Muon with Finite Newton–Schulz: The Smoothing Benefit in Nonsmooth Nonconvex Optimization

arXiv:2608.26288

Mingyi Li[1] **Taira Tsuchiya**[1,2]

[1]The University of Tokyo [2]RIKEN

RIKEN AIP Workshop on LLM×ML 2026, November 11, 2026

Muon optimizer

Muon (Jordan et al., 2024): an optimizer for matrix-shaped parameters that steps along the **orthogonalized momentum**

Muon optimizer

Muon (Jordan et al., 2024): an optimizer for matrix-shaped parameters that steps along the **orthogonalized momentum**

A strong alternative to AdamW in LLM pretraining

- substantial efficiency gains in scaling studies (Liu et al., 2025)
- variants used to train frontier models such as Kimi K2 and GLM-4.5 (Kimi Team, 2025; GLM-4.5 Team, 2025)
- among the strongest in a systematic pretraining benchmark (Wen et al., 2026)

The Muon update

For a matrix parameter $W_t \in \mathbb{R}^{m \times n}$ and a stochastic gradient G_t:

$$M_t = \beta M_{t-1} + (1-\beta) G_t \qquad \text{(momentum, } M_0 = 0\text{)}$$

$$W_{t+1} = W_t - D \ \text{polar}(M_t) \qquad (D > 0)$$

The Muon update

For a matrix parameter $W_t \in \mathbb{R}^{m \times n}$ and a stochastic gradient G_t:

$$M_t = \beta M_{t-1} + (1-\beta) G_t \qquad \text{(momentum, } M_0 = 0\text{)}$$
$$W_{t+1} = W_t - D\ \mathrm{polar}(M_t) \qquad (D > 0)$$

For a thin SVD $M = U\Sigma V^\top$ ($\Sigma \succ 0$), the **polar factor** (orthogonalized M) is

$$\mathrm{polar}(M) = UV^\top.$$

The polar map keeps the singular vectors and sets every positive singular value to 1 (zero singular values stay 0).

The Muon update

For a matrix parameter $W_t \in \mathbb{R}^{m \times n}$ and a stochastic gradient G_t:

$$M_t = \beta M_{t-1} + (1-\beta) G_t \qquad \text{(momentum, } M_0 = 0\text{)}$$
$$W_{t+1} = W_t - D \ \text{polar}(M_t) \qquad (D > 0)$$

For a thin SVD $M = U\Sigma V^\top$ ($\Sigma \succ 0$), the **polar factor** (orthogonalized M) is

$$\text{polar}(M) = UV^\top.$$

The polar map keeps the singular vectors and sets every positive singular value to 1 (zero singular values stay 0).

Computing polar(M) with an SVD is too expensive at LLM scale.

Computing the polar factor: Newton–Schulz

Muon approximates polar(Y) with a few **Newton–Schulz** iterations (Kovarik, 1970; Björck and Bowie, 1971), using only matrix–matrix multiplications:

$$Y \leftarrow \tfrac{15}{8} Y - \tfrac{5}{4}(YY^\top)Y + \tfrac{3}{8}(YY^\top)^2 Y \qquad (Y \text{ normalized so that } \|Y\|_{\mathrm{op}} \le 1)$$

Computing the polar factor: Newton–Schulz

Muon approximates polar(Y) with a few **Newton–Schulz** iterations (Kovarik, 1970; Björck and Bowie, 1971), using only matrix–matrix multiplications:

$$Y \leftarrow \tfrac{15}{8} Y - \tfrac{5}{4}(YY^\top)Y + \tfrac{3}{8}(YY^\top)^2 Y \qquad (Y \text{ normalized so that } \|Y\|_{\mathrm{op}} \le 1)$$

Idea: $\mathrm{polar}(Y) = (YY^\top)^{-1/2}\, Y \xrightarrow{\text{Taylor expansion around } \lambda = 1} \lambda^{-1/2} \approx \tfrac{15}{8} - \tfrac{5}{4}\lambda + \tfrac{3}{8}\lambda^2$

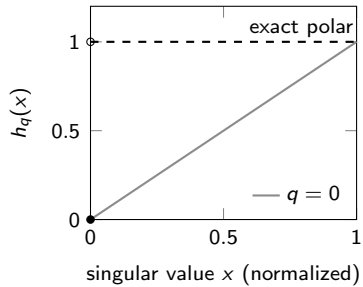
Computing the polar factor: Newton–Schulz

Muon approximates polar(Y) with a few **Newton–Schulz** iterations (Kovarik, 1970; Björck and Bowie, 1971), using only matrix–matrix multiplications:

$$Y \leftarrow \tfrac{15}{8} Y - \tfrac{5}{4}(YY^\top)Y + \tfrac{3}{8}(YY^\top)^2 Y \qquad (Y \text{ normalized so that } \|Y\|_{\mathrm{op}} \le 1)$$

Idea: $\mathrm{polar}(Y) = (YY^\top)^{-1/2} Y \xrightarrow{\text{Taylor expansion around } \lambda = 1} \lambda^{-1/2} \approx \tfrac{15}{8} - \tfrac{5}{4}\lambda + \tfrac{3}{8}\lambda^2$

One step on a singular value x: $f(x) := \tfrac{15}{8}x - \tfrac{5}{4}x^3 + \tfrac{3}{8}x^5$, $\quad$ q **steps**: $h_q := f^{\circ q}$



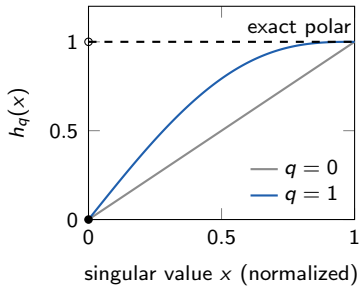
Computing the polar factor: Newton–Schulz

Muon approximates polar(Y) with a few **Newton–Schulz** iterations (Kovarik, 1970; Björck and Bowie, 1971), using only matrix–matrix multiplications:

$$Y \leftarrow \tfrac{15}{8} Y - \tfrac{5}{4}(YY^\top)Y + \tfrac{3}{8}(YY^\top)^2 Y \qquad (Y \text{ normalized so that } \|Y\|_{\mathrm{op}} \le 1)$$

