussian-smoothed polar map (Section 8.4).

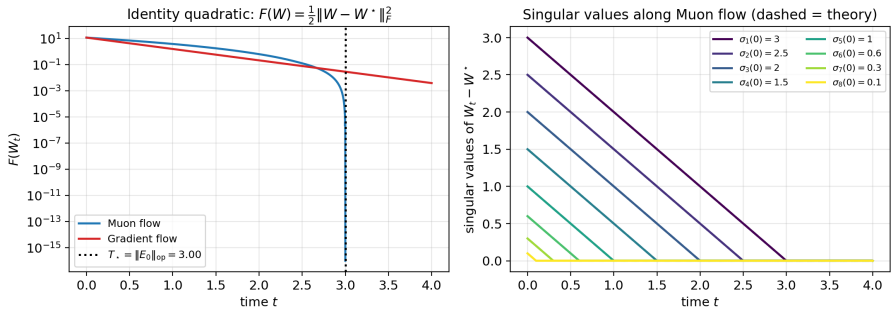

Figure 1: Identity quadratic. **Left**: Muon reaches numerical precision near $t = T_\star = 3$; gradient flow decays exponentially. **Right**: each singular value of $W_t - W^\star$ along Muon tracks $(\sigma_i(0) - t)_+$ (dashed).

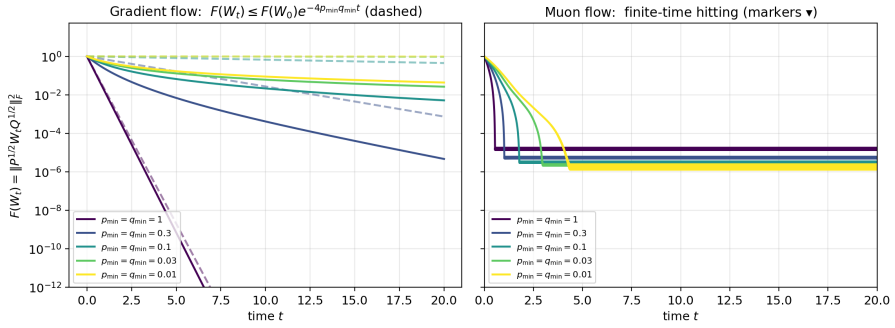

Figure 2: Continuous-time flows on $F(W) = \|P^{1/2}WQ^{1/2}\|_F^2$, $N = 12$, $p_{\max} = q_{\max} = 1$, varying $p_{\min} = q_{\min}$, with $F(W_0) = 1$. **Left**: gradient flow vs. the bound $F(W_0)e^{-4p_{\min}q_{\min}t}$ (dashed); the GD half-life grows by 10^4 as $p_{\min}q_{\min}$ shrinks by 10^4. **Right**: Muon flow on the same axis; empirical hitting times grow only $\approx 6\times$ across the same range, in line with $1/\sqrt{p_{\min}q_{\min}}$.

8.1 Finite-time convergence on the identity quadratic

For $F(W) = \frac{1}{2}\|W - W^\star\|_F^2$ on $\mathbb{R}^{12\times 8}$ with $\|E_0\|_{\mathrm{op}} = 3$, integrating both flows by forward Euler ($\eta = 10^{-3}$), theory predicts $\sigma_i(t) = (\sigma_i(0) - t)_+$ along Muon and $W_t = W^\star$ for $t \ge T_\star = 3$.

Figure 1 matches both predictions up to numerical precision: the singular values decay at unit rate until they reach the discretization floor, and the objective is $\approx 10^{-15}$ near $t = T_\star$.

8.2 Continuous-time convergence on the matrix quadratic

We use the matrix quadratic $\min_W F(W) = \|AWB - C\|_F^2$ in centered form: at the optimum $W^\star$ with $AW^\star B = C$, the residual (12) reduces to $F(W) - F^\star = \|P^{1/2}EQ^{1/2}\|_F^2$ with $E = W - W^\star$, $P = A^\top A$, $Q = BB^\top$. Taking $C = 0$ and $W^\star = 0$ for simplicity, the objective becomes $F(W) = \|P^{1/2}WQ^{1/2}\|_F^2$, and we can directly sweep the curvature spectrum to isolate the dependence of the deterministic rate on $p_{\min}q_{\min}$.

