

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

arXiv:2608.26288

Mingyi Li¹ Taira Tsuchiya^{1,2}

¹The University of Tokyo ²RIKEN

Multi-agent seminar @ UTokyo · October 5, 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)

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)

Its practicality depends on cheap orthogonalization.

The Muon update

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

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

The Muon update

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

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

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

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

The map `polar` 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 :

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

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

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

The map `polar` 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

The **Newton–Schulz iteration** (Kovarik, 1970; Björck and Bowie, 1971) converges to $\text{polar}(Y)$ using only matrix products, after normalizing Y so that $\|Y\|_{\text{op}} \leq 1$:

$$Y \leftarrow \frac{15}{8}Y - \frac{5}{4}(YY^{\top})Y + \frac{3}{8}(YY^{\top})^2Y.$$

This is $p(YY^{\top})Y$: in $\text{polar}(Y) = (YY^{\top})^{-1/2}Y$, the factor $\lambda^{-1/2}$ is replaced by its degree-2 Taylor polynomial at $\lambda = 1$, $p(\lambda) = \frac{15}{8} - \frac{5}{4}\lambda + \frac{3}{8}\lambda^2$.

Computing the polar factor: Newton–Schulz

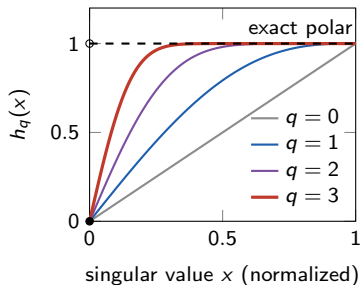
The **Newton–Schulz iteration** (Kovarik, 1970; Björck and Bowie, 1971) converges to $\text{polar}(Y)$ using only matrix products, after normalizing Y so that $\|Y\|_{\text{op}} \leq 1$:

$$Y \leftarrow \frac{15}{8}Y - \frac{5}{4}(YY^\top)Y + \frac{3}{8}(YY^\top)^2Y.$$

This is $p(YY^\top)Y$: in $\text{polar}(Y) = (YY^\top)^{-1/2}Y$, the factor $\lambda^{-1/2}$ is replaced by its degree-2 Taylor polynomial at $\lambda = 1$, $p(\lambda) = \frac{15}{8} - \frac{5}{4}\lambda + \frac{3}{8}\lambda^2$.

Action on the singular values: write $Y = U\Sigma V^\top$.

- ① $(YY^\top)^k Y = U\Sigma^{2k+1}V^\top$ for every k .
- ② One step $Y \mapsto p(YY^\top)Y$ keeps the singular vectors and maps each singular value x to $f(x) := x p(x^2) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$.
- ③ After q steps, $Y_q = U h_q(\Sigma) V^\top$ with $h_q := f^{\circ q}$.



Computing the polar factor: Newton–Schulz

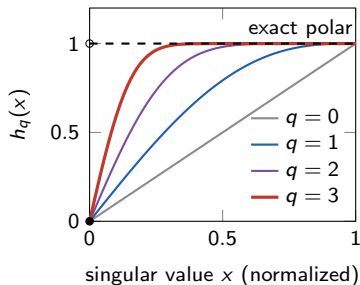
The **Newton–Schulz iteration** (Kovarik, 1970; Björck and Bowie, 1971) converges to $\text{polar}(Y)$ using only matrix products, after normalizing Y so that $\|Y\|_{\text{op}} \leq 1$:

$$Y \leftarrow \frac{15}{8}Y - \frac{5}{4}(YY^\top)Y + \frac{3}{8}(YY^\top)^2Y.$$

This is $p(YY^\top)Y$: in $\text{polar}(Y) = (YY^\top)^{-1/2}Y$, the factor $\lambda^{-1/2}$ is replaced by its degree-2 Taylor polynomial at $\lambda = 1$, $p(\lambda) = \frac{15}{8} - \frac{5}{4}\lambda + \frac{3}{8}\lambda^2$.

Action on the singular values: write $Y = U\Sigma V^\top$.

- ① $(YY^\top)^k Y = U\Sigma^{2k+1}V^\top$ for every k .
- ② One step $Y \mapsto p(YY^\top)Y$ keeps the singular vectors and maps each singular value x to $f(x) := x p(x^2) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$.
- ③ After q steps, $Y_q = U h_q(\Sigma) V^\top$ with $h_q := f^{\circ q}$.



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)
→ **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, which orthogonalizes the momentum M_t exactly, uses the update

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

since $-D \operatorname{polar}(S) \in \arg \min_{\|X\|_{\text{op}} \leq D} \langle S, X \rangle$ and $M_t = (1 - \beta) \sum_{s \leq t} \beta^{t-s} G_s$

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, which orthogonalizes the momentum M_t exactly, uses the update

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

since $-D \operatorname{polar}(S) \in \arg \min_{\|X\|_{\text{op}} \leq D} \langle S, X \rangle$ and $M_t = (1 - \beta) \sum_{s \leq t} \beta^{t-s} G_s$

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

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

Exact-polar Muon can fail to converge even on convex Lipschitz objectives (Parshakova et al., 2026).

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, which orthogonalizes the momentum M_t exactly, uses the update

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

since $-D \operatorname{polar}(S) \in \arg \min_{\|X\|_{\text{op}} \leq D} \langle S, X \rangle$ and $M_t = (1 - \beta) \sum_{s \leq t} \beta^{t-s} G_s$

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

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

Exact-polar Muon can fail to converge even on convex Lipschitz objectives (Parshakova et al., 2026).

Jiang et al. (2026) propose Pion and Leon, which replace the learner

→ **extra ingredients that Muon lacks; finite Newton–Schulz is not covered**

Existing analyses:

one line of work treats finite Newton–Schulz only as an **error** and needs **smoothness**;
the other handles nonsmooth objectives but **replaces the update Muon actually performs**.

Muon never computes the exact polar factor. It runs **finite** Newton–Schulz.

How does the finite Newton–Schulz iteration affect optimization?

Can it be a **benefit** rather than an error?

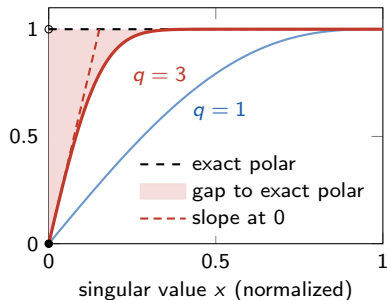
Can Muon find stationary points of nonsmooth nonconvex objectives?

Main result

Theorem (informal). Let $\mathcal{L}$ be Lipschitz, possibly nonsmooth and nonconvex, with unbiased, bounded stochastic gradients. Muon with momentum and $q = O(\log(1/\varepsilon))$ Newton–Schulz steps finds a (ρ, ε) -stationary point using $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ stochastic gradients.

Main result

Theorem (informal). Let $\mathcal{L}$ be Lipschitz, possibly nonsmooth and nonconvex, with unbiased, bounded stochastic gradients. Muon with momentum and $q = O(\log(1/\varepsilon))$ Newton–Schulz steps finds a (ρ, ε) -stationary point using $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ stochastic gradients.



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

- Exact polar ($q = \infty$) = FTL
- No iteration ($q = 0$)
= FTRL with the squared Frobenius norm
- $q = O(\log(1/\varepsilon))$ balances the two
- Also (nearly) optimal for smooth $\mathcal{L}$

1. Introduction
2. **Problem setup**
3. Online-to-nonconvex conversion
4. Discounted regret for general spectral maps
5. Muon with finite Newton–Schulz
6. Beyond Newton–Schulz
7. Conclusion

Problem and assumptions

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

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

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

Nondifferentiable points are handled by a random perturbation (next slide).

