Preconditioning benefits of spectral orthogonalization in Muon

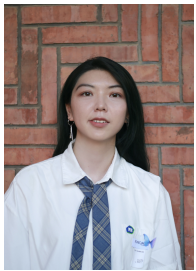
Jianhao Ma

Tsinghua University

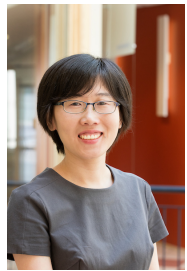
CFAI 2026

August 22, 2026

Coauthors



Yu Huang
UPenn

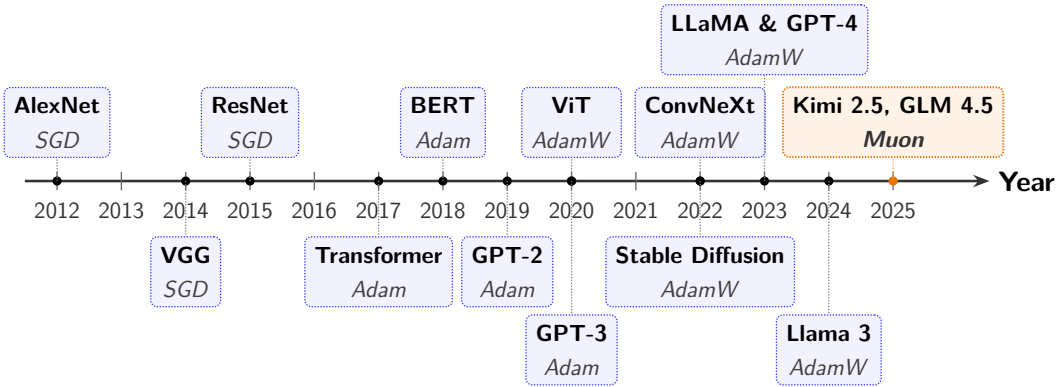


Yuejie Chi
Yale

Yuxin Chen
UPenn

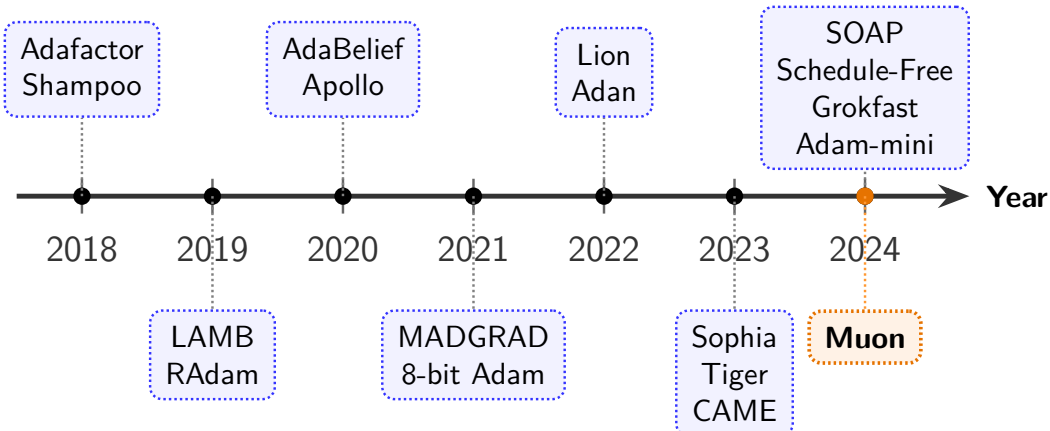
"Preconditioning benefits of spectral orthogonalization in Muon," J. Ma, Y. Huang, Y. Chi, Y. Chen, arXiv:2601.13474, 2026