Let $P = \mathrm{diag}(p_1, \dots, p_N)$ and $Q = \mathrm{diag}(q_1, \dots, q_N)$ with $N = 12$ and p_i, q_i geometrically spaced in $[p_{\min}, 1]$. Rescaling W_0 so $F(W_0) = 1$ and integrating $\dot W = -2PWQ$ (GD) and $\dot W = -\,\mathrm{msgn}(2PWQ)$ (Muon) by forward Euler, theory predicts $F(W_t) \le e^{-4p_{\min}q_{\min}t}$ for GD and $T_\star \le 1/\sqrt{p_{\min}q_{\min}}$ for Muon: GD degrades *linearly* in $p_{\min}q_{\min}$ inside the exponent, whereas Muon's hitting time degrades only as $1/\sqrt{p_{\min}q_{\min}}$.

Figure 2 shows the predicted asymmetry: at $p_{\min} = 10^{-2}$ the GD curve has barely moved at $t = 20$, while Muon has hit the discretization floor by $t \approx 4$. The empirical Muon hitting time grows only $\approx 6\times$ as $p_{\min}q_{\min}$ drops $10^4\times$, well below the $10^2\times$ growth allowed by the upper bound.

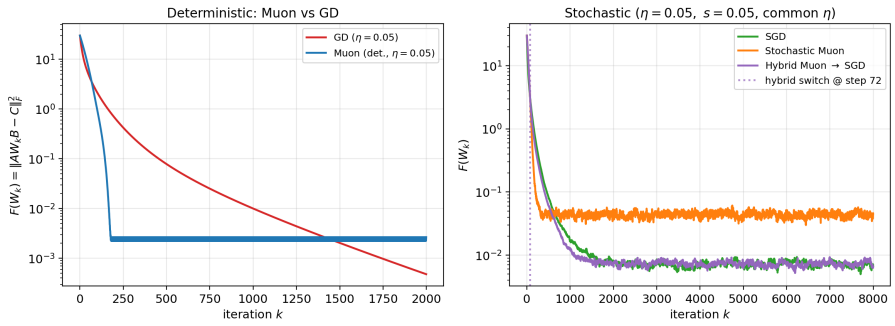

Figure 3: Matrix quadratic on $\mathbb{R}^{15\times15}$. **Left** (deterministic, $\eta = 0.05$): Muon reaches its discretization floor in $\sim$150 steps; GD needs $\sim$1500. **Right** (stochastic, $\eta = 0.05$, $s = 0.05$): SGD's stationary floor is $\approx 7\times$ lower than that of stochastic Muon, close to the predicted $\frac{3\pi}{8s\sqrt{N}} \approx 6.1$ from (35); the hybrid follows Muon's fast transient and converges to SGD's lower floor after switching.

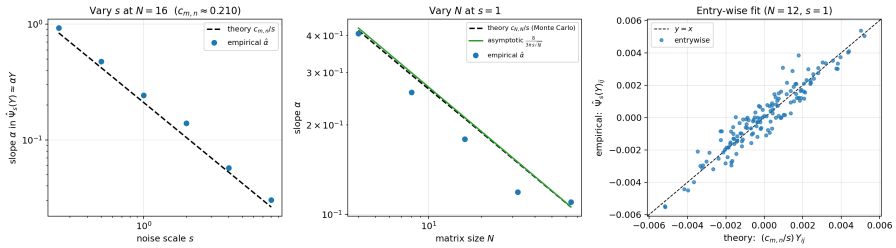

Figure 4: Noise-layer verification of $\Psi_s(Y) \approx (c_{m,n}/s)Y$. **Left**: $\widehat{\alpha}$ vs. the prediction $c_{m,n}/s$ (dashed) as s varies over a decade at $N = 16$; the empirical slope is within 10–15% of the prediction. **Middle**: slope vs. $N \in \{4, 8, 16, 32, 64\}$ at $s = 1$; both the Monte Carlo value of $c_{N,N}/s$ and the random-matrix asymptotic $8/(3\pi s\sqrt{N})$ agree with the empirical points, showing the predicted $N^{-1/2}$ decay. **Right**: entrywise scatter of $\widehat{\Psi}_s(Y)_{ij}$ against $(c_{m,n}/s)Y_{ij}$ at $(N, s, M) = (12, 1, 2\times10^5)$; points lie close to the diagonal $y = x$.

