Convergence of Muon with Newton–Schulz

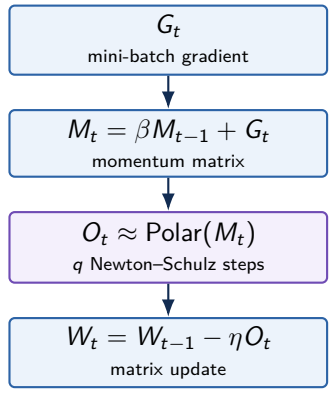
Gyu Yeol Kim **Min-hwan Oh**

Seoul National University | ICLR 2026

What is Muon? (Jordan et al., 2024)

Core mechanism

- Muon is a matrix-aware optimizer for hidden-layer weight matrices.
- It updates weights with an orthogonalized momentum direction instead of raw momentum.

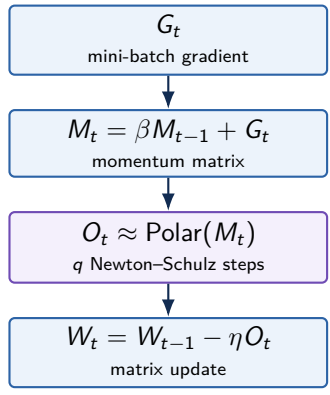


What is Muon? (Jordan et al., 2024)

Core mechanism

- Muon is a matrix-aware optimizer for hidden-layer weight matrices.
- It updates weights with an orthogonalized momentum direction instead of raw momentum.
- If $M_t = U\Sigma V^\top$, the ideal direction is the polar factor

$$\mathrm{Polar}(M_t) = UV^\top.$$



cheap
SVD-free
orthogonalization

What is Muon? (Jordan et al., 2024)

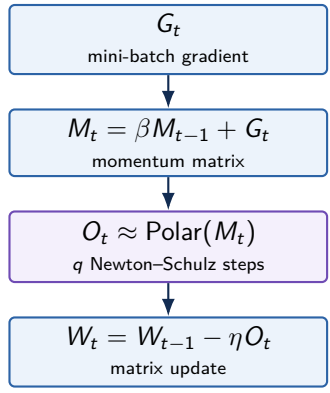
Core mechanism

- Muon is a matrix-aware optimizer for hidden-layer weight matrices.
- It updates weights with an orthogonalized momentum direction instead of raw momentum.
- If $M_t = U\Sigma V^\top$, the ideal direction is the polar factor

$$\mathrm{Polar}(M_t) = UV^\top.$$

Why Muon is attractive

- Respects the native matrix geometry of layers
- Empirically strong in large-scale training
- Practical implementation is SVD-free



Muon computes O_t from the scaled matrix

$$X_{t,0} = \frac{M_t}{\max\{1, \|M_t\|_F\}} \to X_{t,1} \to \cdots \to X_{t,q} := O_t$$

using only matrix multiplications and additions.

The practice–theory gap

What *actual* Muon does

- Finite-step Newton–Schulz orthogonalization
- Cheap, GPU-friendly, and used in real training
- Actual update is $O_t = X_{t,q}$, not exact SVD-polar

Open question

Does the **actual SVD-free Muon** converge in stochastic nonconvex optimization?

The practice–theory gap

What *actual* Muon does

- Finite-step Newton–Schulz orthogonalization
- Cheap, GPU-friendly, and used in real training
- Actual update is $O_t = X_{t,q}$, not exact SVD-polar

What most prior theory studies

- Replace Newton–Schulz by the exact polar factor $\mathrm{Polar}(M_t)$
- Cleaner analysis, but not the optimizer used in practice

Li & Hong, 2025; Shen et al., 2025; Sato et al., 2025

Open question

Does the **actual SVD-free Muon** converge in stochastic nonconvex optimization?

The practice–theory gap

What *actual* Muon does

- Finite-step Newton–Schulz orthogonalization
- Cheap, GPU-friendly, and used in real training
- Actual update is $O_t = X_{t,q}$, not exact SVD-polar

What most prior theory studies

- Replace Newton–Schulz by the exact polar factor $\mathrm{Polar}(M_t)$
- Cleaner analysis, but not the optimizer used in practice

Li & Hong, 2025; Shen et al., 2025; Sato et al., 2025

Open question

Does the **actual SVD-free Muon** converge in stochastic nonconvex optimization?