Idea: $\mathrm{polar}(Y) = (YY^\top)^{-1/2} Y \xrightarrow{\text{Taylor expansion around } \lambda = 1} \lambda^{-1/2} \approx \tfrac{15}{8} - \tfrac{5}{4}\lambda + \tfrac{3}{8}\lambda^2$

One step on a singular value x: $f(x) := \tfrac{15}{8}x - \tfrac{5}{4}x^3 + \tfrac{3}{8}x^5$, $\quad$ q **steps**: $h_q := f^{\circ q}$



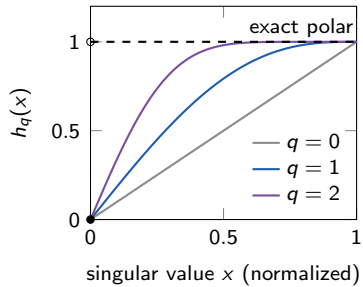
Computing the polar factor: Newton–Schulz

Muon approximates polar(Y) with a few **Newton–Schulz** iterations (Kovarik, 1970; Björck and Bowie, 1971), using only matrix–matrix multiplications:

$$Y \leftarrow \tfrac{15}{8} Y - \tfrac{5}{4}(YY^\top)Y + \tfrac{3}{8}(YY^\top)^2 Y \qquad (Y \text{ normalized so that } \|Y\|_{\mathrm{op}} \le 1)$$

Idea: $\mathrm{polar}(Y) = (YY^\top)^{-1/2} Y \xrightarrow{\text{Taylor expansion around } \lambda = 1} \lambda^{-1/2} \approx \tfrac{15}{8} - \tfrac{5}{4}\lambda + \tfrac{3}{8}\lambda^2$

One step on a singular value x: $f(x) := \tfrac{15}{8}x - \tfrac{5}{4}x^3 + \tfrac{3}{8}x^5$, $\qquad$ q **steps**: $h_q := f^{\circ q}$



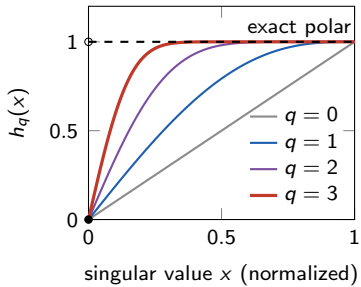
Computing the polar factor: Newton–Schulz

Muon approximates polar(Y) with a few **Newton–Schulz** iterations (Kovarik, 1970; Björck and Bowie, 1971), using only matrix–matrix multiplications:

$$Y \leftarrow \tfrac{15}{8} Y - \tfrac{5}{4}(YY^\top)Y + \tfrac{3}{8}(YY^\top)^2 Y \qquad (Y \text{ normalized so that } \|Y\|_{\mathrm{op}} \le 1)$$

Idea: $\mathrm{polar}(Y) = (YY^\top)^{-1/2} Y \xrightarrow{\text{Taylor expansion around } \lambda = 1} \lambda^{-1/2} \approx \tfrac{15}{8} - \tfrac{5}{4}\lambda + \tfrac{3}{8}\lambda^2$

One step on a singular value x: $f(x) := \tfrac{15}{8}x - \tfrac{5}{4}x^3 + \tfrac{3}{8}x^5$, $\quad$ q **steps**: $h_q := f^{\circ q}$



Muon runs only a few steps: **finite Newton–Schulz**.

Existing convergence analyses of Muon

Almost all convergence analyses assume a smooth objective.

- Most analyses replace Newton–Schulz with the exact polar factor (Li and Hong, 2025; Kovalev, 2025; Pethick et al., 2025; Shen et al., 2026).
- Analyses that keep finite Newton–Schulz treat it as an approximation error (Kim and Oh, 2026; Shulgin et al., 2026; Choudhury et al., 2026)
  $\rightarrow$ **guarantees improve as the iteration approaches the exact polar factor**

Existing convergence analyses of Muon (cont'd)

What about nonsmooth objectives? ReLU, max-pooling, ℓ_1 penalties, hinge and absolute losses

Existing convergence analyses of Muon (cont'd)

What about nonsmooth objectives? ReLU, max-pooling, ℓ_1 penalties, hinge and absolute losses

View the optimizer as an online learner that chooses the update $X_t = W_t - W_{t-1}$ (online-to-nonconvex conversion, explained later)

Exact-polar Muon $W_{t+1} = W_t - D\ \mathrm{polar}(M_t)$ chooses the update

$$X_{t+1} = -D\ \mathrm{polar}(M_t) \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \langle M_t, X \rangle = \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \sum_{s=1}^{t} \beta^{t-s} \langle G_s, X \rangle$$

Existing convergence analyses of Muon (cont'd)

What about nonsmooth objectives? ReLU, max-pooling, ℓ_1 penalties, hinge and absolute losses

View the optimizer as an online learner that chooses the update $X_t = W_t - W_{t-1}$ (online-to-nonconvex conversion, explained later)

Exact-polar Muon $W_{t+1} = W_t - D\ \mathrm{polar}(M_t)$ chooses the update

$$X_{t+1} = -D\ \mathrm{polar}(M_t) \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \langle M_t, X \rangle = \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \sum_{s=1}^{t} \beta^{t-s} \langle G_s, X \rangle$$

exact-polar Muon = follow-the-leader (FTL)

FTL can suffer linear regret, so there is no stationarity guarantee

Existing convergence analyses of Muon (cont'd)

What about nonsmooth objectives? ReLU, max-pooling, ℓ_1 penalties, hinge and absolute losses

View the optimizer as an online learner that chooses the update $X_t = W_t - W_{t-1}$
(online-to-nonconvex conversion, explained later)

Exact-polar Muon $W_{t+1} = W_t - D\ \mathrm{polar}(M_t)$ chooses the update

$$X_{t+1} = -D\ \mathrm{polar}(M_t) \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \langle M_t, X \rangle = \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \sum_{s=1}^{t} \beta^{t-s} \langle G_s, X \rangle$$

exact-polar Muon = follow-the-leader (FTL)

FTL can suffer linear regret, so there is no stationarity guarantee

Jiang et al. (2026) propose Pion and Leon, which replace the learner
→ **extra ingredients that Muon lacks, and finite Newton–Schulz is not covered**

Research question

Existing analyses either
treat finite Newton–Schulz only as an error and need smoothness, or
handle nonsmooth objectives but replace the update Muon actually performs.