8.3 Matrix quadratic: deterministic, stochastic, and hybrid

For $F(W) = \|AWB - C\|_F^2$ on $\mathbb{R}^{15\times15}$ with $P = A^\top A$ and $Q = BB^\top$ diagonal, with entries geometrically spaced in $[0.1, 1]$, we run GD, deterministic Muon, stochastic Muon, SGD (same noise), and a hybrid that switches from Muon to SGD when $\|\nabla F(W_k)\|_{\mathrm{op}}$ drops below a threshold.

In Figure 3, deterministic Muon is insensitive to the curvature spread while GD is limited by its slowest mode, the empirical Muon/SGD floor ratio is close to the prediction of Theorem 7.1, and the hybrid combines Muon's fast transient with SGD's lower noise floor.

8.4 Noise-layer verification of the smoothed polar slope

We test Theorem 6.1 on the noise layer alone (no optimizer). For square $N \times N$ matrices Y with $\|Y\|_F = 0.1\, s$, we estimate the smoothed polar drift by Monte Carlo,

$$\widehat{\Psi}_s(Y) = \frac{1}{M}\sum_{\ell=1}^{M} \mathrm{msgn}(Y + sZ_\ell), \qquad Z_\ell \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_{N\times N}), \tag{27}$$

and fit the best scalar slope $\widehat{\alpha} = \langle \widehat{\Psi}_s, Y\rangle / \|Y\|_F^2$, comparing it to $c_{m,n}/s$. The choice $\|Y\|_F \ll s$ stays inside the linearization regime while keeping $\|\alpha Y\|_F$ above the Monte Carlo noise $O(1/\sqrt{M})$; $\mathbb{E}\|Z\|_*$ is estimated by an independent run, so the curves are predictions, not fits.

Figure 4 supports the linearization in three ways: the $1/s$ scaling at fixed N, the $N^{-1/2}$ scaling at fixed s, and the entrywise identity $\widehat{\Psi}_s(Y) \approx (c_{m,n}/s)\, Y$. The empirical $c_{N,N}$ and the random-

matrix asymptotic $8/(3\pi\sqrt{N})$ agree down to $N = 8$, consistent with (24). Noise thus turns the discontinuous matrix sign into a smooth, locally linear map at the origin, which is why finite-time behavior is lost in the stochastic regime.

9 Limitations and discussion

Our model is intentionally idealized. Practical Muon uses momentum, Newton–Schulz approximations to the polar factor, layerwise normalization conventions, mixed precision, and often decoupled weight decay. Weight decay changes the problem by introducing an implicit spectral-norm constraint, which is outside the no-decay focus of this paper. Momentum changes the noise filtering and may reduce the effective s in the polar argument. We expect the basic mechanism to persist in these settings: noise inside the matrix sign still smooths the polar map near stationarity.

The finite-time theorem is a continuous-time idealization. Discrete Muon with fixed learning rate cannot converge exactly to a point in finite steps; it reaches a discretization-scale neighborhood and may oscillate. Decaying learning rates or switching to an SGD-like method are natural ways to obtain fine convergence.

Our experiments use controlled objectives (identity quadratic and $\|AWB - C\|_F^2$) where hitting times, the slope $c_{m,n}/s$, and modewise stationary variances can be measured exactly. On neural networks the matrix quadratic is only a local linearization, so the natural next step is to test per-layer hitting times, the empirical drift $\mathbb{E}[\mathrm{msgn}(g(W) + \xi)]$ versus $(c_{m,n}/s)\, g(W)$ at checkpoints, and Muon→SGD switching directly.

10 Conclusion

Deterministic Muon obeys an exact nuclear-norm dissipation identity and converges in finite time under a nuclear Łojasiewicz error bound. On matrix quadratics its hitting time grows only as $1/\sqrt{p_{\min} q_{\min}}$, whereas the exponential rate of gradient flow shrinks in proportion to $p_{\min} q_{\min}$. With minibatch noise, its drift is $\mathbb{E}[\mathrm{msgn}(g(W) + \xi)]$; for isotropic Gaussian noise, this is the gradient of a smoothed nuclear norm and linearizes near an optimum to $(\mathbb{E}\|Z\|_*/(mns))\nabla F$. The local variance comparison on matrix quadratics supports switching from Muon to SGD near stationarity.

Ethics statement

This work studies optimization theory for the Muon optimizer and uses only synthetic matrix problems. It involves no human subjects, personal data, or released models, and we are not aware of ethical concerns specific to it.