Problem Setting & Metric

$$\min_{W \in \mathbb{R}^{m \times n}} f(W) = \mathbb{E}_\xi[f(W;\xi)]$$

under standard smoothness and bounded-variance assumptions, with stationarity measured by

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(W_{t-1})\|_*\right].$$

Key technical bridge: from Newton–Schulz residual to polar error

Invariant target and subspace

Newton–Schulz preserves the target polar direction and the relevant column space:

$$\mathrm{Polar}(X_{t,j}) = \mathrm{Polar}(M_t), \qquad \mathrm{range}(X_{t,j}) = \mathrm{range}(M_t).$$

So the analysis reduces to controlling how close $X_{t,j}X_{t,j}^\top$ is to being a projector.

Key technical bridge: from Newton–Schulz residual to polar error

Invariant target and subspace

Newton–Schulz preserves the target polar direction and the relevant column space:

$$\mathrm{Polar}(X_{t,j}) = \mathrm{Polar}(M_t), \qquad \mathrm{range}(X_{t,j}) = \mathrm{range}(M_t).$$

So the analysis reduces to controlling how close $X_{t,j}X_{t,j}^\top$ is to being a projector.

Residual contraction

Define the orthogonality residual

$$\delta_{t,j} = \left\| \Pi_t - X_{t,j}X_{t,j}^\top \right\|_{\mathrm{op}}.$$

Then the multi-step decay is

$$\delta_{t,q} \le \delta_{t,0}^{(\kappa+1)^q}.$$

Key technical bridge: from Newton–Schulz residual to polar error

Invariant target and subspace

Newton–Schulz preserves the target polar direction and the relevant column space:

$$\mathrm{Polar}(X_{t,j}) = \mathrm{Polar}(M_t), \qquad \mathrm{range}(X_{t,j}) = \mathrm{range}(M_t).$$

So the analysis reduces to controlling how close $X_{t,j}X_{t,j}^\top$ is to being a projector.

Residual contraction

Define the orthogonality residual

$$\delta_{t,j} = \left\| \Pi_t - X_{t,j}X_{t,j}^\top \right\|_{\mathrm{op}}.$$

Then the multi-step decay is

$$\delta_{t,q} \le \delta_{t,0}^{(\kappa+1)^q}.$$

Bridge to the exact polar factor

Let

$$\varepsilon_q = \left\| X_{t,q} - \mathrm{Polar}(M_t) \right\|_{\mathrm{op}}.$$

The residual bound implies

$$\varepsilon_q \le 1 - \sqrt{1 - \delta_0^{(\kappa+1)^q}}.$$

Interpretation

After only a few Newton–Schulz steps, the practical update direction is already **very close** to the exact SVD-polar direction.

The approximation improves doubly exponentially in the number of steps q.

Main theorem: Muon matches the idealized rate up to a fast-vanishing factor

Convergence guarantee

With suitable η and β,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(W_{t-1})\|_*\right] \le \chi_q\left[\mathcal{O}\left(\left(\frac{r\sigma^2 LD}{BT}\right)^{1/4}\right)+\mathcal{O}\left(\sqrt{\frac{LD}{T}}+\frac{\sigma r}{\sqrt{BT}}\right)\right].$$

$$\chi_q=\frac{1}{1-\varepsilon_q}\le\frac{1}{\sqrt{1-\delta_0^{(\kappa+1)^q}}}\xrightarrow[q\uparrow]{}1.$$

$r=\min\{m,n\}$ and $D=f(W_0)-f^\star$.

Main theorem: Muon matches the idealized rate up to a fast-vanishing factor

Convergence guarantee

With suitable η and β,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(W_{t-1})\|_*\right] \le \chi_q\left[\mathcal{O}\left(\left(\frac{r\sigma^2 LD}{BT}\right)^{1/4}\right)+\mathcal{O}\left(\sqrt{\frac{LD}{T}}+\frac{\sigma r}{\sqrt{BT}}\right)\right].$$

$$\chi_q=\frac{1}{1-\varepsilon_q}\le\frac{1}{\sqrt{1-\delta_0^{(\kappa+1)^q}}}\xrightarrow[q\uparrow]{}1.$$

$r=\min\{m,n\}$ and $D=f(W_0)-f^\star$.