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

Points of nondifferentiability

The assumption needs differentiability, but ReLU networks have kinks. For locally Lipschitz $\mathcal{L}$, a **random perturbation** restores it (Cutkosky et al., 2023):

$$\hat{\mathcal{L}}_\delta(W) := \mathbb{E}_Z[\mathcal{L}(W + \delta Z)], \quad Z \text{ uniform on the unit Frobenius ball.}$$

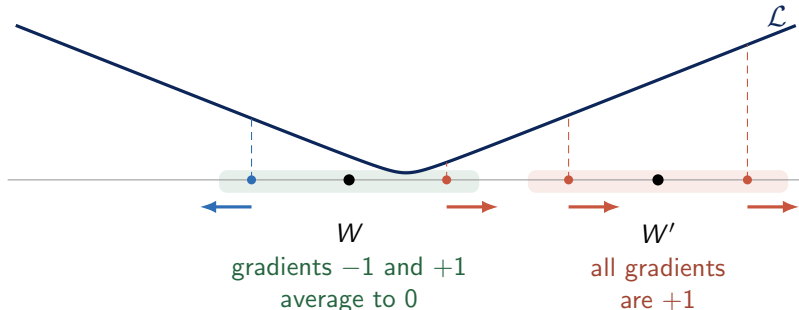
- $\hat{\mathcal{L}}_\delta$ satisfies the assumption: differentiable and Lipschitz (its value is never needed).
- Stochastic gradient: draw Z and use the output G of the original oracle at $W + \delta Z$,
 $\mathbb{E}[G \mid W] = \mathbb{E}_Z[\nabla \mathcal{L}(W + \delta Z)] = \nabla \hat{\mathcal{L}}_\delta(W)$. $\mathcal{L}$ is differentiable at $W + \delta Z$ with probability 1
- Stationarity of $\hat{\mathcal{L}}_\delta$ with radius $\rho \geq \delta$ ((ρ, ε) -stationarity, defined later) gives stationarity of $\mathcal{L}$ with radius 2ρ , as in Cutkosky et al. (2023).

Cutkosky et al. (2023): “In practice we suspect that such perturbation arguments are unnecessary: intuitively an algorithm is unlikely to query a point of non-differentiability.”

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 about ρ , has norm at most ε .

Definition (Jiang et al., 2026).

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

W is a (ρ, ε) -stationary point if

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

- $\mathcal{P}(W; \rho)$: mean W , average distance $\leq \rho$ from W (e.g., the two points on the previous slide).
- A Goldstein-type relaxation (Goldstein, 1977), unavoidable for general Lipschitz objectives (Zhang et al., 2020; Kornowski and Shamir, 2022).

Finite supports follow the definition of Jiang et al.; the restriction is inessential, since finitely many moment conditions can be matched by a finitely supported distribution.

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**: $\|\nabla\mathcal{L}(W)\|_* \leq \varepsilon$.

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.

Lemma. If $\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(W')\|_* \leq L_{\text{op}}\|W - W'\|_{\text{op}}$ ($\forall W, W'$), then

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

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

Proof: for any $p \in \mathcal{P}(W; \rho)$ and $Y \sim p$,

$$\begin{aligned} \|\nabla\mathcal{L}(W)\|_* &\leq \|\mathbb{E}\nabla\mathcal{L}(Y)\|_* + \|\mathbb{E}[\nabla\mathcal{L}(W) - \nabla\mathcal{L}(Y)]\|_* && \text{triangle inequality} \\ &\leq \|\mathbb{E}\nabla\mathcal{L}(Y)\|_* + \mathbb{E}\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(Y)\|_* && \text{Jensen's inequality} \\ &\leq \|\mathbb{E}\nabla\mathcal{L}(Y)\|_* + L_{\text{op}}\mathbb{E}\|W - Y\|_{\text{op}} \leq \|\mathbb{E}\nabla\mathcal{L}(Y)\|_* + L_{\text{op}}\rho, && \text{smoothness, } \mathbb{E}\|W - Y\|_{\text{op}} \leq \rho \end{aligned}$$

and the claim follows by taking the infimum over p on the right.

We use this lemma later for smooth objectives.

Muon in our analysis

Input: momentum $\beta \in (0, 1)$, radius $D > 0$, operator-norm bound $G_{\text{op}} > 0$, depth q

$M_0 = 0, X_1 = 0$

for $t = 1, 2, \dots$:

 receive G_t

$M_t \leftarrow \beta M_{t-1} + (1 - \beta) G_t$

momentum

$Y \leftarrow M_t / G_{\text{op}}$

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

$X_{t+1} \leftarrow -D Y$ ($W_{t+1} \leftarrow W_t - D Y$)

next update

Muon in our analysis

Input: momentum $\beta \in (0, 1)$, radius $D > 0$, operator-norm bound $G_{\text{op}} > 0$, depth q

$M_0 = 0, X_1 = 0$

for $t = 1, 2, \dots$:

 receive G_t

$M_t \leftarrow \beta M_{t-1} + (1 - \beta) G_t$

momentum

$Y \leftarrow M_t / G_{\text{op}}$

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

$X_{t+1} \leftarrow -D Y$ ($W_{t+1} \leftarrow W_t - D Y$)

next update

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

$$X_{t+1} = -D \mathcal{H}_{h_q}(M_t / G_{\text{op}}), \quad \mathcal{H}_h(U \Sigma V^\top) := U \text{diag}(h(\sigma_1), \dots, h(\sigma_r)) V^\top$$

($U \Sigma V^\top$: thin SVD) h_q : f composed q times ($= 1$ on $[1, \infty)$). Exact polar: $h(x) = \mathbf{1}\{x > 0\}$.

Outline

1. Introduction
2. Problem setup
3. **Online-to-nonconvex conversion**
4. Discounted regret for general spectral maps
5. Muon with finite Newton–Schulz
6. Beyond Newton–Schulz
7. Conclusion

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

receive the loss $X \mapsto \langle G_t, X \rangle$

suffer the loss $\langle G_t, X_t \rangle$

The learner outputs the update X_t , not a point W_t in the feasible set

(in online-to-batch conversion, the learner outputs a point in the feasible set)

The loss tracks the change of $\mathcal{L}$ through the first-order approximation

$\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) \approx \langle \nabla \mathcal{L}(W_{t-1}), X_t \rangle$ (exact in O2NC).

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

receive the loss $X \mapsto \langle G_t, X \rangle$

suffer the loss $\langle G_t, X_t \rangle$

The learner outputs the update X_t , not a point W_t in the feasible set

(in online-to-batch conversion, the learner outputs a point in the feasible set)

The loss tracks the change of $\mathcal{L}$ through the first-order approximation

$\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) \approx \langle \nabla \mathcal{L}(W_{t-1}), X_t \rangle$ (exact in O2NC).

Existing optimizers correspond to known online learners:

SGD with momentum to online gradient descent, Adam to discounted FTRL, and Muon (exact polar) to follow-the-leader.

Up to small modifications (Zhang and Cutkosky, 2024; Ahn et al., 2024; Jiang et al., 2026).

Online-to-nonconvex conversion (O2NC)

O2NC (Cutkosky et al., 2023): a framework that lets an online learner choose the updates and converts a regret bound into a guarantee of stationarity.

for $t = 1, 2, \dots$:

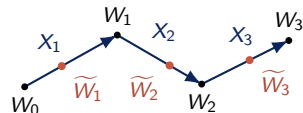
the learner chooses an update X_t , $\|X_t\|_{\text{op}} \leq D$

update $W_t \leftarrow W_{t-1} + X_t$

query a stochastic gradient G_t at $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$,

$u_t \sim \text{Unif}[0, 1]$

feed the linear loss $X \mapsto \langle G_t, X \rangle$ to the learner



query point $\widetilde{W}_t$:
uniformly random on the segment

Online-to-nonconvex conversion (O2NC)

O2NC (Cutkosky et al., 2023): a framework that lets an online learner choose the updates and converts a regret bound into a guarantee of stationarity.

for $t = 1, 2, \dots$:

the learner chooses an update X_t , $\|X_t\|_{\text{op}} \leq D$

update $W_t \leftarrow W_{t-1} + X_t$

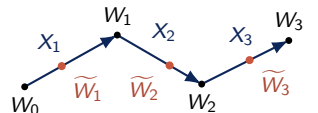
query a stochastic gradient G_t at $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$,