Evolution of Deep Learning Optimizers



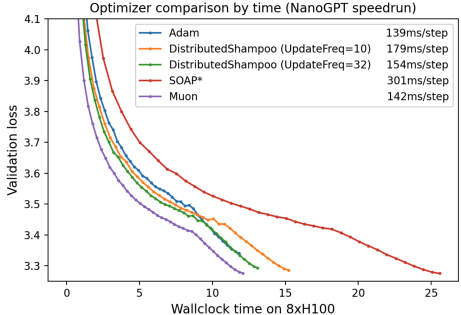
Muon (**M**oment**U**m **O**rthogonalized by **N**ewton-Schulz) optimizer
— Keller Jordan blog '24

The Quest to Dethrone Adam

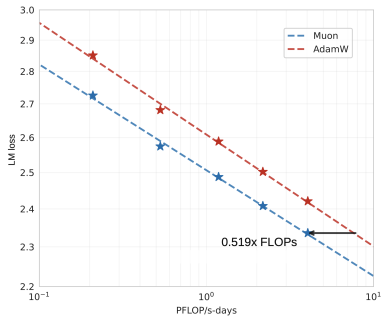


Muon performs well in practice

Companies are using Muon to train LLMs: OpenAI, DeepSeek, Moonshot AI, Z.ai, ...



(a) Muon's original blog



(b) Kimi report

Muon algorithm

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{m \times n}} \; f(\boldsymbol{\theta})$$

Muon algorithm

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{m\times n}} \; f(\boldsymbol{\theta})$$

Muon: for $t = 0, 1, \cdots$

$$\boldsymbol{B}_t = \nabla f(\boldsymbol{\theta}_t) + \mu \boldsymbol{B}_{t-1} \qquad \text{(gradient + momentum)}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \, \mathsf{msign}(\boldsymbol{B}_t) \qquad \text{(spectral orthogonalization)}$$

Muon algorithm

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{m \times n}} \; f(\boldsymbol{\theta})$$

Muon: for $t = 0, 1, \cdots$

$$\boldsymbol{B}_t = \nabla f(\boldsymbol{\theta}_t) + \mu \boldsymbol{B}_{t-1} \qquad \text{(gradient + momentum)}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \, \mathsf{msign}(\boldsymbol{B}_t) \qquad \text{(spectral orthogonalization)}$$

- $\underbrace{\mathsf{msign}(\boldsymbol{Z}) := \boldsymbol{U}\boldsymbol{V}^\top}_{\text{matrix sign function}}$ if $\boldsymbol{Z}$ has SVD $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

Muon algorithm

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{m \times n}} \; f(\boldsymbol{\theta})$$

Muon: for $t = 0, 1, \cdots$

$$\boldsymbol{B}_t = \nabla f(\boldsymbol{\theta}_t) + \mu \boldsymbol{B}_{t-1} \qquad \text{(gradient + momentum)}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \, \mathsf{msign}(\boldsymbol{B}_t) \qquad \text{(spectral orthogonalization)}$$

Simplified `Muon` (or spectral gradient method): for $t = 0, 1, \cdots$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \, \mathsf{msign}\big(\nabla f(\boldsymbol{\theta}_t)\big) \qquad \text{(no momentum)}$$

Newton-Schulz iteration

In practice, we use Newton-Schulz iteration to approximate the matrix sign, whose overhead is just ~ 1 to 3%.