Research question

Existing analyses either
treat finite Newton–Schulz only as an error and need smoothness, or
handle nonsmooth objectives but replace the update Muon actually performs.

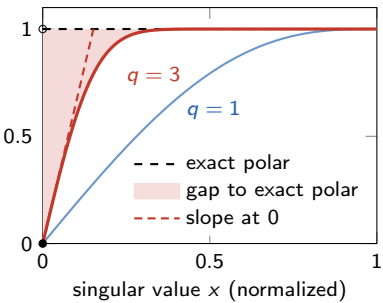
How does the finite Newton–Schulz iteration affect optimization?
Can it be a **benefit** rather than an error?
Can Muon find stationary points of nonsmooth nonconvex objectives?

Main result

Theorem (informal). In stochastic nonsmooth nonconvex optimization, Muon with momentum and $q = O(\log(1/\varepsilon))$ Newton–Schulz steps finds a stationary point with optimal number of gradient queries up to dimension-dependent constants.

Main result

Theorem (informal). In stochastic nonsmooth nonconvex optimization, Muon with momentum and $q = O(\log(1/\varepsilon))$ Newton–Schulz steps finds a stationary point with optimal number of gradient queries up to dimension-dependent constants.



Finite Newton–Schulz **smooths** the polar map.

- Exact polar ($q = \infty$) = FTL
- No iteration ($q = 0$)
  = follow-the-regularized-leader (FTRL)
  with the squared Frobenius regularizer
- $q = O(\log(1/\varepsilon))$ balances the two

Further contributions

1. **Optimal in ε for smooth objectives**
   $O(\varepsilon^{-2})$ gradient evaluations without noise, $O(\varepsilon^{-4})$ with noise

2. **Muon with finite Newton–Schulz is FTRL**
   the number of steps q controls the amount of regularization

3. **The analysis extends beyond Newton–Schulz**
   to other Lipschitz smoothings of the polar map, such as SoftMuon

Setup: Stochastic nonsmooth nonconvex optimization

$$\min_{W \in \mathbb{R}^{m \times n}} \mathcal{L}(W) := \mathbb{E}_{\zeta \sim \mathcal{D}}[\ell(W; \zeta)], \qquad m \leq n,$$

where $\ell(W; \zeta)$ is the loss on a sample ζ from the data distribution $\mathcal{D}$.

Assumption. $\mathcal{L}$ is differentiable (*), and $\Delta_{\mathcal{L}} := \mathcal{L}(W_0) - \inf_W \mathcal{L}(W) < \infty$. At every query point W, the oracle, using randomness independent of the past, returns G with

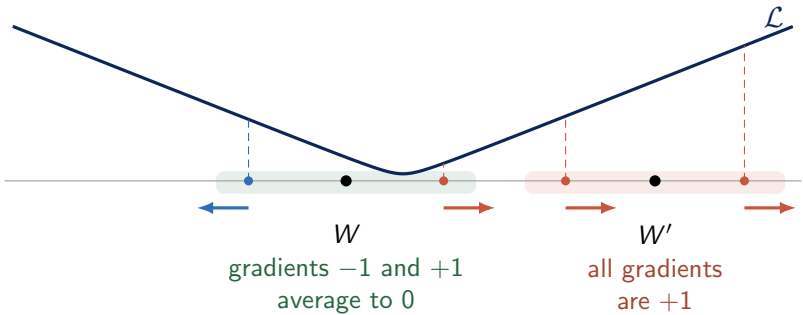
$$\mathbb{E}[G \mid W] = \nabla \mathcal{L}(W), \qquad \mathbb{E}[\|G\|_{\mathrm{F}}^2 \mid W] \leq \Gamma^2, \qquad \mathbb{E}[\|G - \nabla \mathcal{L}(W)\|_{\mathrm{F}}^2 \mid W] \leq \sigma^2.$$

(*) Nondifferentiable points can be handled by a random perturbation.

- In particular, $\|\nabla \mathcal{L}(W)\|_{\mathrm{F}} \leq \Gamma$, so $\mathcal{L}$ is Γ-Lipschitz.
- **No smoothness** (Lipschitz gradient) is assumed.

Stationarity without smoothness

In the smooth case, the target is $\|\nabla \mathcal{L}(W)\|_* \leq \varepsilon$ ($\|\cdot\|_*$: nuclear norm).
Without smoothness, this is too strict: near a kink, the gradient norm stays ≈ 1 even very close to the minimizer.



Gradient norm at the point: ≈ 1 at both W and W'.
Average of the nearby gradients shown: ≈ 0 at W, ≈ 1 at W'.

(ρ, ε)-stationary point

W is (ρ, ε)-stationary if the gradient, averaged over a suitable distribution centered at W with average distance at most ρ, has norm at most ε.

(ρ, ε)-stationary point

W is (ρ, ε)-stationary if the gradient, averaged over a suitable distribution centered at W with average distance at most ρ, has norm at most ε.

Definition (Jiang et al., 2026)**.**

$$\mathcal{P}(W; \rho) \coloneqq \left\{ p \text{ finitely supported} \;\middle|\; \mathbb{E}_{Y \sim p}[Y] = W, \;\; \mathbb{E}_{Y \sim p} \| Y - W \|_{\mathrm{op}} \leq \rho \right\}.$$

W is a (ρ, ε)-**stationary point** if

$$\| \nabla \mathcal{L}(W) \|_*^{[\rho]} \coloneqq \inf_{p \in \mathcal{P}(W; \rho)} \| \mathbb{E}_{Y \sim p} \nabla \mathcal{L}(Y) \|_* \leq \varepsilon.$$

A Goldstein-type relaxation (Goldstein, 1977), unavoidable for general Lipschitz objectives (Zhang et al., 2020; Kornowski and Shamir, 2022).

Comparison with stationarity in the smooth case

In the smooth case, the goal is an ε-**stationary point**: $\|\nabla \mathcal{L}(W)\|_* \leq \varepsilon$.

Comparison with stationarity in the smooth case

In the smooth case, the goal is an ε-**stationary point**: $\|\nabla \mathcal{L}(W)\|_* \leq \varepsilon$.
Since $\|\nabla \mathcal{L}(W)\|_*^{[\rho]} \leq \|\nabla \mathcal{L}(W)\|_*$ (take the point mass at W), every ε-stationary point is (ρ, ε)-stationary.