Reproducibility statement

All assumptions are stated with the corresponding results in Sections 3–7, and complete proofs are given in Appendix A. Section 8 and the figure captions specify the objectives, dimensions, step sizes, and noise levels of every experiment. The Python scripts that generate all figures will be provided as anonymized supplementary material.

References

Abdelrahman Abouzeid. Does your optimizer care how you normalize? normalization-optimizer coupling in LLM training. *arXiv preprint arXiv:2604.01563*, 2026.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations*, 2026.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 2018.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3920–3930. PMLR, 2025.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 2018.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As Lyapunov predicts. In *International Conference on Learning Representations*, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026.

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

Peng Cheng, Jiucheng Zang, Qingnan Li, Liheng Ma, Yufei Cui, Yingxue Zhang, Boxing Chen, Ming Jian, and Wen Tong. TrasMuon: Trust-region adaptive scaling for orthogonalized momentum optimizers. *arXiv preprint arXiv:2602.13498*, 2026.

Michael Crawshaw, Chirag Modi, Mingrui Liu, and Robert M. Gower. An exploration of non-Euclidean gradient descent: Muon and its many variants. In *International Conference on Machine Learning*, 2026.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Zhehang Du and Weijie Su. The Newton-Muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, https://kellerjordan.github.io/posts/muon/, 2024. December 8, 2024.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 9851 of *Lecture Notes in Computer Science*, pp. 795–811. Springer, 2016. doi: 10.1007/978-3-319-46128-1_50.

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

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

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

Kfir Y. Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110. PMLR, 2017.

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

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon^2: Boosting Muon via adaptive second-moment preconditioning. *arXiv preprint arXiv:2604.09967*, 2026.

Stanisław Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pp. 87–89, Paris, 1963. Éditions du Centre National de la Recherche Scientifique.

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

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Ryan Murray, Brian Swenson, and Soummya Kar. Revisiting normalized gradient descent: Fast evasion of saddle points. *IEEE Transactions on Automatic Control*, 64(11):4818–4824, 2019.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *arXiv preprint arXiv:2505.23737*, 2025.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

A Proofs and additional derivations

We give the proofs deferred from the main text, followed by the SDE derivation and formulas for $c_{m,n}$ and the modewise variances. Notation follows Sections 3–7.

A.1 Proof: nuclear-norm subgradient

This is the standard subdifferential formula for the nuclear norm. If $G = U\Sigma V^\top$, then

$$\partial\|G\|_* = \big\{UV^\top + R : U^\top R = 0,\ RV = 0,\ \|R\|_{\mathrm{op}} \le 1\big\}.$$

Taking $R = 0$ proves $\mathrm{msgn}(G) \in \partial\|G\|_*$ and $\langle G, \mathrm{msgn}(G)\rangle = \|G\|_*$.

A.2 Proof of Proposition 4.1 (nuclear dissipation)

By the chain rule and the nuclear-norm subgradient identity,

$$\frac{d}{dt}F(W_t) = \langle \nabla F(W_t), \dot{W}_t \rangle = -\langle \nabla F(W_t), \operatorname{msgn}(\nabla F(W_t)) \rangle = -\|\nabla F(W_t)\|_*.$$

A.3 Proof of Theorem 4.3 (finite-time convergence)

Let $E(t) = F(W_t) - F^\star$. By (9) and Assumption 4.2,

$$\dot{E}(t) = -\|\nabla F(W_t)\|_* \le -\kappa E(t)^\theta.$$

For $E(t) > 0$,

$$\frac{d}{dt}E(t)^{1-\theta} = (1-\theta)E(t)^{-\theta}\dot{E}(t) \le -\kappa(1-\theta).$$

Therefore $E(t)^{1-\theta} \le E(0)^{1-\theta} - \kappa(1-\theta)t$. The right side reaches zero by the stated time bound, and since $E(t) \ge 0$, $E(t) = 0$ for all later times.

A.4 Proof of Corollary 4.5 (matrix quadratic error bound)

In a basis diagonalizing P and Q, with eigenvalues p_i, q_j and entries e_{ij} of E,

$$F(W) - F^\star = \sum_{ij} p_i q_j e_{ij}^2, \qquad \|\nabla F(W)\|_F^2 = 4\sum_{ij} p_i^2 q_j^2 e_{ij}^2 \ge 4 p_{\min} q_{\min} \sum_{ij} p_i q_j e_{ij}^2.$$

Thus $\|\nabla F(W)\|_* \ge \|\nabla F(W)\|_F \ge 2\sqrt{p_{\min} q_{\min}}\sqrt{F(W) - F^\star}$, and Theorem 4.3 gives the hitting-time bound.