$u_t \sim \text{Unif}[0, 1]$

feed the linear loss $X \mapsto \langle G_t, X \rangle$ to the learner

Regret against a fixed U with $\|U\|_{\text{op}} \leq D$:

$$\text{Reg}_T(U) = \sum_{t=1}^T \langle G_t, X_t - U \rangle$$



query point $\widetilde{W}_t$:
uniformly random on the segment

Online-to-nonconvex conversion (O2NC)

O2NC (Cutkosky et al., 2023): a framework that lets an online learner choose the updates and converts a regret bound into a guarantee of stationarity.

for $t = 1, 2, \dots$:

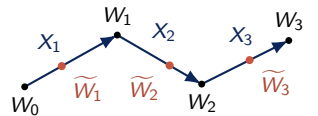
the learner chooses an update X_t , $\|X_t\|_{\text{op}} \leq D$

update $W_t \leftarrow W_{t-1} + X_t$

query a stochastic gradient G_t at $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$,

$u_t \sim \text{Unif}[0, 1]$

feed the linear loss $X \mapsto \langle G_t, X \rangle$ to the learner



query point $\widetilde{W}_t$:
uniformly random on the segment

Regret against a fixed U with $\|U\|_{\text{op}} \leq D$:

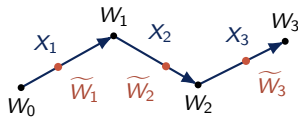
$$\text{Reg}_T(U) = \sum_{t=1}^T \langle G_t, X_t - U \rangle$$

To the learner, the nonsmooth nonconvexity of $\mathcal{L}$ is just an unpredictable gradient sequence $(G_t)_t$

In the smooth case, the predictability of the gradients can be exploited for better rates

An exact identity for one round

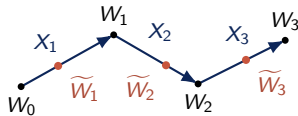
For the update $X_t = W_t - W_{t-1}$ chosen by the learner, the stochastic gradient G_t is queried at $\widetilde{W}_t = W_{t-1} + u_t X_t$, $u_t \sim \text{Unif}[0, 1]$.



$$\begin{aligned}
 \text{fundamental theorem of} & \quad \mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) = \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + u(W_t - W_{t-1})), W_t - W_{t-1} \rangle du \\
 \text{calculus along the segment} & \\
 \text{definition of } X_t & \quad = \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + uX_t), X_t \rangle du \\
 u_t \text{ is uniform} & \quad = \mathbb{E}_{u_t} \langle \nabla \mathcal{L}(\widetilde{W}_t), X_t \rangle \\
 G_t \text{ is unbiased} & \quad = \mathbb{E}[\langle G_t, X_t \rangle \mid W_{t-1}, X_t]
 \end{aligned}$$

An exact identity for one round

For the update $X_t = W_t - W_{t-1}$ chosen by the learner, the stochastic gradient G_t is queried at $\widetilde{W}_t = W_{t-1} + u_t X_t$, $u_t \sim \text{Unif}[0, 1]$.



$$\begin{aligned}
 \text{fundamental theorem of calculus along the segment} \quad \mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) &= \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + u(W_t - W_{t-1})), W_t - W_{t-1} \rangle du \\
 \text{definition of } X_t \quad &= \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + uX_t), X_t \rangle du \\
 u_t \text{ is uniform} \quad &= \mathbb{E}_{u_t} \langle \nabla \mathcal{L}(\widetilde{W}_t), X_t \rangle \\
 G_t \text{ is unbiased} \quad &= \mathbb{E}[\langle G_t, X_t \rangle \mid W_{t-1}, X_t]
 \end{aligned}$$

Without smoothness, the loss $\langle G_t, X_t \rangle$ is an unbiased estimate of the change of $\mathcal{L}$:

$$\mathbb{E}[\mathcal{L}(W_T)] - \mathcal{L}(W_0) = \sum_{t=1}^T \mathbb{E} \langle G_t, X_t \rangle$$

From a regret bound to a bound on the averaged gradient

Take $U := -D \operatorname{polar}(\sum_{t=1}^T G_t)$, the best update in hindsight ($\|U\|_{\text{op}} \leq D$). By definition,

$$\sum_{t=1}^T \langle G_t, X_t \rangle = \operatorname{Reg}_T(U) + \langle \sum_{t=1}^T G_t, U \rangle = \operatorname{Reg}_T(U) - D \left\| \sum_{t=1}^T G_t \right\|_*.$$

In expectation, the left side is $\mathbb{E}[\mathcal{L}(W_T)] - \mathcal{L}(W_0) \geq -\Delta_{\mathcal{L}}$, so

$$\mathbb{E} \left\| \frac{1}{T} \sum_{t=1}^T G_t \right\|_* \leq \frac{\mathbb{E}[\operatorname{Reg}_T(U)] + \Delta_{\mathcal{L}}}{DT}: \text{ the regret bounds the averaged gradient.}$$

With OGD, $\operatorname{Reg}_T(U) = O(\sqrt{T})$ for every U , so the right side vanishes as $T \rightarrow \infty$.

From a regret bound to a bound on the averaged gradient

Take $U := -D \operatorname{polar}(\sum_{t=1}^T G_t)$, the best update in hindsight ($\|U\|_{\text{op}} \leq D$). By definition,

$$\sum_{t=1}^T \langle G_t, X_t \rangle = \operatorname{Reg}_T(U) + \langle \sum_{t=1}^T G_t, U \rangle = \operatorname{Reg}_T(U) - D \left\| \sum_{t=1}^T G_t \right\|_*.$$

In expectation, the left side is $\mathbb{E}[\mathcal{L}(W_T)] - \mathcal{L}(W_0) \geq -\Delta_{\mathcal{L}}$, so

$$\mathbb{E} \left\| \frac{1}{T} \sum_{t=1}^T G_t \right\|_* \leq \frac{\mathbb{E}[\operatorname{Reg}_T(U)] + \Delta_{\mathcal{L}}}{DT}: \text{ the regret bounds the averaged gradient.}$$

With OGD, $\operatorname{Reg}_T(U) = O(\sqrt{T})$ for every U , so the right side vanishes as $T \rightarrow \infty$.

The query points $\widetilde{W}_t$ spread over the whole trajectory (up to DT apart).

For stationarity at a point, average the gradients of **recent rounds** only:

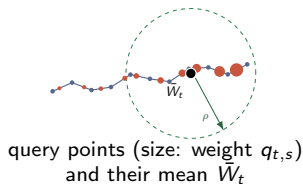
- over blocks (Cutkosky et al., 2023) (appendix);
- with the momentum weights β^{t-s} **starting on the next slide**

Small momentum gives a stationary point

Weights $q_{t,s} := (1 - \beta)\beta^{t-s}/(1 - \beta^t)$: they sum to 1 and **favor recent rounds** (largest at $s = t$, decaying by β per round).

$$\frac{M_t}{1 - \beta^t} = \sum_{s \leq t} q_{t,s} G_s \quad \text{average of the recent gradients}$$

$$\bar{W}_t := \sum_{s \leq t} q_{t,s} \tilde{W}_s \quad \text{average of the recent query points}$$



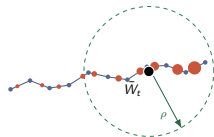
Small momentum gives a stationary point

Weights $q_{t,s} := (1 - \beta)\beta^{t-s}/(1 - \beta^t)$: they sum to 1 and **favor recent rounds** (largest at $s = t$, decaying by β per round).

$$\frac{M_t}{1 - \beta^t} = \sum_{s \leq t} q_{t,s} G_s \quad \text{average of the recent gradients}$$

$$\bar{W}_t := \sum_{s \leq t} q_{t,s} \widetilde{W}_s \quad \text{average of the recent query points}$$

Distribution for stationarity: let p pick $\widetilde{W}_s$ with probability $q_{t,s}$; then p has mean $\bar{W}_t$ and $\mathbb{E}_{Y \sim p} \|Y - \bar{W}_t\|_{\text{op}} \leq 2D \mathbb{E}_{s,s' \sim q_{t,\cdot}} |s - s'| \leq 4D\beta/(1 - \beta)$ (consecutive query points are within $2D$). With $D = (1 - \beta)\rho/(4\beta)$, this is $\leq \rho$, i.e., $p \in \mathcal{P}(\bar{W}_t; \rho)$.