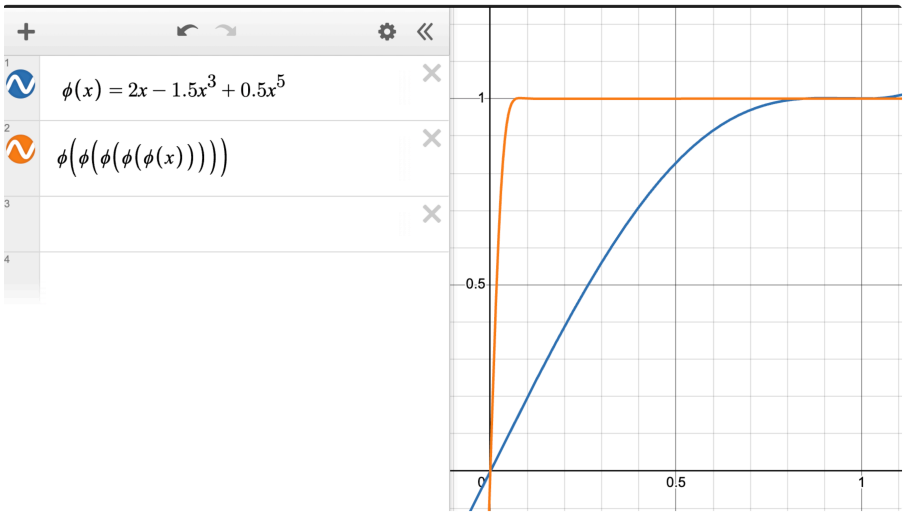
Newton-Schulz(5): given matrix $\boldsymbol{Z}$ and coefficients (a, b, c)

Set $\boldsymbol{O}_0 = \boldsymbol{Z} / \|\boldsymbol{Z}\|_{\mathsf{F}}$

For $k = 0, \ldots, 4$:

$$\boldsymbol{O}_{k+1} \leftarrow \phi(\boldsymbol{O}_k) := a\,\boldsymbol{O}_k + b\left(\boldsymbol{O}_k\boldsymbol{O}_k^\top\right)\boldsymbol{O}_k + c\left(\boldsymbol{O}_k\boldsymbol{O}_k^\top\right)^2\boldsymbol{O}_k$$

Newton-Schulz iteration (cont'd)

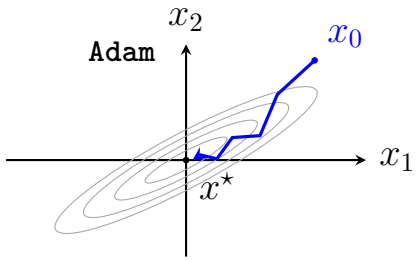


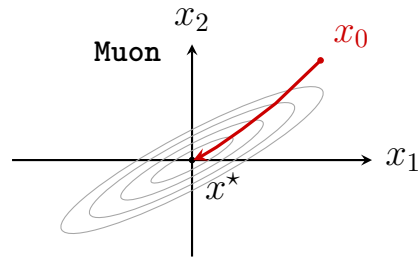
Existing theoretical analyses of Muon

- **Generic analysis:** Bernstein, Newhouse '24, Kovalev '25, Li, Hong '25, Chen et al. '25, An et al. '25, ...
- **One-step analysis:** Davis, Drusvyatskiy '25, Su '25, ...
- **Generalization:** Tveit et al. '25, Vasudeva et al. '25, Wang et al. '25, Zhang et al. '25, ...
- **Earlier work:** Carlson et al. '15, Gupta et al. '18, Tuddenham et al. '22, Vyas et al. '25, ...

Can we develop end-to-end theory to show provable benefits of Muon over other optimizers?

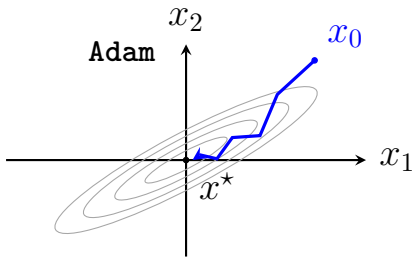
This talk: end-to-end theory of simplified `Muon`

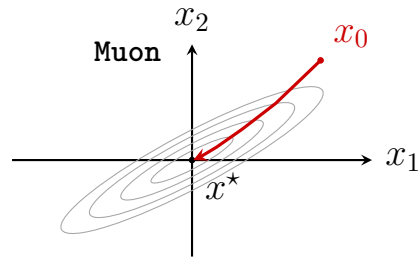




This talk: end-to-end theory of simplified `Muon`

preconditioning effect





This talk: end-to-end theory of simplified `Muon` (case studies)

preconditioning effect

- matrix factorization
- in-context learning of linear transformers

Matrix factorization

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad f(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathsf{F}}^2$$