A.5 Proof of Theorem 6.1 (linearization of the Gaussian-smoothed polar map)

We first record the Gaussian shift identity. For $h : \mathbb{R}^d \to \mathbb{R}$ integrable under the standard Gaussian and $q(u) = \mathbb{E}h(Z+u)$,

$$\nabla^2 q(0)[v,v] = \mathbb{E}h(Z)\big((Z^\top v)^2 - \|v\|^2\big), \tag{28}$$

which follows from $q(u) = \int h(z)\varphi(z-u)\,dz$ and differentiating the Gaussian density.

Now let $\phi_s(Y) = \mathbb{E}\|Y + sZ\|_*$ and $\Psi_s = \nabla\phi_s$. For $U \in O(m)$ and $V \in O(n)$, isotropy of Z and unitary invariance of $\|\cdot\|_*$ give

$$\begin{aligned}\phi_s(UYV^\top) &= \mathbb{E}\|UYV^\top + sZ\|_* = \mathbb{E}\|U(Y + sU^\top ZV)V^\top\|_* \\ &= \mathbb{E}\|Y + sU^\top ZV\|_* = \mathbb{E}\|Y + sZ\|_* = \phi_s(Y),\end{aligned}$$

so ϕ_s is invariant under $Y \mapsto UYV^\top$. Let $B(H,K) := D^2\phi_s(0)[H,K]$. Differentiating $\phi_s(UYV^\top) = \phi_s(Y)$ twice at $Y = 0$ gives $B(UHV^\top, UKV^\top) = B(H,K)$ for all H, K and orthogonal U, V. Permutation/sign invariance forces $B(H,K) = a_s\langle H, K\rangle_F$ for some scalar a_s.

Since $\phi_s(tH) = s\mathbb{E}\|Z + (t/s)H\|_*$, applying (28) with $h(Z) = \|Z\|_*$ and $v = H/s$ yields

$$D^2\phi_s(0)[H,H] = \frac{1}{s}\mathbb{E}\big[\|Z\|_*\big(\langle Z, H\rangle^2 - \|H\|_F^2\big)\big]. \tag{29}$$

Averaging (29) over an orthonormal basis $\{E_a\}_{a=1}^d$ of $\mathbb{R}^{m\times n}$ ($d = mn$) and using $\sum_a \langle Z, E_a\rangle^2 = \|Z\|_F^2$ gives

$$da_s = \frac{1}{s}\mathbb{E}\big[\|Z\|_*\big(\|Z\|_F^2 - d\big)\big].$$

Writing $Z = R\Theta$ with $R = \|Z\|_F \sim \chi_d$ and Θ uniform on the sphere, $\|Z\|_* = R\|\Theta\|_*$ with $R \perp \Theta$; using $\mathbb{E}R^3 = (d+1)\mathbb{E}R$ for $R \sim \chi_d$,

$$\mathbb{E}\big[\|Z\|_*\big(\|Z\|_F^2 - d\big)\big] = \mathbb{E}\|Z\|_*.$$

Therefore $a_s = \mathbb{E}\|Z\|_*/(ds) = c_{m,n}/s$. Since $\Psi_s = \nabla\phi_s$, we obtain $D\Psi_s(0)[H] = (c_{m,n}/s)H$ and (23).

A.6 Proof: covariance of the matrix sign at the optimum

Lemma A.1 (Covariance of $\mathrm{msgn}(sZ)$)**.** *Let $Z \in \mathbb{R}^{m\times n}$ have i.i.d. standard Gaussian entries and $r = \min(m,n)$. Then for every $s \neq 0$,*

$$\mathbb{E}\,\mathrm{msgn}(sZ) = 0, \qquad \mathrm{Cov}(\mathrm{vec}\ \mathrm{msgn}(sZ)) = \frac{r}{mn} I_{mn}, \tag{30}$$

independently of s.