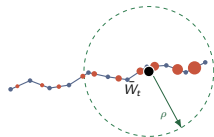
query points (size: weight $q_{t,s}$)
and their mean $\bar{W}_t$

Small momentum gives a stationary point

Weights $q_{t,s} := (1 - \beta)\beta^{t-s}/(1 - \beta^t)$: they sum to 1 and **favor recent rounds** (largest at $s = t$, decaying by β per round).

$$\frac{M_t}{1 - \beta^t} = \sum_{s \leq t} q_{t,s} G_s \quad \text{average of the recent gradients}$$

$$\bar{W}_t := \sum_{s \leq t} q_{t,s} \bar{W}_s \quad \text{average of the recent query points}$$



query points (size: weight $q_{t,s}$)
and their mean $\bar{W}_t$

Distribution for stationarity: let p pick $\bar{W}_s$ with probability $q_{t,s}$; then p has mean $\bar{W}_t$ and $\mathbb{E}_{Y \sim p} \|Y - \bar{W}_t\|_{\text{op}} \leq 2D \mathbb{E}_{s,s' \sim q_{t,\cdot}} |s - s'| \leq 4D\beta/(1 - \beta)$ (consecutive query points are within $2D$). With $D = (1 - \beta)\rho/(4\beta)$, this is $\leq \rho$, i.e., $p \in \mathcal{P}(\bar{W}_t; \rho)$.

$$\text{Without noise } (G_s = \nabla \mathcal{L}(\bar{W}_s)): \quad \|\nabla \mathcal{L}(\bar{W}_t)\|_*^{[\rho]} \leq \left\| \sum_{s \leq t} q_{t,s} \nabla \mathcal{L}(\bar{W}_s) \right\|_* = \frac{\|M_t\|_*}{1 - \beta^t}$$

With noise $E_s := G_s - \nabla \mathcal{L}(\bar{W}_s)$, the right side has the extra term $(1 - \beta) \left\| \sum_{s \leq t} \beta^{t-s} E_s \right\|_* / (1 - \beta^t)$ (bounded later). It remains to show that $\|M_t\|_*$ is small for most t .

Discounted regret and the momentum: an identity

Goal: bound $\|M_t\|_*$ by the regret of the learner.

Definition (Zhang and Cutkosky, 2024; Jiang et al., 2026). **Discounted regret** at round t :

$$\text{Reg}_t^{[\beta]}(D) := \max_{\|U\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s - U \rangle. \quad \text{regret over the last } \approx 1/(1 - \beta) \text{ rounds}$$

Discounted regret and the momentum: an identity

Goal: bound $\|M_t\|_*$ by the regret of the learner.

Definition (Zhang and Cutkosky, 2024; Jiang et al., 2026). **Discounted regret** at round t :

$$\text{Reg}_t^{[\beta]}(D) := \max_{\|U\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s - U \rangle. \quad \text{regret over the last } \approx 1/(1-\beta) \text{ rounds}$$

$$\begin{aligned} \text{Reg}_t^{[\beta]}(D) &= \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s \rangle + \max_{\|U\|_{\text{op}} \leq D} \langle \sum_{s=1}^t \beta^{t-s} G_s, -U \rangle \\ &= \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s \rangle + \frac{D}{1-\beta} \|M_t\|_* \quad \text{duality and } M_t = (1-\beta) \sum_{s \leq t} \beta^{t-s} G_s \end{aligned}$$

For any learner,
$$\frac{D}{1-\beta} \|M_t\|_* = \text{Reg}_t^{[\beta]}(D) + \underbrace{\left(- \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s \rangle \right)}_{\text{in expectation, the recent decrease of } \mathcal{L}}$$

The momentum is large only if $\mathcal{L}$ decreased recently or the discounted regret is large. $\mathcal{L}$ can decrease by at most $\Delta_{\mathcal{L}}$, so **small regret leaves few rounds with large momentum.**

Summing over the rounds

Goal: show that $\|M_t\|_*$ is small for most t , in terms of the regret and $\Delta_{\mathcal{L}}$.

① Multiply the identity at round t by $a_t := 1 - \beta$ ($t < T$), $a_T := 1$, and sum:

$$\frac{D}{1 - \beta} \sum_{t=1}^T a_t \|M_t\|_* = \sum_{t=1}^T a_t \text{Reg}_t^{[\beta]}(D) - \sum_{s=1}^T \langle G_s, X_s \rangle.$$

Round s enters the identities at $t \geq s$ with weight β^{t-s} ; in total $\sum_{t \geq s} a_t \beta^{t-s} = (1 - \beta^{T-s}) + \beta^{T-s} = 1$.

② The last term has expectation $\mathcal{L}(W_0) - \mathbb{E}[\mathcal{L}(W_T)] \leq \Delta_{\mathcal{L}}$,
so with small regret, $\sum_t a_t \|M_t\|_*$ is small.

Summing over the rounds

Goal: show that $\|M_t\|_*$ is small for most t , in terms of the regret and $\Delta_{\mathcal{L}}$.

① Multiply the identity at round t by $a_t := 1 - \beta$ ($t < T$), $a_T := 1$, and sum:

$$\frac{D}{1-\beta} \sum_{t=1}^T a_t \|M_t\|_* = \sum_{t=1}^T a_t \text{Reg}_t^{[\beta]}(D) - \sum_{s=1}^T \langle G_s, X_s \rangle.$$

Round s enters the identities at $t \geq s$ with weight β^{t-s} ; in total $\sum_{t \geq s} a_t \beta^{t-s} = (1 - \beta^{T-s}) + \beta^{T-s} = 1$.

② The last term has expectation $\mathcal{L}(W_0) - \mathbb{E}[\mathcal{L}(W_T)] \leq \Delta_{\mathcal{L}}$,

so with small regret, $\sum_t a_t \|M_t\|_*$ is small.

③ Output: draw τ with $\mathbb{P}(\tau = t) = \frac{a_t(1-\beta^t)}{(1-\beta)T}$ and return $\bar{W}_\tau$ (**discounted O2NC**).

The weights a_t match ②, and $1 - \beta^t$ cancels the denominator of the stationarity bound.

$$\mathbb{E} \|\nabla \mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]} \leq \frac{1}{DT} \left(\sum_{t=1}^T a_t \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] + \Delta_{\mathcal{L}} \right) + \text{noise term}$$

Conversion guarantee

Lemma (Jiang et al., 2026). With $D = (1 - \beta)\rho/(4\beta)$, discounted O2NC satisfies

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

$\text{Reg}_t^{[\beta]}(1)$: discounted regret of X_s/D against $\|U\|_{\text{op}} \leq 1$; $\text{Reg}_t^{[\beta]}(D) = D \cdot \text{Reg}_t^{[\beta]}(1)$.

Proof: the previous slides, plus a martingale bound on the noise (appendix).

Conversion guarantee

Lemma (Jiang et al., 2026). With $D = (1 - \beta)\rho/(4\beta)$, discounted O2NC satisfies

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

$\text{Reg}_t^{[\beta]}(1)$: discounted regret of X_s/D against $\|U\|_{\text{op}} \leq 1$; $\text{Reg}_t^{[\beta]}(D) = D \cdot \text{Reg}_t^{[\beta]}(1)$.

Proof: the previous slides, plus a martingale bound on the noise (appendix).