What the theorem says

- **First nonconvex convergence guarantee** for Muon with finite Newton–Schulz steps
- Same iteration-rate form as exact-polar Muon, up to χ_q
- χ_q approaches 1 **doubly exponentially fast** in q

Main theorem: Muon matches the idealized rate up to a fast-vanishing factor

Convergence guarantee

With suitable η and β,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(W_{t-1})\|_*\right] \le \chi_q\left[\mathcal{O}\left(\left(\frac{r\sigma^2 LD}{BT}\right)^{1/4}\right)+\mathcal{O}\left(\sqrt{\frac{LD}{T}}+\frac{\sigma r}{\sqrt{BT}}\right)\right].$$

$$\chi_q=\frac{1}{1-\varepsilon_q}\le\frac{1}{\sqrt{1-\delta_0^{(\kappa+1)^q}}}\xrightarrow[q\uparrow]{}1.$$

$r=\min\{m,n\}$ and $D=f(W_0)-f^\star$.

What the theorem says

- **First nonconvex convergence guarantee** for Muon with finite Newton–Schulz steps
- Same iteration-rate form as exact-polar Muon, up to χ_q
- χ_q approaches 1 **doubly exponentially fast** in q

Main message

Finite-step Newton–Schulz changes only a **rapidly vanishing constant factor**, not the essential convergence rate.

Why can Muon improve over SGD with momentum?

Compare the leading term under the same nuclear-norm stationarity metric

Method	Leading term	Rank dependence
SGD with momentum	$\mathcal{O}\left(\left(\frac{r^2\sigma^2LD}{BT}\right)^{1/4}\right)$	$r^{1/2}$
Muon	$\mathcal{O}\left(\left(\frac{r\sigma^2LD}{BT}\right)^{1/4}\right)$	$r^{1/4}$

Why can Muon improve over SGD with momentum?

Compare the leading term under the same nuclear-norm stationarity metric

Method	Leading term	Rank dependence
SGD with momentum	$\mathcal{O}\left(\left(\frac{r^2\sigma^2 LD}{BT}\right)^{1/4}\right)$	$r^{1/2}$
Muon	$\mathcal{O}\left(\left(\frac{r\sigma^2 LD}{BT}\right)^{1/4}\right)$	$r^{1/4}$

Geometric intuition

Orthogonalizing the momentum aligns the update with the spectral–nuclear geometry of matrix parameters instead of treating the parameter as a long vector.

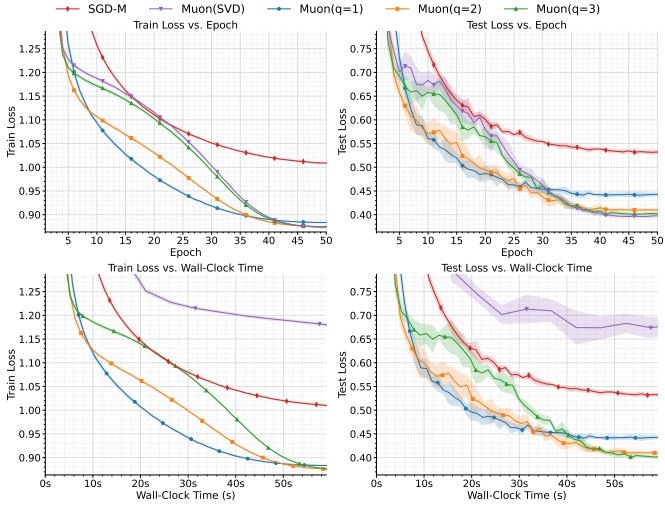
Implication

Muon sharpens the dependence on rank in the dominant $T^{-1/4}$ term, and Muon inherits this benefit up to the factor χ_q.

Experiments: a few Newton–Schulz steps are enough

Observed in the paper

- $q = 1$ already improves over SGD-M

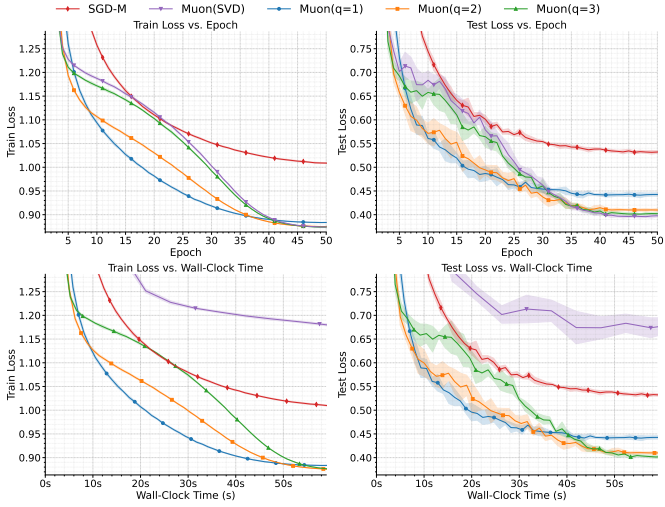


CIFAR-10 / CifarNet, $\kappa = 2$, averaged over 5 runs.