Proof. Write $M := \mathrm{msgn}(Z)$ (well defined a.s. since Z has rank r a.s.). By Gaussianity $Z \overset{d}{=} -Z$ and oddness of msgn, $\mathbb{E}M = 0$; for $s \neq 0$, $\mathrm{msgn}(sZ) = \mathrm{sgn}(s)\, M$ and $\mathbb{E}\,\mathrm{msgn}(sZ) = 0$. For diagonal sign matrices D_1, D_2, $D_1 Z D_2 \overset{d}{=} Z$ and $\mathrm{msgn}(D_1 Z D_2) = D_1 M D_2$, so $M \overset{d}{=} D_1 M D_2$. For $(i,j) \neq (k,\ell)$, choosing signs gives $\mathbb{E}[M_{ij} M_{k\ell}] = 0$. Permutation invariance equalizes all entry variances, so $\mathrm{Cov}(\mathrm{vec}\, M) = c\, I_{mn}$. Almost surely $\|M\|_F^2 = r$, hence $\mathrm{tr}\,\mathrm{Cov}(\mathrm{vec}\, M) = \mathbb{E}\|M\|_F^2 = r = cmn$ and $c = r/(mn)$. Sign-flipping by $\mathrm{sgn}(s)$ preserves the covariance. □

A.7 Proof of Theorem 7.1 (near-optimum SDE for $\|AWB - C\|_F^2$)

By Theorem 6.1,

$$\mathbb{E}\,\mathrm{msgn}(2PEQ + sZ) = \frac{c_{m,n}}{s}\, 2PEQ + o(\|PEQ\|_F/s).$$

Vectorizing yields the drift $-(2c_{m,n}/s)(Q \otimes P)\,\mathrm{vec}(E_t)$. Lemma A.1 gives the leading covariance $(r/(mn))I_{mn}$ of $\mathrm{vec}\,\mathrm{msgn}(sZ)$ at the optimum, producing the diffusion in (25).

A.8 Mean polar perturbation away from the optimum

Lemma A.2 (Mean polar perturbation)**.** *Let $Y \in \mathbb{R}^{m\times n}$ have full rank $r = \min(m,n)$, set $\tau := \sigma_{\min}(Y) > 0$, and let $Z \in \mathbb{R}^{m\times n}$ have i.i.d. standard Gaussian entries. There is a constant $C_{m,n} < \infty$, depending only on m and n, such that if $s \le \tau/\big(4(\sqrt{m}+\sqrt{n})\big)$, then*

$$\big\|\mathbb{E}\,\mathrm{msgn}(Y + sZ) - \mathrm{msgn}(Y)\big\|_F \le C_{m,n}\,\frac{s^2}{\tau^2}. \tag{31}$$

Proof. Write $\mathcal{P}(X) = \mathrm{msgn}(X)$ and let A be the event $\{\|sZ\|_{\mathrm{op}} \le \tau/2\}$. If $\|X - Y\|_{\mathrm{op}} \le \tau/2$, Weyl's inequality gives $\sigma_{\min}(X) \ge \tau/2$, so $\mathcal{P}$ is C^2 on this neighborhood with second derivative bounded by $K_{m,n}/\tau^2$. On A, Taylor's theorem gives

$$\mathcal{P}(Y + sZ) = \mathcal{P}(Y) + s\, D\mathcal{P}_Y[Z] + R_2(Y, sZ), \qquad \|R_2(Y,sZ)\|_F \le K_{m,n}\,\frac{s^2\|Z\|_F^2}{\tau^2}.$$

Split $\mathbb{E}[\mathcal{P}(Y+sZ) - \mathcal{P}(Y)]$ over A and its complement A^c.

On A. The event A is invariant under $Z \mapsto -Z$ and $D\mathcal{P}_Y$ is linear, so $\mathbb{E}[D\mathcal{P}_Y[Z]\mathbf{1}_A] = 0$. Hence

$$\big\|\mathbb{E}[(\mathcal{P}(Y+sZ) - \mathcal{P}(Y))\mathbf{1}_A]\big\|_F \le K_{m,n}\,\frac{s^2\,\mathbb{E}\|Z\|_F^2}{\tau^2} = K_{m,n}\,\frac{s^2 mn}{\tau^2}.$$