- Only the discounted regrets depend on the learner.
- If $\mathbb{E}[\text{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}$ 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 satisfies

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

$\text{Reg}_t^{[\beta]}(1)$: discounted regret of X_s/D against $\|U\|_{\text{op}} \leq 1$; $\text{Reg}_t^{[\beta]}(D) = D \cdot \text{Reg}_t^{[\beta]}(1)$.

Proof: the previous slides, plus a martingale bound on the noise (appendix).

- Only the discounted regrets depend on the learner.
- If $\mathbb{E}[\text{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}$ 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.

Outline

1. Introduction
2. Problem setup
3. Online-to-nonconvex conversion
4. **Discounted regret for general spectral maps**
5. Muon with finite Newton–Schulz
6. Beyond Newton–Schulz
7. Conclusion

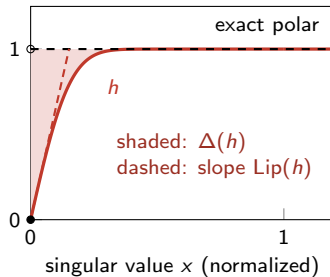
Muon with a general spectral map

Replace h_q by a **spectral map** h ; fix a normalization scale $C > 0$ (Muon: $C = G_{\text{op}}$):

$$X_{t+1} = -D \mathcal{H}_h(M_t/C). \quad \mathcal{H}_h(U\Sigma V^\top) = U \text{diag}(h(\sigma_i)) V^\top$$

Assumption. $h: [0, \infty) \rightarrow [0, 1]$ is continuous and nondecreasing, and

- (i) $h(0) = 0$;
- (ii) $\text{Lip}(h) := \sup_{x \neq y} |h(x) - h(y)| / |x - y| < \infty$;
- (iii) $\Delta(h) := \int_0^\infty (1 - h(x)) \, dx < \infty$.



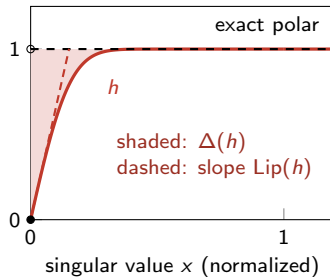
Muon with a general spectral map

Replace h_q by a **spectral map** h ; fix a normalization scale $C > 0$ (Muon: $C = G_{\text{op}}$):

$$X_{t+1} = -D \mathcal{H}_h(M_t/C). \quad \mathcal{H}_h(U\Sigma V^\top) = U \text{diag}(h(\sigma_i)) V^\top$$

Assumption. $h: [0, \infty) \rightarrow [0, 1]$ is continuous and nondecreasing, and

- (i) $h(0) = 0$;
- (ii) $\text{Lip}(h) := \sup_{x \neq y} |h(x) - h(y)| / |x - y| < \infty$;
- (iii) $\Delta(h) := \int_0^\infty (1 - h(x)) \, dx < \infty$.



$\Delta(h)$: the approximation error of h relative to the exact polar map.

Exact polar: $\Delta = 0$, but $\text{Lip} = \infty$.

Finite Newton–Schulz h_q ($h_q = 1$ on $[1, \infty)$): both finite.

Discounted regret for general spectral maps

Theorem.

$$\mathbb{E} \left[\text{Reg}_t^{[\beta]}(D) \right] \leq D \left(\underbrace{\frac{rC \Delta(h)}{1 - \beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h)}{C}}_{\text{stability}} \right).$$

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

Discounted regret for general spectral maps

Theorem.

$$\mathbb{E} \left[\text{Reg}_t^{[\beta]}(D) \right] \leq D \left(\underbrace{\frac{rC \Delta(h)}{1 - \beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h)}{C}}_{\text{stability}} \right).$$

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

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

Discounted regret for general spectral maps

Theorem.

$$\mathbb{E} \left[\text{Reg}_t^{[\beta]}(D) \right] \leq D \left(\underbrace{\frac{rC \Delta(h)}{1 - \beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h)}{C}}_{\text{stability}} \right).$$

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

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

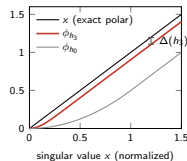
Finite Newton–Schulz: a larger depth q decreases $\Delta(h_q)$ but increases $\text{Lip}(h_q)$;
the depth q balances the penalty and stability terms.

Muon with finite Newton–Schulz is itself FTRL (shown later).

Proof idea: the update is the gradient of a potential

Smoothed potential

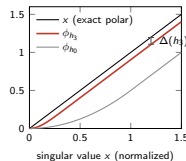
$$\tilde{\Phi}(M) := C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) := \int_0^x h(u) \, du$$



Proof idea: the update is the gradient of a potential

Smoothed potential

$$\tilde{\Phi}(M) := C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) := \int_0^x h(u) du$$



Lemma. $\tilde{\Phi}$ is convex and continuously differentiable with $\nabla \tilde{\Phi} = \mathcal{H}_h(\cdot/C)$.

Hence the update is $X_{t+1} = -D \mathcal{H}_h(M_t/C) = -D \nabla \tilde{\Phi}(M_t)$.

Proof. For any convex differentiable ϕ with $\phi'(0) = 0$, the function $F(S) := \sum_i \phi(\sigma_i(S))$ is convex and differentiable, with

$$\nabla F(U\Sigma V^\top) = U \text{diag}(\phi'(\sigma_i(S))) V^\top.$$

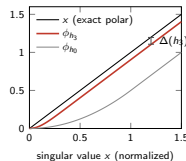
Apply this with $\phi = \phi_h$.

Such a learner is called a gradient-based prediction algorithm (Abernethy et al., 2014).

Proof idea: the update is the gradient of a potential

Smoothed potential

$$\tilde{\Phi}(M) := C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) := \int_0^x h(u) du$$



Lemma. $\tilde{\Phi}$ is convex and continuously differentiable with $\nabla \tilde{\Phi} = \mathcal{H}_h(\cdot/C)$.

Hence the update is $X_{t+1} = -D \mathcal{H}_h(M_t/C) = -D \nabla \tilde{\Phi}(M_t)$.

Proof. For any convex differentiable ϕ with $\phi'(0) = 0$, the function $F(S) := \sum_i \phi(\sigma_i(S))$ is convex and differentiable, with

$$\nabla F(U\Sigma V^\top) = U \operatorname{diag}(\phi'(\sigma_i(S))) V^\top.$$

Apply this with $\phi = \phi_h$.

Such a learner is called a gradient-based prediction algorithm (Abernethy et al., 2014).

Exact-polar Muon: since $\operatorname{polar}(M) \in \partial \|M\|_*$, the corresponding potential is $M \mapsto \|M\|_*$
 (for $h = \mathbf{1}\{x > 0\}$: $\phi_h(x) = x$ and $\tilde{\Phi} = \|\cdot\|_*$)

Proof: a momentum-based regret decomposition

Lemma. With $B_{\tilde{\Phi}}$ the Bregman divergence of $\tilde{\Phi}$,

$$\text{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta} \left(\underbrace{\|M_t\|_* - \tilde{\Phi}(M_t)}_{\text{penalty}} + \underbrace{\sum_{s=1}^t \beta^{t-s} B_{\tilde{\Phi}}(M_s \| M_{s-1})}_{\text{stability}} + \underbrace{\sum_{s=1}^{t-1} \beta^{t-s-1} V_s}_{\leq 0} \right),$$

where $V_s := (1 - \beta)(\tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle)$.

Proof: a momentum-based regret decomposition

Lemma. With $B_{\tilde{\Phi}}$ the Bregman divergence of $\tilde{\Phi}$,

$$\text{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta} \left(\underbrace{\|M_t\|_* - \tilde{\Phi}(M_t)}_{\text{penalty}} + \underbrace{\sum_{s=1}^t \beta^{t-s} B_{\tilde{\Phi}}(M_s \| M_{s-1})}_{\text{stability}} + \underbrace{\sum_{s=1}^{t-1} \beta^{t-s-1} V_s}_{\leq 0} \right),$$

where $V_s := (1-\beta)(\tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle)$.