Experiments: a few Newton–Schulz steps are enough

Observed in the paper

- $q = 1$ already improves over SGD-M
- $q = 2, 3$ nearly match exact SVD by epoch

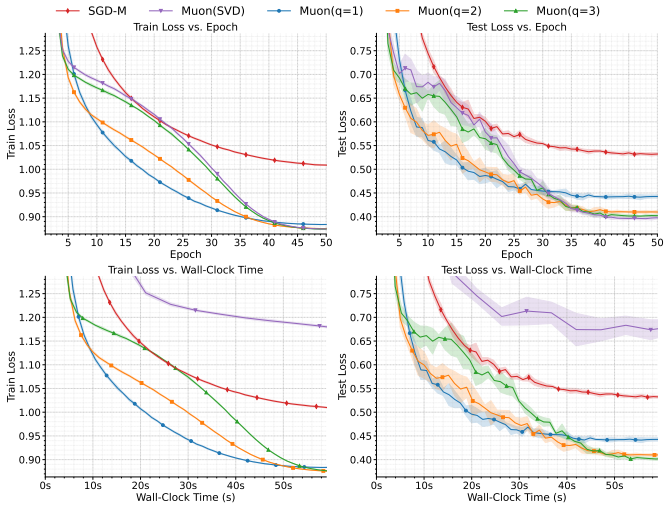


CIFAR-10 / CifarNet, $\kappa = 2$, averaged over 5 runs.

Experiments: a few Newton–Schulz steps are enough

Observed in the paper

- $q = 1$ already improves over SGD-M
- $q = 2, 3$ nearly match exact SVD by epoch
- In wall-clock time, NS-based Muon is much faster than SVD-based Muon



CIFAR-10 / CifarNet, $\kappa = 2$, averaged over 5 runs.

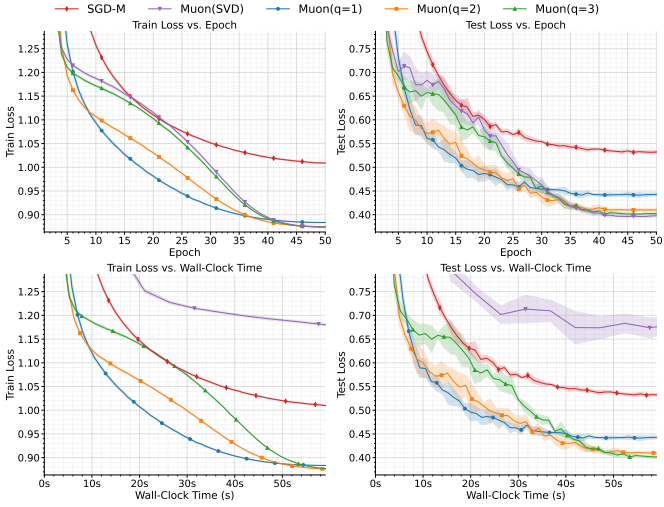
Experiments: a few Newton–Schulz steps are enough

Observed in the paper

- $q = 1$ already improves over SGD-M
- $q = 2, 3$ nearly match exact SVD by epoch
- In wall-clock time, NS-based Muon is much faster than SVD-based Muon

Why this matches theory

Theorems 1 and 2 predict near-SVD behavior once the factor χ_q becomes close to 1, which happens after only a few Newton–Schulz steps.



CIFAR-10 / CifarNet, $\kappa = 2$, averaged over 5 runs.

Conclusion

Main message

- We analyze Actual Muon, not an SVD idealization.

Conclusion

Main message

- We analyze Actual Muon, not an SVD idealization.
- Finite-step Newton–Schulz Muon has the same essential nonconvex convergence rate up to a factor that vanishes doubly exponentially in q.

Conclusion

Main message

- We analyze Actual Muon, not an SVD idealization.
- Finite-step Newton–Schulz Muon has the same essential nonconvex convergence rate up to a factor that vanishes doubly exponentially in q.
- Muon also gives a sharper rank dependence than SGD with momentum.

Conclusion

Main message

- We analyze Actual Muon, not an SVD idealization.
- Finite-step Newton–Schulz Muon has the same essential nonconvex convergence rate up to a factor that vanishes doubly exponentially in q.
- Muon also gives a sharper rank dependence than SGD with momentum.

A few Newton–Schulz steps are enough.

They make Muon both theoretically justified and practically efficient.

Thank you!