On A^c. Since $\|\mathcal{P}(X)\|_F \le \sqrt{r}$ for every X, $\|\mathcal{P}(Y+sZ) - \mathcal{P}(Y)\|_F \le 2\sqrt{r}$. Gordon's bound $\mathbb{E}\|Z\|_{\mathrm{op}} \le \sqrt{m} + \sqrt{n}$ (Vershynin, 2018, Section 7.3) and Gaussian concentration for the 1-Lipschitz map $Z \mapsto \|Z\|_{\mathrm{op}}$ (Boucheron et al., 2013, Theorem 5.6) give $\mathbb{P}(\|Z\|_{\mathrm{op}} \ge \sqrt{m}+\sqrt{n}+t) \le e^{-t^2/2}$ for $t \ge 0$. The assumption on s gives $\tau/(2s) \ge 2(\sqrt{m}+\sqrt{n})$, so $t := \tau/(2s) - (\sqrt{m}+\sqrt{n}) \ge \tau/(4s)$ and

$$\mathbb{P}(A^c) = \mathbb{P}\big(\|Z\|_{\mathrm{op}} > \tau/(2s)\big) \le e^{-\tau^2/(32s^2)} \le \frac{32\,s^2}{\tau^2},$$

using $e^{-u} \le 1/u$ for $u > 0$. Hence $\|\mathbb{E}[(\mathcal{P}(Y+sZ) - \mathcal{P}(Y))\mathbf{1}_{A^c}]\|_F \le 64\sqrt{r}\,s^2/\tau^2$.

Combining the two parts gives (31) with $C_{m,n} = K_{m,n}mn + 64\sqrt{r}$. □

Interpretation. Far from the optimum, $\sigma_{\min}(g(W))$ is large compared with $s(\sqrt{m}+\sqrt{n})$ and the bias is second order in s. Near the optimum the expansion fails, and Theorem 6.1 replaces it.

A.9 STOCHASTIC-MODIFIED-EQUATION DERIVATION

We give the moment-matching derivation of (15)–(16). Fix $W_k = W$ and let ξ have the conditional law of ξ_k. For stochastic Muon (14),

$$\mathbb{E}[W_{k+1} - W_k \mid W_k = W] = -\eta\,\Psi(W), \qquad \mathrm{Cov}\big(\mathrm{vec}(W_{k+1} - W_k) \mid W_k = W\big) = \eta^2 C_M(W),$$

where $\Psi(W) = \mathbb{E}[\mathrm{msgn}(g(W)+\xi)]$ and $C_M(W) = \mathrm{Cov}(\mathrm{vec}\,\mathrm{msgn}(g(W)+\xi))$. Setting $t = k\eta$ and $dt = \eta$, matching these to an Itô diffusion $d\,\mathrm{vec}(W_t) = -\,\mathrm{vec}(\Psi)\,dt + \sigma_M\,dB_t$ with $\sigma_M\sigma_M^\top = \eta C_M$ recovers (15). For SGD,

$$\mathbb{E}[W_{k+1} - W_k \mid W_k = W] = -\eta g(W), \qquad \mathrm{Cov}\big(\mathrm{vec}(W_{k+1} - W_k) \mid W_k = W\big) = \frac{\eta^2}{B}\Sigma_g(W),$$

which by the same argument gives (16).

A.10 CLOSED FORMS AND ASYMPTOTICS FOR $c_{m,n}$

The coefficient $c_{m,n} = \mathbb{E}\|Z\|_*/(mn)$ is the expected nuclear norm per matrix entry. The square-matrix asymptotic $c_{N,N} \sim 8/(3\pi\sqrt{N})$ is given in the main text (24); the remaining cases are recorded here.

Scalar case ($m = n = 1$). $c_{1,1} = \mathbb{E}|Z| = \sqrt{2/\pi}$. In particular, $\mathbb{E}\,\mathrm{sign}(y + sZ) = \sqrt{2/\pi}\,(y/s) + o(y/s)$.

Vector case ($m = 1$, $n = d$).

$$c_{1,d} = \frac{1}{d}\mathbb{E}\chi_d = \frac{\sqrt{2}}{d}\frac{\Gamma((d+1)/2)}{\Gamma(d/2)} \sim \frac{1}{\sqrt{d}}.$$

Rectangular case ($m \le n$, $m/n \to \gamma \in (0,1]$). Set $a = (1-\sqrt{\gamma})^2$, $b = (1+\sqrt{\gamma})^2$. Then

$$\mathbb{E}\|Z\|_* \sim m\sqrt{n}\,I(\gamma), \qquad I(\gamma) = \int_a^b \sqrt{x}\,\frac{\sqrt{(b-x)(x-a)}}{2\pi\gamma x}\,dx,$$

so $c_{m,n} \sim I(\gamma)/\sqrt{n}$, with $I(1) = 8/(3\pi)$ recovering the square case and $I(\gamma) \to 1$ as $\gamma \to 0$.