Idea:

1. By duality, $\text{Reg}_t^{[\beta]}(D) = \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s \rangle + \frac{D}{1-\beta} \|M_t\|_*$ (the momentum identity).
2. Write $(1-\beta)G_s = (M_s - M_{s-1}) + (1-\beta)M_{s-1}$: Bregman divergences and differences $\tilde{\Phi}(M_s) - \tilde{\Phi}(M_{s-1})$ appear; rearranging the discounted sum of the latter leaves V_s .

Proof: a momentum-based regret decomposition

Lemma. With $B_{\tilde{\Phi}}$ the Bregman divergence of $\tilde{\Phi}$,

$$\text{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta} \left(\underbrace{\|M_t\|_* - \tilde{\Phi}(M_t)}_{\text{penalty}} + \underbrace{\sum_{s=1}^t \beta^{t-s} B_{\tilde{\Phi}}(M_s \| M_{s-1})}_{\text{stability}} + \underbrace{\sum_{s=1}^{t-1} \beta^{t-s-1} V_s}_{\leq 0} \right),$$

where $V_s := (1-\beta)(\tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle)$.

Idea:

1. By duality, $\text{Reg}_t^{[\beta]}(D) = \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s \rangle + \frac{D}{1-\beta} \|M_t\|_*$ (the momentum identity).
2. Write $(1-\beta)G_s = (M_s - M_{s-1}) + (1-\beta)M_{s-1}$: Bregman divergences and differences $\tilde{\Phi}(M_s) - \tilde{\Phi}(M_{s-1})$ appear; rearranging the discounted sum of the latter leaves V_s .

Unlike the decomposition via cumulative gradients (Abernethy et al., 2014; Jiang et al., 2026), this one is in terms of the **momentum**; it yields the optimal ε^{-2} dependence for deterministic smooth objectives.

Proof: the penalty term

$$\tilde{\Phi}(M) = C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) = \int_0^x h(u) \, du$$

Since $\|M\|_* = \sum_{i=1}^r \sigma_i(M)$,

$$\begin{aligned} \|M\|_* - \tilde{\Phi}(M) &= \sum_{i=1}^r \left(\sigma_i(M) - C \phi_h(\sigma_i(M)/C) \right) = C \sum_{i=1}^r \int_0^{\sigma_i(M)/C} (1 - h(u)) \, du \\ &\leq C \sum_{i=1}^r \int_0^\infty (1 - h(u)) \, du = rC \Delta(h). \end{aligned}$$

This difference is nonnegative since $h \leq 1$.

Proof: the penalty term

$$\tilde{\Phi}(M) = C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) = \int_0^x h(u) \, du$$

$$\text{Since } \|M\|_* = \sum_{i=1}^r \sigma_i(M),$$

$$\begin{aligned} \|M\|_* - \tilde{\Phi}(M) &= \sum_{i=1}^r \left(\sigma_i(M) - C \phi_h(\sigma_i(M)/C) \right) = C \sum_{i=1}^r \int_0^{\sigma_i(M)/C} (1 - h(u)) \, du \\ &\leq C \sum_{i=1}^r \int_0^\infty (1 - h(u)) \, du = rC \Delta(h). \end{aligned}$$

This difference is nonnegative since $h \leq 1$.

$$\text{Penalty term: } \frac{D}{1-\beta} (\|M_t\|_* - \tilde{\Phi}(M_t)) \leq \frac{D rC \Delta(h)}{1-\beta}.$$

Proof: the stability term

Fact. If h is Lipschitz and $h(0) = 0$, then $\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_F \leq \text{Lip}(h) \|X - Y\|_F$.
Hence $\nabla \tilde{\Phi} = \mathcal{H}_h(\cdot/C)$ is $(\text{Lip}(h)/C)$ -Lipschitz in $\|\cdot\|_F$.

Stability term.

if $\nabla \tilde{\Phi}$ is L -Lipschitz, then
 $B_{\tilde{\Phi}}(X \| Y) \leq \frac{L}{2} \|X - Y\|_F^2$

$$\mathbb{E} \|M_s - M_{s-1}\|_F^2 \leq 4(1 - \beta)^2 \Gamma^2$$

$$\begin{aligned} \mathbb{E} B_{\tilde{\Phi}}(M_s \| M_{s-1}) &\leq \frac{\text{Lip}(h)}{2C} \mathbb{E} \|M_s - M_{s-1}\|_F^2 \\ &\leq \frac{2 \text{Lip}(h)(1 - \beta)^2 \Gamma^2}{C} \end{aligned}$$

Proof: the stability term

Fact. If h is Lipschitz and $h(0) = 0$, then $\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_F \leq \text{Lip}(h) \|X - Y\|_F$.
Hence $\nabla\tilde{\Phi} = \mathcal{H}_h(\cdot/C)$ is $(\text{Lip}(h)/C)$ -Lipschitz in $\|\cdot\|_F$.

Stability term. if $\nabla\tilde{\Phi}$ is L -Lipschitz, then $B_{\tilde{\Phi}}(X\|Y) \leq \frac{L}{2} \|X - Y\|_F^2$

$$\begin{aligned} \mathbb{E}B_{\tilde{\Phi}}(M_s\|M_{s-1}) &\leq \frac{\text{Lip}(h)}{2C} \mathbb{E}\|M_s - M_{s-1}\|_F^2 \\ &\leq \frac{2\text{Lip}(h)(1-\beta)^2\Gamma^2}{C} \end{aligned}$$

Summing with the weights β^{t-s} , where $(1-\beta)\sum_{s=1}^t \beta^{t-s} \leq 1$,

Stability term:
$$\frac{D}{1-\beta} \sum_{s=1}^t \beta^{t-s} \mathbb{E}B_{\tilde{\Phi}}(M_s\|M_{s-1}) \leq D \frac{2\Gamma^2 \text{Lip}(h)}{C}.$$

Proof: the stability term

Fact. If h is Lipschitz and $h(0) = 0$, then $\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_F \leq \text{Lip}(h) \|X - Y\|_F$.
Hence $\nabla \tilde{\Phi} = \mathcal{H}_h(\cdot/C)$ is $(\text{Lip}(h)/C)$ -Lipschitz in $\|\cdot\|_F$.

Stability term. if $\nabla \tilde{\Phi}$ is L -Lipschitz, then $B_{\tilde{\Phi}}(X\|Y) \leq \frac{L}{2} \|X - Y\|_F^2$

$$\begin{aligned} \mathbb{E} B_{\tilde{\Phi}}(M_s \| M_{s-1}) &\leq \frac{\text{Lip}(h)}{2C} \mathbb{E} \|M_s - M_{s-1}\|_F^2 \\ &\leq \frac{2 \text{Lip}(h)(1 - \beta)^2 \Gamma^2}{C} \end{aligned}$$

Summing with the weights β^{t-s} , where $(1 - \beta) \sum_{s=1}^t \beta^{t-s} \leq 1$,

$$\text{Stability term: } \frac{D}{1 - \beta} \sum_{s=1}^t \beta^{t-s} \mathbb{E} B_{\tilde{\Phi}}(M_s \| M_{s-1}) \leq D \frac{2\Gamma^2 \text{Lip}(h)}{C}.$$

Last term. $\tilde{\Phi}$ is convex with $\tilde{\Phi}(0) = 0$, so

$$0 = \tilde{\Phi}(0) \geq \tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle = \frac{V_s}{1 - \beta}.$$

Outline

1. Introduction
2. Problem setup
3. Online-to-nonconvex conversion
4. Discounted regret for general spectral maps
5. **Muon with finite Newton–Schulz**
6. Beyond Newton–Schulz
7. Conclusion

Finite Newton–Schulz as a spectral map

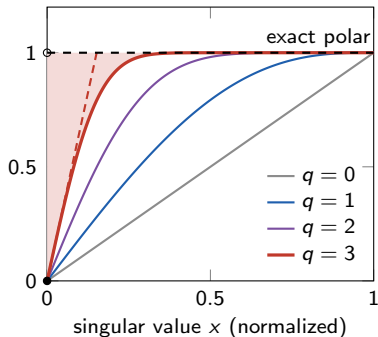
With $f(x) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$, set $h_q = f^{\circ q}$ on $[0, 1]$ and $h_q = 1$ on $[1, \infty)$: continuous, nondecreasing, $h_q(0) = 0$.