Comparison with stationarity in the smooth case

In the smooth case, the goal is an ε-**stationary point**: $\|\nabla \mathcal{L}(W)\|_* \le \varepsilon$.
Since $\|\nabla \mathcal{L}(W)\|_*^{[\rho]} \le \|\nabla \mathcal{L}(W)\|_*$ (take the point mass at W), every ε-stationary point is (ρ, ε)-stationary.

Lemma. If $\|\nabla \mathcal{L}(W) - \nabla \mathcal{L}(W')\|_* \le L_{\mathrm{op}} \|W - W'\|_{\mathrm{op}}$ for all W, W', then

$$\|\nabla \mathcal{L}(W)\|_* \le \|\nabla \mathcal{L}(W)\|_*^{[\rho]} + L_{\mathrm{op}} \rho.$$

In particular, with $\rho = \varepsilon / L_{\mathrm{op}}$, a (ρ, ε)-stationary point is 2ε-stationary.

Muon in our analysis

Input: momentum $\beta \in (0,1)$, radius $D > 0$, operator-norm bound $G_{\mathrm{op}} > 0$, depth q
$M_0 = 0, \quad X_1 = 0$
for $t = 1, 2, \ldots$:
receive G_t
$M_t \leftarrow \beta M_{t-1} + (1-\beta) G_t$ — momentum
$Y \leftarrow M_t / G_{\mathrm{op}}$ — normalization
repeat q times:
$Y \leftarrow \frac{15}{8} Y - \frac{5}{4} (YY^\top) Y + \frac{3}{8} (YY^\top)^2 Y$ — Newton–Schulz step
$X_{t+1} \leftarrow -D\, Y \qquad (W_{t+1} \leftarrow W_t - D\, Y)$ — next update

Muon in our analysis

Input: momentum $\beta \in (0,1)$, radius $D > 0$, operator-norm bound $G_{\mathrm{op}} > 0$, depth q
$M_0 = 0, \;\; X_1 = 0$
for $t = 1, 2, \ldots$**:**
  receive G_t
  $M_t \leftarrow \beta M_{t-1} + (1-\beta) G_t$ — momentum
  $Y \leftarrow M_t / G_{\mathrm{op}}$ — normalization
  repeat q times:
    $Y \leftarrow \frac{15}{8} Y - \frac{5}{4}(YY^\top)Y + \frac{3}{8}(YY^\top)^2 Y$ — Newton–Schulz step
  $X_{t+1} \leftarrow -D\,Y \qquad (W_{t+1} \leftarrow W_t - D\,Y)$ — next update

Assume $\|G_t\|_{\mathrm{op}} \le G_{\mathrm{op}}$ a.s. Then $\|M_t\|_{\mathrm{op}} \le G_{\mathrm{op}}$, so the singular values of M_t / G_{op} lie in $[0,1]$, and

$$X_{t+1} = -D\,U\,\mathrm{diag}\big(h_q(\sigma_1), \ldots, h_q(\sigma_r)\big)\,V^\top \qquad \text{for the thin SVD } M_t/G_{\mathrm{op}} = U\Sigma V^\top$$

$h_q = f^{\circ q}$ with $f(x) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$. Exact polar: $h(x) = \mathbf{1}\{x > 0\}$.

An optimizer as an online learner

Optimizer
- choose an update X_t
- $W_t \leftarrow W_{t-1} + X_t$
- receive a stochastic gradient G_t

$=$

Online learner
- play an action (the update) X_t
- suffer the loss $\langle G_t, X_t \rangle$
- observe the loss $X \mapsto \langle G_t, X \rangle$

An optimizer as an online learner

Optimizer
- choose an update X_t
- $W_t \leftarrow W_{t-1} + X_t$
- receive a stochastic gradient G_t

$=$

Online learner
- play an action (the update) X_t
- suffer the loss $\langle G_t, X_t \rangle$
- observe the loss $X \mapsto \langle G_t, X \rangle$

Where the linear loss comes from:

$\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) = \mathcal{L}(W_{t-1} + X_t) - \mathcal{L}(W_{t-1}) \approx \langle G_t, X_t \rangle$

$\rightarrow$ total loss $\sum_t \langle G_t, X_t \rangle \approx \mathcal{L}(W_T) - \mathcal{L}(W_0)$: the smaller the loss, the more $\mathcal{L}$ decreases

An optimizer as an online learner

Optimizer
- choose an update X_t
- $W_t \leftarrow W_{t-1} + X_t$
- receive a stochastic gradient G_t

$=$

Online learner
- play an action (the update) X_t
- suffer the loss $\langle G_t, X_t \rangle$
- observe the loss $X \mapsto \langle G_t, X \rangle$

Where the linear loss comes from:

$\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) = \mathcal{L}(W_{t-1} + X_t) - \mathcal{L}(W_{t-1}) \approx \langle G_t, X_t \rangle$

$\rightarrow$ total loss $\sum_t \langle G_t, X_t \rangle \approx \mathcal{L}(W_T) - \mathcal{L}(W_0)$: the smaller the loss, the more $\mathcal{L}$ decreases

Online-to-nonconvex conversion (O2NC) (Cutkosky et al., 2023)
lets an online learner choose the updates and converts a regret bound into a stationarity guarantee

- the first optimal rate for stochastic nonsmooth nonconvex optimization
- optimal rates for smooth objectives as well, both deterministic and stochastic

Discounted regret

Standard goal in online linear optimization

Regret

$$\mathrm{Reg}_t(D) := \sum_{s=1}^{t} \langle G_s,\, X_s \rangle \quad - \quad \min_{\|U\|_{\mathrm{op}} \le D} \sum_{s=1}^{t} \langle G_s,\, U \rangle$$

cumulative loss of the learner

cumulative loss of the best fixed update

Discounted regret

Standard goal in online linear optimization

Regret

$$\mathrm{Reg}_t(D) := \sum_{s=1}^{t} \langle G_s,\, X_s \rangle \quad - \quad \min_{\|U\|_{\mathrm{op}} \le D} \sum_{s=1}^{t} \langle G_s,\, U \rangle$$

cumulative loss of the learner

cumulative loss of the best fixed update

Measure the regret with the same weights β^{t-s} as Muon's momentum

Discounted regret

$$\mathrm{Reg}_t^{[\beta]}(D) := \sum_{s=1}^{t} \beta^{t-s} \langle G_s,\, X_s \rangle \quad - \quad \min_{\|U\|_{\mathrm{op}} \le D} \sum_{s=1}^{t} \beta^{t-s} \langle G_s,\, U \rangle$$