- $\boldsymbol{M}^\star \succeq \boldsymbol{0}$: rank-$r$

Matrix factorization

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad f(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathsf{F}}^2$$

- $\boldsymbol{M}^\star \succeq \boldsymbol{0}$: rank-$r$
- Condition number: $\kappa := \sigma_{\max}(\boldsymbol{M}^\star)/\sigma_{\min}(\boldsymbol{M}^\star)$

Matrix factorization

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad f(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathsf{F}}^2$$

- $\boldsymbol{M}^\star \succeq \boldsymbol{0}$: rank-$r$
- Condition number: $\kappa := \sigma_{\max}(\boldsymbol{M}^\star)/\sigma_{\min}(\boldsymbol{M}^\star)$
- $k = r$: exact parameterization; $k > r$: over-parameterization

Matrix factorization

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad f(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathsf{F}}^2$$

simplified `Muon`: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

Matrix factorization

$$\underset{\boldsymbol{U}\in\mathbb{R}^{d\times k}}{\text{minimize}} \quad f(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}^\star\right\|_{\mathsf{F}}^2$$

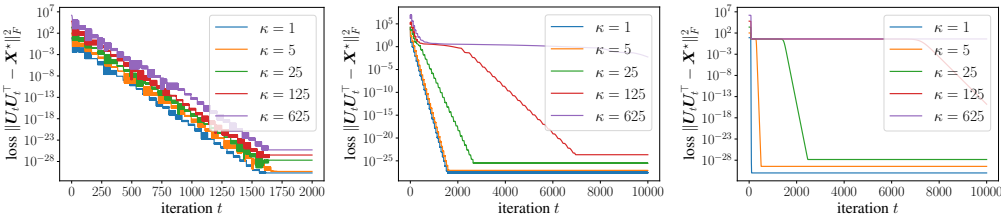
simplified `Muon`: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

vs.

$$\texttt{GD}: \ \boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$$

$$\underbrace{\texttt{SignGD}}_{\text{simplified Adam}}: \ \boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{sign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$$

Simulation results



(a) `Muon` (b) `SignGD` (c) `GD`

Loss vs. iterations for matrix factorization with varying condition numbers κ

Convergence theory of `Muon`

simplified `Muon`: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

Theorem 1 (exactly parameterized ($k = r$))

- Stepsize: $\eta_t = C_{\eta,t}\sqrt{\sigma^\star_{\max}}\,\rho^t$ for $2/3 \le \rho < 1$, $C_{\eta,t} \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(1,2)$
- Initialization: $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ w/ random orthonormal $\boldsymbol{O}$ and small α

Convergence theory of Muon

simplified Muon: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathrm{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

Theorem 1 (exactly parameterized ($k = r$))

- Stepsize: $\eta_t = C_{\eta,t}\sqrt{\sigma^\star_{\max}}\,\rho^t$ for $2/3 \le \rho < 1$, $C_{\eta,t} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(1,2)$
- Initialization: $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ w/ random orthonormal $\boldsymbol{O}$ and small α

With prob. at least 0.99, one has $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| \le \varepsilon$ if

$$T = \frac{1}{1-\rho}\log\left(\frac{16\sigma^\star_{\max}}{(1-\rho)^2\varepsilon}\right)$$

Convergence theory of Muon (cont'd)

simplified Muon: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathrm{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

Theorem 1 (mildly over-parameterized ($r \le k < d$))

- Stepsize: $\eta_t = C_{\eta,t}\sqrt{\sigma^\star_{\max}}\,\rho^t$ for $2/3 \le \rho < 1$, $C_{\eta,t} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(1,2)$
- Initialization: $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ w/ random orthonormal $\boldsymbol{O}$ and small α