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

$$\text{Lip}(h_q) = A^q, \quad \frac{1}{2A^q} \leq \Delta(h_q) \leq \frac{1}{A^q + 1}.$$

$$\Delta(h) = \int_0^\infty (1 - h(x)) \, dx$$

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

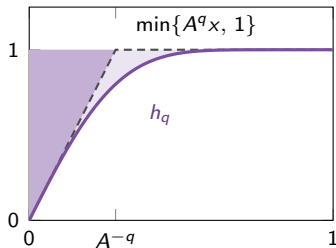


shaded: $\Delta(h_3)$; dashed: slope A^3 of h_3 at 0

Proof of the lemma

Lower bound. Since $\text{Lip}(h_q) = A^q$, we have $h_q(x) \leq \min\{A^q x, 1\}$ for every $x \geq 0$, so the gap contains a triangle:

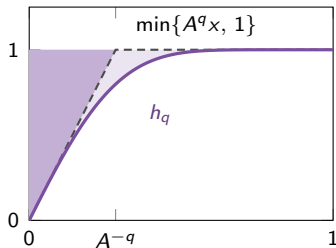
$$\Delta(h_q) \geq \int_0^{A^{-q}} (1 - A^q x) dx = \frac{1}{2A^q}.$$



Proof of the lemma

Lower bound. Since $\text{Lip}(h_q) = A^q$, we have $h_q(x) \leq \min\{A^q x, 1\}$ for every $x \geq 0$, so the gap contains a triangle:

$$\Delta(h_q) \geq \int_0^{A^{-q}} (1 - A^q x) \, dx = \frac{1}{2A^q}.$$



Upper bound (idea). With $g_L(x) := 1 - (1 - x)^L$: $f \geq g_A$ and $g_L \circ g_{L'} = g_{LL'}$, so by the monotonicity of f , $h_q = f^{\circ q} \geq g_A^{\circ q} = g_{A^q}$ on $[0, 1]$, and

$$\Delta(h_q) \leq \int_0^1 (1 - g_{A^q}(x)) \, dx = \int_0^1 (1 - x)^{A^q} \, dx = \frac{1}{A^q + 1}.$$

Discounted regret of Muon with finite Newton–Schulz

By the general theorem with $C = G_{\text{op}}$ and the lemma ($\Delta(h_q) \leq A^{-q}$, $\text{Lip}(h_q) = A^q$),

$$\begin{aligned} \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] &\leq D \left(\underbrace{\frac{rG_{\text{op}} \Delta(h_q)}{1 - \beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h_q)}{G_{\text{op}}}}_{\text{stability}} \right) \\ &\leq D \left(\frac{rG_{\text{op}}}{(1 - \beta) A^q} + \frac{2 A^q \Gamma^2}{G_{\text{op}}} \right). \end{aligned}$$

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

Discounted regret of Muon with finite Newton–Schulz

By the general theorem with $C = G_{\text{op}}$ and the lemma ($\Delta(h_q) \leq A^{-q}$, $\text{Lip}(h_q) = A^q$),

$$\begin{aligned} \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] &\leq D \left(\underbrace{\frac{rG_{\text{op}} \Delta(h_q)}{1-\beta}}_{\text{penalty}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h_q)}{G_{\text{op}}}}_{\text{stability}} \right) \\ &\leq D \left(\frac{rG_{\text{op}}}{(1-\beta) A^q} + \frac{2 A^q \Gamma^2}{G_{\text{op}}} \right). \end{aligned}$$

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

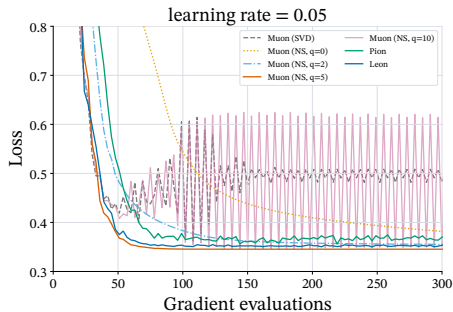
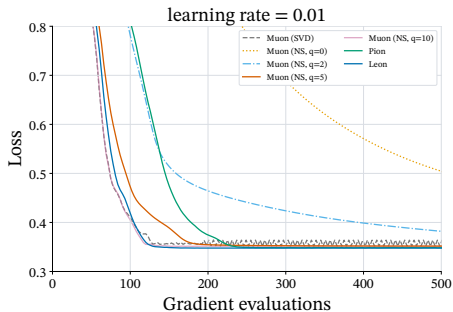
Theorem.

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

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

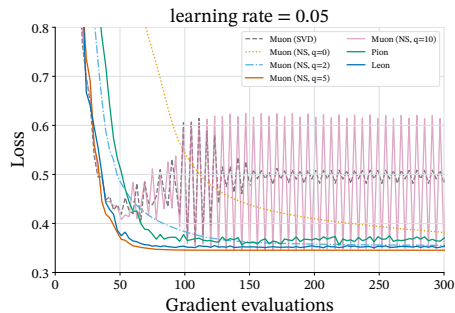
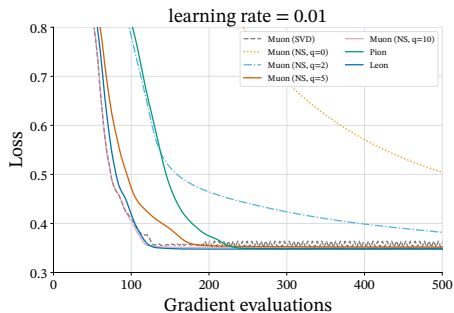
Numerical illustration

A nonsmooth nonconvex problem (a variant of Jiang et al., 2026), $W \in \mathbb{R}^{20 \times 20}$, $\beta = 0.9$; learning rate = D .



Numerical illustration

A nonsmooth nonconvex problem (a variant of Jiang et al., 2026), $W \in \mathbb{R}^{20 \times 20}$, $\beta = 0.9$; learning rate = D .



- Learning rate 0.01 (left): small depths ($q = 0, 2$) are slow.
- Learning rate 0.05 (right): exact-polar Muon (SVD) and $q = 10$ oscillate; $q = 0$ is slow; $q = 2, 5$ are fast and stable.

Consistent with the penalty–stability tradeoff. (Gradients at the iterates; not a benchmark.)

Main result: nonsmooth nonconvex objectives

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

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

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

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

$\bar{W}_\tau$: 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 $\varepsilon \leq B$. Run discounted O2NC with Muon (finite Newton–Schulz) as the learner, with

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

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

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

$\bar{W}_\tau$: 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

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	$O\left(\max\left\{\frac{\nu^2 \ Q\ _*^2 \Delta_{\mathcal{L}}}{\rho \varepsilon^3}, \frac{\nu^2 \ Q\ _*^2}{\varepsilon^2}, \frac{\nu r G}{\varepsilon}\right\}\right)$
Leon (Jiang et al., 2026)	hyperbolic smoothing	exact	$O\left(\max\left\{\frac{\ Q\ _*^2 \Delta_{\mathcal{L}}}{\rho \varepsilon^3}, \frac{\ Q\ _*^2}{\varepsilon^2}, \frac{r G}{\varepsilon}\right\}\right)$
This work	finite NS	$q = O(\log(1/\varepsilon))$ NS steps	$O\left(\max\left\{\frac{r(\Gamma + \sigma)^2 \Delta_{\mathcal{L}}}{\rho \varepsilon^3}, \frac{r(\Gamma + \sigma)^2}{\varepsilon^2}\right\}\right)$

- Same leading term $\rho^{-1}\varepsilon^{-3}$. Pion and Leon assume $\mathbb{E}[G_t G_t^\top] \preceq Q^2$ and $\|G_t\|_{\text{op}} \leq G$ a.s.; $\|Q\|_*^2$ and $r(\Gamma + \sigma)^2$ come from different assumptions and are not directly comparable.
- Pion's bound assumes an exact expectation of polar factors; its Monte Carlo version uses T samples per round to keep this bound.
- $\nu = (n/(n - r - 1))^{1/4}$; the bounds for Pion and Leon hold for sufficiently small ε .