Conversion guarantee

Lemma (Jiang et al., 2026). With $D = (1-\beta)\rho/(4\beta)$, discounted O2NC outputs $\bar{W}_\tau$ with

$$\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]} \le \underbrace{\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}}_{\text{initial suboptimality}} + \underbrace{\frac{1}{T}\mathbb{E}\Big[\mathrm{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t<T}\mathrm{Reg}_t^{[\beta]}(1)\Big]}_{\text{discounted regret of the learner}}$$
$$+ \underbrace{\Big(1-\beta+\frac{\beta}{T}\Big)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}}}_{\text{noise}}.$$

$\mathrm{Reg}_t^{[\beta]}(1)$: discounted regret of X_s/D against $\|U\|_{\mathrm{op}} \le 1$, so that $\mathrm{Reg}_t^{[\beta]}(D) = D \cdot \mathrm{Reg}_t^{[\beta]}(1)$.

Conversion guarantee

Lemma (Jiang et al., 2026). With $D = (1-\beta)\rho/(4\beta)$, discounted O2NC outputs $\bar{W}_\tau$ with

$$\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]} \le \underbrace{\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}}_{\text{initial suboptimality}} + \underbrace{\frac{1}{T}\mathbb{E}\Big[\mathrm{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t<T}\mathrm{Reg}_t^{[\beta]}(1)\Big]}_{\text{discounted regret of the learner}}$$
$$+ \underbrace{\Big(1-\beta+\frac{\beta}{T}\Big)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}}}_{\text{noise}}.$$

$\mathrm{Reg}_t^{[\beta]}(1)$: discounted regret of X_s/D against $\|U\|_{\mathrm{op}} \le 1$, so that $\mathrm{Reg}_t^{[\beta]}(D) = D \cdot \mathrm{Reg}_t^{[\beta]}(1)$.

If $\boxed{\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}}$ for all t, where $B \asymp \sqrt{r}\,(\Gamma+\sigma)$, then

$1-\beta \asymp (\varepsilon/B)^2$ and $\boxed{T \asymp \Delta_{\mathcal{L}} B^2/(\rho\varepsilon^3) + B^2/\varepsilon^2}$ make the bound $\lesssim \varepsilon$.

Conversion guarantee

Lemma (Jiang et al., 2026). With $D = (1-\beta)\rho/(4\beta)$, discounted O2NC outputs $\bar{W}_\tau$ with

$$\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]} \le \underbrace{\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}}_{\text{initial suboptimality}} + \underbrace{\frac{1}{T}\mathbb{E}\Big[\mathrm{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t<T}\mathrm{Reg}_t^{[\beta]}(1)\Big]}_{\text{discounted regret of the learner}}$$
$$+ \underbrace{\Big(1-\beta+\frac{\beta}{T}\Big)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}}}_{\text{noise}}.$$

$\mathrm{Reg}_t^{[\beta]}(1)$: discounted regret of X_s/D against $\|U\|_{\mathrm{op}} \le 1$, so that $\mathrm{Reg}_t^{[\beta]}(D) = D \cdot \mathrm{Reg}_t^{[\beta]}(1)$.

If $\boxed{\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}}$ for all t, where $B \asymp \sqrt{r}\,(\Gamma + \sigma)$, then

$1-\beta \asymp (\varepsilon/B)^2$ and $\boxed{T \asymp \Delta_{\mathcal{L}} B^2/(\rho\varepsilon^3) + B^2/\varepsilon^2}$ make the bound $\lesssim \varepsilon$.

Next: the discounted regret of Muon with finite Newton–Schulz.

Muon with a general spectral map

Replace h_q by a **spectral map** h.
For the thin SVD $M_t/G_{\mathrm{op}} = U\Sigma V^\top$,

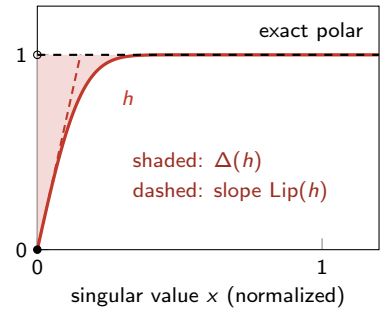
$$X_{t+1} = -D\, U\, \mathrm{diag}\big(h(\sigma_1), \ldots, h(\sigma_r)\big)\, V^\top.$$

Assumption. $h\colon [0,\infty) \to [0,1]$ is continuous and nondecreasing, with

(i) $h(0) = 0$

(ii) Lipschitz constant $\mathrm{Lip}(h) < \infty$

(iii) $\Delta(h) := \int_0^\infty \big(1 - h(x)\big)\, \mathrm{d}x < \infty$



Muon with a general spectral map

Replace h_q by a **spectral map** h.
For the thin SVD $M_t/G_{\mathrm{op}} = U\Sigma V^\top$,

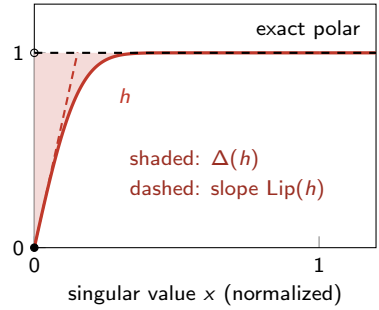
$$X_{t+1} = -D\, U\, \mathrm{diag}\big(h(\sigma_1), \ldots, h(\sigma_r)\big)\, V^\top.$$

Assumption. $h\colon [0,\infty) \to [0,1]$ is continuous and nondecreasing, with

(i) $h(0) = 0$

(ii) Lipschitz constant $\mathrm{Lip}(h) < \infty$

(iii) $\Delta(h) := \int_0^\infty \big(1 - h(x)\big)\,\mathrm{d}x < \infty$



Finite Newton–Schulz: $h_q = f^{\circ q}$ on $[0,1]$ and $h_q = 1$ on $[1,\infty)$
Exact polar: $h = \mathbf{1}\{x > 0\}$, with $\Delta = 0$ but $\mathrm{Lip} = \infty$.