With prob. at least 0.99, one has $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| \le \varepsilon$ if

$$T = \frac{1}{1-\rho}\log\left(\frac{16\sigma^\star_{\max}}{(1-\rho)^2\varepsilon}\right)$$

Convergence theory of Muon (cont'd)

simplified Muon: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

Theorem 2 (heavily over-parameterized ($k \geq d$))

- Stepsize: $\eta_t = C_\eta\sqrt{\sigma^\star_{\max}}\,\rho^t$ for $1/2 \leq \rho < 1$, $C_\eta \sim \mathsf{Unif}(1,2)$
- Initialization: $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ w/ orthonormal $\boldsymbol{O}$, $0 < \alpha \leq C_\eta\sqrt{\lambda^\star_{\max}}$

Convergence theory of Muon (cont'd)

simplified Muon: $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right)$

Theorem 2 (heavily over-parameterized ($k \geq d$))

- Stepsize: $\eta_t = C_\eta\sqrt{\sigma^\star_{\max}}\,\rho^t$ for $1/2 \leq \rho < 1$, $C_\eta \sim \mathsf{Unif}(1,2)$
- Initialization: $\boldsymbol{U}_0 = \alpha\boldsymbol{O}$ w/ orthonormal $\boldsymbol{O}$, $0 < \alpha \leq C_\eta\sqrt{\lambda^\star_{\max}}$

With prob. 1, one has $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{M}^\star\right\| \leq \varepsilon$ if

$$T \geq \frac{1}{1-\rho}\log\left(\frac{8\sigma^\star_{\max}}{\varepsilon}\right)$$

Comparison with other optimizers

- GD: needs $\kappa \log \frac{1}{\varepsilon}$ iterations

Comparison with other optimizers

- GD: needs $\kappa \log \frac{1}{\varepsilon}$ iterations
- SignGD (simplified Adam): needs $\Omega(\kappa)$ iterations

Theorem 3 (lower bound for SignGD (simplified Adam))

Consider SignGD w/ non-increasing stepsizes and $\eta_0 \leq 1/16$. There exists a problem instance such that: for any $\varepsilon \lesssim \kappa^{-2}$, SignGD w/ warm start needs at least

$$\frac{\kappa - 1}{4} \quad \text{iterations to yield } \varepsilon \text{ accuracy}$$

Comparison with other optimizers (cont'd)

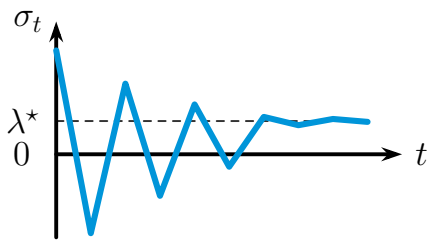
algorithm	parameterization	# of iterations	paper
simplified `Muon`	exactly-parameterized	$\log \frac{1}{\varepsilon}$	this work
	over-parameterized	$\log \frac{1}{\varepsilon}$	this work
`GD`	exactly-parameterized	$\kappa \log \frac{1}{\varepsilon}$	Chi et al. '19
	over-parameterized	$\kappa^3 \log \frac{1}{\varepsilon}$	Stöger et al. '21
	lower bound	$\kappa \log \frac{1}{\varepsilon}$	folklore
`SignGD`	lower bound	κ	this work

Analysis: scalar case ($d = 1$)

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta_t \ \mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right), \qquad t = 0, 1, \cdots$$

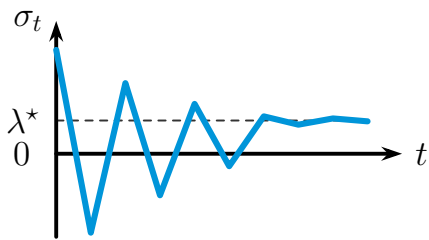
Analysis: scalar case ($d = 1$)

$$\sigma_{t+1} = \sigma_t - \eta_t \ \mathsf{sign}\left(\sigma_t^3 - \lambda^\star \sigma_t\right), \qquad t = 0, 1, \cdots$$