A.11 MODEWISE OU FORMULAS

Diagonalize $P = U\,\mathrm{diag}(p_i)U^\top$ and $Q = V\,\mathrm{diag}(q_j)V^\top$, and write $\widetilde{E} = U^\top E V$ with entries $\widetilde{e}_{ij}$. Linearizing (25)–(26) yields the scalar OU processes

$$d\widetilde{e}_{ij} = -\kappa_{ij}^{\mathrm{Muon}}\widetilde{e}_{ij}\,dt + \sqrt{\eta\,r/(mn)}\,dB_{ij,t}, \qquad \kappa_{ij}^{\mathrm{Muon}} = \frac{2c_{m,n}}{s}p_iq_j, \tag{32}$$

$$d\widetilde{e}_{ij} = -2p_iq_j\widetilde{e}_{ij}\,dt + \sqrt{\eta}\,s\,dB_{ij,t}, \qquad \kappa_{ij}^{\mathrm{SGD}} = 2p_iq_j. \tag{33}$$

For an OU process $dX = -\kappa X\,dt + \sigma\,dB$ the stationary variance is $\sigma^2/(2\kappa)$, hence

$$\mathrm{Var}_{\mathrm{Muon}}(\widetilde{e}_{ij}) = \frac{\eta\,r/(mn)}{2\kappa_{ij}^{\mathrm{Muon}}} = \frac{\eta\,r\,s}{4\,mn\,c_{m,n}\,p_iq_j}, \qquad \mathrm{Var}_{\mathrm{SGD}}(\widetilde{e}_{ij}) = \frac{\eta\,s^2}{4\,p_iq_j}, \tag{34}$$

so the contraction-rate and stationary-variance ratios

$$\frac{\kappa_{ij}^{\mathrm{Muon}}}{\kappa_{ij}^{\mathrm{SGD}}} = \frac{c_{m,n}}{s}, \qquad \frac{\mathrm{Var}_{\mathrm{Muon}}(\widetilde{e}_{ij})}{\mathrm{Var}_{\mathrm{SGD}}(\widetilde{e}_{ij})} = \frac{r}{s\,\mathbb{E}\|Z\|_*}, \tag{35}$$

are both independent of the mode curvature p_iq_j. For square $N \times N$ matrices, $c_{N,N} \sim 8/(3\pi\sqrt{N})$ gives

$$\frac{\mathrm{Var}_{\mathrm{Muon}}}{\mathrm{Var}_{\mathrm{SGD}}} \sim \frac{3\pi}{8\,s\sqrt{N}}.$$

These ratios compare the two methods at the same learning rate η. Since the learning rates of Muon and SGD have different natural scales, a comparison that avoids this choice runs SGD with learning rate $\eta' = \eta\, c_{m,n}/s$, so that both methods have the same local contraction rate $\kappa_{ij}^{\mathrm{Muon}}\eta = \kappa_{ij}^{\mathrm{SGD}}\eta'$ in every mode. Then $\mathrm{Var}_{\mathrm{SGD}}(\widetilde{e}_{ij}) = \eta' s^2/(4p_i q_j)$ and

$$\left.\frac{\mathrm{Var}_{\mathrm{Muon}}(\widetilde{e}_{ij})}{\mathrm{Var}_{\mathrm{SGD}}(\widetilde{e}_{ij})}\right|_{\text{matched rate}} = \frac{r}{mn\, c_{m,n}^2} = \frac{r\, mn}{(\mathbb{E}\|Z\|_*)^2} \geq 1, \tag{36}$$

where the inequality follows from $\|Z\|_* \leq \sqrt{r}\,\|Z\|_F$ and $\mathbb{E}\|Z\|_F \leq (\mathbb{E}\|Z\|_F^2)^{1/2} = \sqrt{mn}$. The ratio is independent of s, η, and the mode; it equals $\pi/2$ for $m = n = 1$, tends to 1 in the vector case, and tends to $9\pi^2/64 \approx 1.39$ for square matrices.

The local quadratic objective floor $F(W^\star + E) - F^\star = \sum_{ij} p_i q_j \widetilde{e}_{ij}^2$ inherits the same ratios, because each mode contributes $p_i q_j\, \mathrm{Var}(\widetilde{e}_{ij})$, in which the factor $p_i q_j$ cancels.

B AI USE STATEMENT

We used generative AI tools to proofread and edit the manuscript, to check the proofs, and to check the bibliographic details of the references. The authors reviewed all AI-assisted changes and checks, and take responsibility for the full content of the paper.