Discounted regret for general spectral maps

Theorem.

$$\mathbb{E}\Big[\mathrm{Reg}_t^{[\beta]}(D)\Big] \le D\Big(\underbrace{\frac{rG_{\mathrm{op}}\,\Delta(h)}{1-\beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2\,\mathrm{Lip}(h)}{G_{\mathrm{op}}}}_{\text{stability}}\Big).$$

$\mathrm{Reg}_t^{[\beta]}(D) = \max_{\|U\|_{\mathrm{op}}\le D} \sum_{s\le t} \beta^{t-s}\langle G_s, X_s - U\rangle$

Discounted regret for general spectral maps

Theorem.

$$\mathbb{E}\Big[\mathrm{Reg}_t^{[\beta]}(D)\Big] \le D\Big(\underbrace{\frac{rG_{\mathrm{op}}\,\Delta(h)}{1-\beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2\,\mathrm{Lip}(h)}{G_{\mathrm{op}}}}_{\text{stability}}\Big).$$

$\mathrm{Reg}_t^{[\beta]}(D) = \max_{\|U\|_{\mathrm{op}}\le D} \sum_{s\le t} \beta^{t-s}\langle G_s, X_s - U\rangle$

The same penalty–stability tradeoff as in the analysis of follow-the-regularized-leader (FTRL) and online mirror descent (Shalev-Shwartz, 2012; McMahan, 2017).

Discounted regret for general spectral maps

Theorem.

$$\mathbb{E}\Big[\mathrm{Reg}_t^{[\beta]}(D)\Big] \le D\Big(\underbrace{\frac{rG_{\mathrm{op}}\,\Delta(h)}{1-\beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2\,\mathrm{Lip}(h)}{G_{\mathrm{op}}}}_{\text{stability}}\Big).$$

$\mathrm{Reg}_t^{[\beta]}(D) = \max_{\|U\|_{\mathrm{op}}\le D} \sum_{s\le t} \beta^{t-s}\langle G_s, X_s - U\rangle$

The same penalty–stability tradeoff as in the analysis of follow-the-regularized-leader (FTRL) and online mirror descent (Shalev-Shwartz, 2012; McMahan, 2017).

Finite Newton–Schulz: a larger depth q decreases $\Delta(h_q)$ but increases $\mathrm{Lip}(h_q)$
$\rightarrow$ the depth q balances the penalty and stability terms
Muon with finite Newton–Schulz is itself FTRL

Finite Newton–Schulz as a spectral map

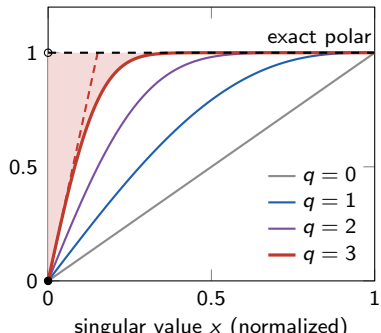
Recall $f(x) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$ and $h_q = f^{\circ q}$ on $[0,1]$, $h_q = 1$ on $[1,\infty)$: continuous, nondecreasing, $h_q(0) = 0$.

Lemma. With $A := f'(0) = 15/8$,

$$\mathrm{Lip}(h_q) = A^q, \qquad \frac{1}{2A^q} \le \Delta(h_q) \le \frac{1}{A^q + 1}.$$

$\Delta(h) = \int_0^\infty (1 - h(x))\,\mathrm{d}x$

Increasing q shrinks the gap and raises the slope, both geometrically. The uniform error does not shrink: $\sup_{x>0}\big(1 - h_q(x)\big) = 1$ for every q.



shaded: $\Delta(h_3)$ dashed: slope A^3 of h_3 at 0

Discounted regret of Muon with finite Newton–Schulz

By the general theorem and the lemma ($\Delta(h_q) \le A^{-q}$, $\mathrm{Lip}(h_q) = A^q$),

$$\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(D)] \le D\Big(\underbrace{\frac{rG_{\mathrm{op}}\,\Delta(h_q)}{1-\beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2\,\mathrm{Lip}(h_q)}{G_{\mathrm{op}}}}_{\text{stability}}\Big) \le D\Big(\frac{rG_{\mathrm{op}}}{(1-\beta)\,A^q} + \frac{2\,A^q\,\Gamma^2}{G_{\mathrm{op}}}\Big).$$

A larger q decreases the penalty and increases the stability term, which blows up as $q \to \infty$, reflecting the jump of the polar map at 0.

Discounted regret of Muon with finite Newton–Schulz

By the general theorem and the lemma ($\Delta(h_q) \le A^{-q}$, $\mathrm{Lip}(h_q) = A^q$),

$$\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(D)] \le D\Big(\underbrace{\frac{rG_{\mathrm{op}}\,\Delta(h_q)}{1-\beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2\,\mathrm{Lip}(h_q)}{G_{\mathrm{op}}}}_{\text{stability}}\Big) \le D\Big(\frac{rG_{\mathrm{op}}}{(1-\beta)\,A^q} + \frac{2\,A^q\,\Gamma^2}{G_{\mathrm{op}}}\Big).$$

A larger q decreases the penalty and increases the stability term, which blows up as $q \to \infty$, reflecting the jump of the polar map at 0. Balancing the two terms:

Theorem.

$$q = \left\lceil \log_A \frac{G_{\mathrm{op}}\sqrt{r}}{\Gamma\sqrt{1-\beta}} \right\rceil \quad \text{gives} \quad \mathbb{E}[\mathrm{Reg}_t^{[\beta]}(D)] \le (1+2A)\,D\,\Gamma\sqrt{\frac{r}{1-\beta}}.$$

Dividing by D: $\boxed{\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}}$ for all t,
as required by the conversion guarantee.