Smooth nonconvex objectives

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

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

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

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

Smooth nonconvex objectives

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

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

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

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

Proof idea:

- (ρ, ε) to ε : $\|\nabla\mathcal{L}(W)\|_* \leq \|\nabla\mathcal{L}(W)\|_*^{[\rho]} + L_{\text{op}}\rho$. With $\rho = \varepsilon/L_{\text{op}}$, the nonsmooth bound alone gives only ε^{-4} , even without noise.
- Stability: split $M_s - M_{s-1}$ into signal and noise. The signal part is bounded independently of $\text{Lip}(h_q)$, and it is small since the momentum tracks the gradient (smoothness); A^q multiplies only σ^2 .

Comparison: smooth nonconvex objectives

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

- Leading terms coincide with those for exact-polar Muon. Versus [Kim and Oh \(2026\)](#): no factor χ_q , and $r^2\sigma^2/\varepsilon^2$ improves to $r\sigma^2/\varepsilon^2$ (slightly different algorithms).
- $\delta < 1$: uniform error of the inexact linear minimization oracle (LMO); $\chi_q = (1 - \varepsilon_q)^{-1}$, ε_q : uniform error of q NS steps.
- This work additionally assumes $\|G_t\|_{\text{op}} \leq G_{\text{op}}$ a.s. and a Frobenius second-moment bound.

Muon with finite Newton–Schulz is FTRL

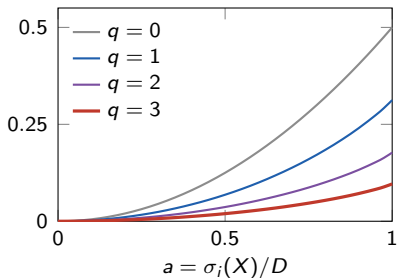
Theorem. With the regularizer $R_{h,C}(X) := DC \sum_i \phi_h^*(\sigma_i(X)/D)$, where ϕ_h^* is the Fenchel conjugate of ϕ_h ($\phi_h(x) = \int_0^x h(u) \, du$),

$$X_{t+1} \in \arg \min_{\|X\|_{\text{op}} \leq D} \left\{ \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle + \frac{R_{h,C}(X)}{1-\beta} \right\}.$$

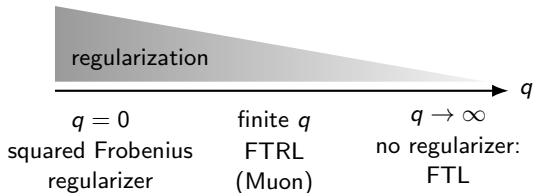
Muon with finite Newton–Schulz is FTRL

Theorem. With the regularizer $R_{h,C}(X) := DC \sum_i \phi_h^*(\sigma_i(X)/D)$, where ϕ_h^* is the Fenchel conjugate of ϕ_h ($\phi_h(x) = \int_0^x h(u) du$),

$$X_{t+1} \in \arg \min_{\|X\|_{\text{op}} \leq D} \left\{ \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle + \frac{R_{h,C}(X)}{1-\beta} \right\}.$$



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



The depth controls the amount of regularization.

Outline

1. Introduction
2. Problem setup
3. Online-to-nonconvex conversion
4. Discounted regret for general spectral maps
5. Muon with finite Newton–Schulz
6. **Beyond Newton–Schulz**
7. Conclusion

The gap–slope tradeoff is unavoidable

Every h with $0 \leq h \leq 1$ and $h(0) = 0$ has $h(x) \leq \min\{\text{Lip}(h)x, 1\}$, so the gap contains a triangle:

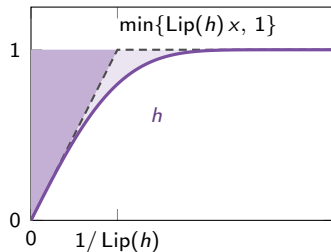
$$\Delta(h) \geq \frac{1}{2 \text{Lip}(h)}.$$

A smaller gap forces a larger slope.

$$\text{gap } \Delta(h) = \int_0^\infty (1 - h(x)) \, dx; \quad \text{slope } \text{Lip}(h)$$

Finite Newton–Schulz has $\Delta(h_q) \leq 1/\text{Lip}(h_q)$:
optimal up to a factor of 2.

The same holds for Newton–Schulz from a Taylor polynomial of any degree $\kappa \geq 1$ (Muon: $\kappa = 2$), with A replaced by the slope $A_\kappa = f'_\kappa(0)$ at 0 of the degree- κ polynomial f_κ .



light: the gap $\Delta(h)$; dark: the triangle

Another spectral map: SoftMuon

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

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

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 $M_t = U \text{diag}(\sigma_i) V^\top$, sets

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

with a smoothing parameter $a > 0$.

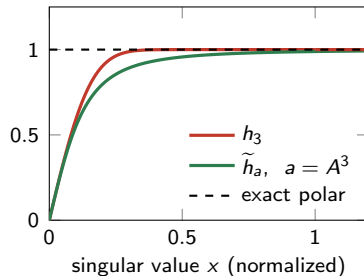
The regularized Muon of Mustafi et al. (2026) uses the same map.

Our assumption holds, with

$$\text{Lip}(\tilde{h}_a) = a, \quad \Delta(\tilde{h}_a) = \frac{1}{a}.$$

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

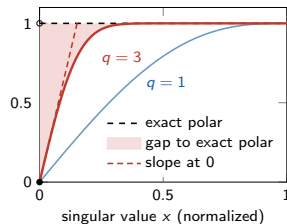
SoftMuon adjusts a during training and uses gradients at the iterates; our guarantee is for fixed a inside discounted O2NC.



same slope at 0; h_3 has the smaller gap

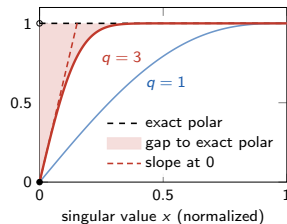
Finite Newton–Schulz is a smoothing, not an error:
it turns Muon from follow-the-leader into follow-the-regularized-leader.

1. **Smoothing.** A jump at 0 becomes a finite slope; exact-polar Muon ($q = \infty$) is FTL.
2. **Tradeoff.** Discounted regret $\leq$ penalty $\propto \Delta(h)$ + stability $\propto \text{Lip}(h)$; the depth q balances the two.
3. **Guarantees.** $q = O(\log(1/\varepsilon))$ inside O2NC: (ρ, ε) -stationary points with $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ stochastic gradients; optimal ε -dependence for smooth objectives.



Finite Newton–Schulz is a smoothing, not an error:
it turns Muon from follow-the-leader into follow-the-regularized-leader.

1. **Smoothing.** A jump at 0 becomes a finite slope; exact-polar Muon ($q = \infty$) is FTL.
2. **Tradeoff.** Discounted regret $\leq$ penalty $\propto \Delta(h)$ + stability $\propto \text{Lip}(h)$; the depth q balances the two.
3. **Guarantees.** $q = O(\log(1/\varepsilon))$ inside O2NC: (ρ, ε) -stationary points with $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ stochastic gradients; optimal ε -dependence for smooth objectives.



Differences from deployed Muon (future work)

- normalization: a constant G_{op} vs. a data-dependent one (e.g., $\|M_t\|_F$);
- coefficients: Taylor vs. an empirically tuned quintic;
- random query points and an averaged output (O2NC) vs. gradients at the iterates.

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.
- Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 619–640, 2024.
- 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.

References II

- 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.
- 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/>.

References III

- 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.
- 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.

References IV

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

References V

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.