Main result: nonsmooth nonconvex objectives

Theorem. Let $B := \sqrt{r}\,((1+2A)\Gamma+\sigma) \asymp \sqrt{r}(\Gamma+\sigma)$ and $\varepsilon \le B$. Run discounted O2NC with Muon (finite Newton–Schulz) as the learner, with

$$1-\beta = \Big(\frac{\varepsilon}{3B}\Big)^2, \qquad D = \frac{(1-\beta)\rho}{4\beta}, \qquad q = \Big\lceil \log_A \frac{3BG_{\mathrm{op}}\sqrt{r}}{\Gamma\varepsilon} \Big\rceil.$$

Then $\bar{W}_\tau$ is a (ρ,ε)-stationary point in expectation, using

$$T = O\Big(\frac{\Delta_{\mathcal{L}}\, r(\Gamma+\sigma)^2}{\rho\,\varepsilon^3} + \frac{r(\Gamma+\sigma)^2}{\varepsilon^2}\Big) \quad \text{stochastic gradients.}$$

$\bar{W}_\tau$: the output of discounted O2NC, a weighted mean of the query points

Main result: nonsmooth nonconvex objectives

Theorem. Let $B := \sqrt{r}\,((1+2A)\Gamma+\sigma) \asymp \sqrt{r}(\Gamma+\sigma)$ and $\varepsilon \le B$. Run discounted O2NC with Muon (finite Newton–Schulz) as the learner, with

$$1-\beta = \Big(\frac{\varepsilon}{3B}\Big)^2, \qquad D = \frac{(1-\beta)\rho}{4\beta}, \qquad q = \Big\lceil \log_A \frac{3BG_{\mathrm{op}}\sqrt{r}}{\Gamma\varepsilon} \Big\rceil.$$

Then $\bar{W}_\tau$ is a (ρ,ε)-stationary point in expectation, using

$$T = O\Big(\frac{\Delta_{\mathcal{L}}\, r(\Gamma+\sigma)^2}{\rho\,\varepsilon^3} + \frac{r(\Gamma+\sigma)^2}{\varepsilon^2}\Big) \quad \text{stochastic gradients.}$$

$\bar{W}_\tau$: the output of discounted O2NC, a weighted mean of the query points

Proof: discounted regret $O(\Gamma\sqrt{r/(1-\beta)})$ + conversion guarantee, with $1-\beta \asymp (\varepsilon/B)^2$.

Comparison: nonsmooth nonconvex objectives

Number of stochastic gradients to find a (ρ, ε)-stationary point (NS: Newton–Schulz):

Method	Smoothing	Polar	Complexity
Exact-polar Muon	none	exact	not covered by discounted O2NC
Pion (Jiang et al., 2026)	Gaussian perturbation	exact	$O\Big(\max\Big\{\frac{\nu^2\|Q\|_*^2\Delta_{\mathcal{L}}}{\rho\varepsilon^3}, \frac{\nu^2\|Q\|_*^2}{\varepsilon^2}, \frac{\nu r G}{\varepsilon}\Big\}\Big)$
Leon (Jiang et al., 2026)	hyperbolic smoothing	exact	$O\Big(\max\Big\{\frac{\|Q\|_*^2\Delta_{\mathcal{L}}}{\rho\varepsilon^3}, \frac{\|Q\|_*^2}{\varepsilon^2}, \frac{rG}{\varepsilon}\Big\}\Big)$
This work	**finite NS**	$q = O(\log(1/\varepsilon))$ NS steps	$O\Big(\max\Big\{\frac{r(\Gamma+\sigma)^2\Delta_{\mathcal{L}}}{\rho\varepsilon^3}, \frac{r(\Gamma+\sigma)^2}{\varepsilon^2}\Big\}\Big)$

$\nu = (n/(n-r-1))^{1/4}$

Smooth nonconvex objectives

Assumption. In addition, $\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(W')\|_* \le L_{\mathrm{op}}\|W - W'\|_{\mathrm{op}}$ for all W, W'.

Theorem. Let $\varepsilon \le 12rG_{\mathrm{op}}$. Run discounted O2NC with Muon as the learner, with $q = \lceil \log_A(12rG_{\mathrm{op}}/\varepsilon)\rceil$ and suitable β and D. Then $\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_\tau)\|_* \le \varepsilon$, using

$$T = O\Big(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2} + \frac{r\sigma^2}{\varepsilon^2} + \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\Big) \quad \text{stochastic gradients.}$$

Optimal in ε: ε^{-2} without noise (Carmon et al., 2020), ε^{-4} with noise (Arjevani et al., 2023).

Comparison: smooth nonconvex objectives

Reference	Deterministic	Stochastic	Polar
Kovalev (2025)	$O\Big(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\Big)$	$O\Big(\max\Big\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{\sqrt{r}\sigma}{\varepsilon}, \frac{r^{3/2}\sigma^3}{\varepsilon^3}, \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\Big\}\Big)$	exact
Shen et al. (2026)	$O\Big(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\Big)$	$O\Big(\max\Big\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{r^2\sigma^4}{L_{\mathrm{op}}\Delta_{\mathcal{L}}\varepsilon^2}, \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\Big\}\Big)$	exact
Shulgin et al. (2026)	$O\Big(\frac{(1+\delta)^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{(1-\delta)^2\varepsilon^2}\Big)$	$O\Big(\frac{(1+\delta)r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{(1-\delta)^4\varepsilon^4}\Big)$	inexact LMO
Kim and Oh (2026)	$O\Big(\frac{\chi_q^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\Big)$	$O\Big(\max\Big\{\frac{\chi_q^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{\chi_q^2 r^2\sigma^2}{\varepsilon^2}, \frac{\chi_q^4 r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\Big\}\Big)$	finite NS
This work	$O\Big(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\Big)$	$O\Big(\max\Big\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{r\sigma^2}{\varepsilon^2}, \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\Big\}\Big)$	finite NS

Leading terms coincide with those for exact-polar Muon.
$\chi_q = (1-\varepsilon_q)^{-1}$, where ε_q is the uniform error of q Newton–Schulz steps.

Muon with finite Newton–Schulz is FTRL

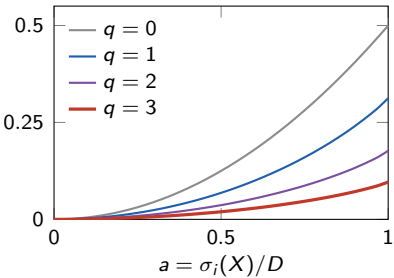
Theorem. Let $\phi_h(x) \coloneqq \int_0^x h(u)\,\mathrm{d}u$, let ϕ_h^* be its Fenchel conjugate, and let $R_h(X) \coloneqq D\, G_{\mathrm{op}} \sum_i \phi_h^*\big(\sigma_i(X)/D\big)$. Then

$$X_{t+1} \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \left\{ \sum_{s=1}^{t} \beta^{t-s} \langle G_s, X \rangle + \frac{R_h(X)}{1-\beta} \right\}.$$

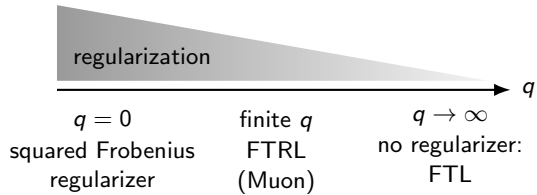
Muon with finite Newton–Schulz is FTRL

Theorem. Let $\phi_h(x) \coloneqq \int_0^x h(u)\,\mathrm{d}u$, let ϕ_h^* be its Fenchel conjugate, and let $R_h(X) \coloneqq D\, G_{\mathrm{op}} \sum_i \phi_h^*\big(\sigma_i(X)/D\big)$. Then

$$X_{t+1} \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \left\{ \sum_{s=1}^{t} \beta^{t-s} \langle G_s, X \rangle + \frac{R_h(X)}{1-\beta} \right\}.$$



regularizer per singular value, $\phi^*_{h_q}(a)$



The depth controls the amount of regularization.

Another spectral map: SoftMuon

SoftMuon (Feoktistov et al., 2026), with the same momentum $M_t = U\,\mathrm{diag}(\sigma_i)V^\top$, sets

$$X_{t+1} = -D\,M_t\big(M_t^\top M_t + a^{-2}I\big)^{-1/2} = -D\,U\,\mathrm{diag}\big(\widetilde{h}_a(\sigma_i)\big)\,V^\top, \qquad \widetilde{h}_a(x) = \frac{ax}{\sqrt{1+a^2x^2}},$$

with a smoothing parameter $a > 0$.
The regularized Muon of Mustafi et al. (2026) uses the same map.

Another spectral map: SoftMuon

SoftMuon (Feoktistov et al., 2026), with the same momentum $M_t = U\,\mathrm{diag}(\sigma_i)V^\top$, sets

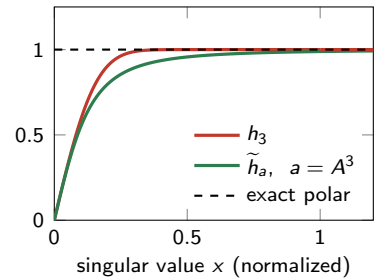
$$X_{t+1} = -D\,M_t\big(M_t^\top M_t + a^{-2}I\big)^{-1/2} = -D\,U\,\mathrm{diag}\big(\widetilde{h}_a(\sigma_i)\big)V^\top, \qquad \widetilde{h}_a(x) = \frac{ax}{\sqrt{1+a^2x^2}},$$

with a smoothing parameter $a > 0$.
The regularized Muon of Mustafi et al. (2026) uses the same map.

$\widetilde{h}_a$ satisfies our assumption, with

$$\mathrm{Lip}(\widetilde{h}_a) = a, \qquad \Delta(\widetilde{h}_a) = \frac{1}{a}.$$

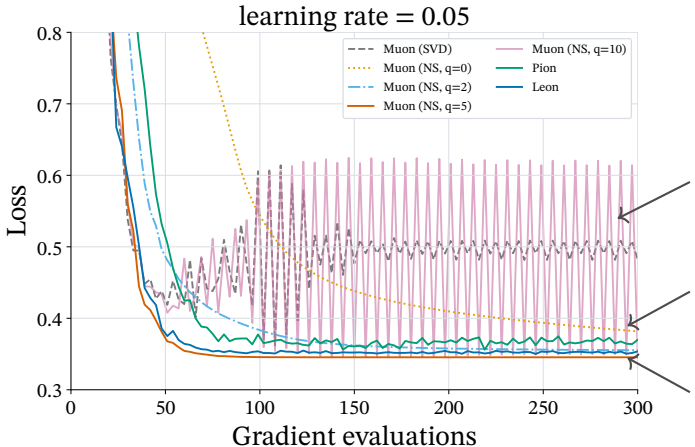
So the same guarantees hold, now controlled by a rather than the depth q.



same slope at 0 h_3 has the smaller gap

Numerical illustration

A nonsmooth nonconvex problem adapted from Jiang et al. (2026).



exact polar (SVD) and $q = 10$: oscillate

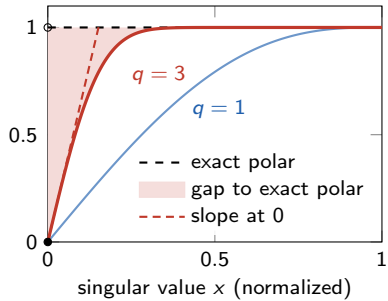
$q = 0$: slow

$q = 2$ and $q = 5$: fast and stable

Consistent with the penalty–stability tradeoff.

Summary

Finite Newton–Schulz is a smoothing mechanism, not an approximation error!



Thanks for listening!

References I

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 807–823, 2014.

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(1–2):165–214, 2023.

Åke Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020.

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

Ashok Cutkosky, 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*, volume 202, pages 6643–6670, 2023.

References II

Dmitrii Feoktistov, Timofey Belinsky, Andrey Veprikov, Amir Zainullin, and Aleksandr Beznosikov. Softsign: Smooth sign in your optimizer for better parameter heterogeneity handling. In *International Conference on Machine Learning*, 2026.

GLM-4.5 Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1): 14–22, 1977.

Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336, pages 3782–3824, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. 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*, 2026.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

References III

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

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

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.

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

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.

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

Aratrika Mustafi, Soumya Mukherjee, and Bharath K. Sriperumbudur. Move on Muon: A Hamiltonian probability gradient flow perspective of Muon optimizer. *arXiv preprint arXiv:2605.23871*, 2026.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 49069–49104, 2025.

References IV

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *International Conference on Artificial Intelligence and Statistics*, volume 300, pages 5167–5175, 2026.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, 2026.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11173–11182, 2020.

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*, volume 235, pages 58780–58799, 2024.