Analysis: scalar case ($d = 1$)

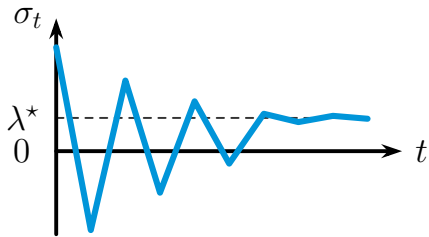
$$\sigma_{t+1} = \sigma_t - \eta_t \,\mathsf{sign}\left(\sigma_t^3 - \lambda^\star \sigma_t\right), \qquad t = 0, 1, \cdots$$



linear convergence: using $\eta_t = C\sqrt{|\lambda^\star|}\,\rho^t$ for $C \geq 1,\ \rho \in [1/2, 1)$:

Analysis: scalar case ($d = 1$)

$$\sigma_{t+1} = \sigma_t - \eta_t \ \mathsf{sign}\left(\sigma_t^3 - \lambda^\star \sigma_t\right), \qquad t = 0, 1, \cdots$$



linear convergence: using $\eta_t = C\sqrt{|\lambda^\star|}\,\rho^t$ for $C \geq 1,\ \rho \in [1/2, 1)$:

$$\left|\sigma_{t+1}^2 - \lambda^\star\right| \ \lesssim\ |\lambda^\star|\rho^t \qquad (\underline{\textbf{constant}}\ \text{convergence rate})$$

The case w/ exact subspace alignment

$$\boldsymbol{M}^\star = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top}$$

The case w/ exact subspace alignment

$$\boldsymbol{M}^\star = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top}$$

if $\boldsymbol{U}_0 = \underbrace{\boldsymbol{V}^\star}_{\text{exact subspace}} \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top$ for ortho $\boldsymbol{R}$, then

The case w/ exact subspace alignment

$$\boldsymbol{M}^\star = \boldsymbol{V}^\star \boldsymbol{\Lambda}^\star \boldsymbol{V}^{\star\top}$$

if $\boldsymbol{U}_0 = \underbrace{\boldsymbol{V}^\star}_{\text{exact subspace}} \boldsymbol{\Sigma}_0 \boldsymbol{R}^\top$ for ortho $\boldsymbol{R}$, then

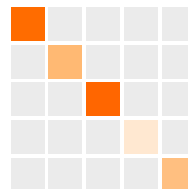
$$\begin{aligned}\boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \,\mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right) \\ &= \underbrace{\boldsymbol{V}^\star \boldsymbol{\Sigma}_t \boldsymbol{R}^\top - \eta_t \boldsymbol{V}^\star \mathsf{diag}\{\mathsf{sign}(\boldsymbol{\Sigma}_t^3 - \boldsymbol{\Lambda}^\star \boldsymbol{\Sigma}_t)\}\boldsymbol{R}^\top}_{\text{subspace is preserved}}\end{aligned}$$

The case w/ exact subspace alignment

$$\begin{aligned} \boldsymbol{U}_{t+1} &= \boldsymbol{U}_t - \eta_t \,\mathsf{msign}\left((\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{M}^\star)\boldsymbol{U}_t\right) \\ &= \underbrace{\boldsymbol{V}^\star\boldsymbol{\Sigma}_t\boldsymbol{R}^\top - \eta_t\boldsymbol{V}^\star\mathsf{diag}\{\mathsf{sign}(\boldsymbol{\Sigma}_t^3 - \boldsymbol{\Lambda}^\star\boldsymbol{\Sigma}_t)\}\boldsymbol{R}^\top}_{\text{subspace is preserved}} \end{aligned}$$

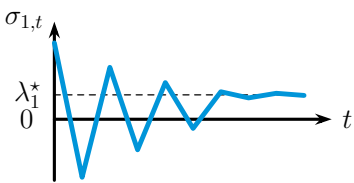
spectral dynamics:

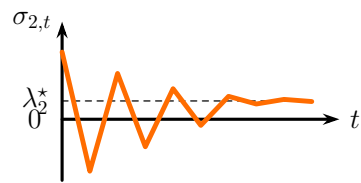
$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \eta_t\mathsf{diag}\{\mathsf{sign}(\boldsymbol{\Sigma}_t^3 - \boldsymbol{\Lambda}^\star\boldsymbol{\Sigma}_t)\}$$



The case w/ exact subspace alignment

Spectral dynamics of `Muon` decouple into d scalar sequences:

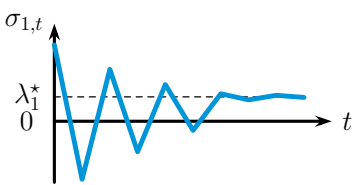


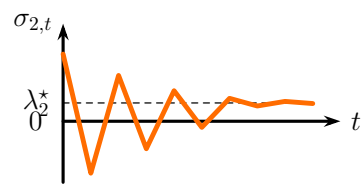


The case w/ exact subspace alignment

Spectral dynamics of `Muon` decouple into d scalar sequences:

- each sequence converges linearly w/ constant conv. rates





Analysis: more general case

Subspace alignment?

Analysis: more general case

Subspace alignment?

- always guaranteed when heavily over-parameterized ($k \geq d$)

Analysis: more general case

Subspace alignment?

- always guaranteed when heavily over-parameterized $(k \geq d)$
- approximately satisfied under small initialization $(r \leq k < d)$

$$\boldsymbol{U}_1 \approx \boldsymbol{U}_0 - \eta_0 \,\mathsf{msign}\left(\cancel{\boldsymbol{U}_0 \boldsymbol{U}_0^\top \boldsymbol{U}_0} - \boldsymbol{M}^\star \boldsymbol{U}_0\right)$$

Analysis: more general case

Subspace alignment?

- always guaranteed when heavily over-parameterized ($k \geq d$)
- approximately satisfied under small initialization ($r \leq k < d$)

$$\boldsymbol{U}_1 \approx \boldsymbol{U}_0 - \eta_0 \,\mathsf{msign}\left(\cancel{\boldsymbol{U}_0 \boldsymbol{U}_0^\top \boldsymbol{U}_0} - \boldsymbol{M}^\star \boldsymbol{U}_0\right)$$

- **challenge:** avoid blow-up of approx. error over time

Hessian perspective: connection to ScaledGD

- Vector form of Muon:

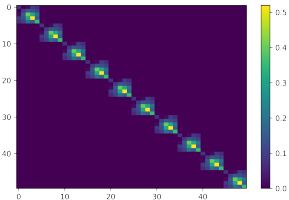
$$\mathsf{vec}(\boldsymbol{U}_{t+1}) = \mathsf{vec}(\boldsymbol{U}_t) - \eta_t \left(\boldsymbol{I} \otimes \left(\nabla f(\boldsymbol{U}_t)^\top \nabla f(\boldsymbol{U}_t) \right)^{1/2} \right)^{-1} \mathsf{vec}(\nabla f(\boldsymbol{U}_t))$$

- ScaledGD update:

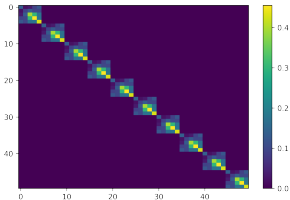
$$\mathsf{vec}(\boldsymbol{U}_{t+1}) = \mathsf{vec}(\boldsymbol{U}_t) - \eta_t \left(\boldsymbol{I} \otimes \left(\boldsymbol{U}_t^\top \boldsymbol{U}_t \right) \right)^{-1} \mathsf{vec}(\nabla f(\boldsymbol{U}_t)).$$

- ScaledGD has condition-number-free convergence rate $\log \frac{1}{\varepsilon}$ (Tong, Ma, Chi '21).

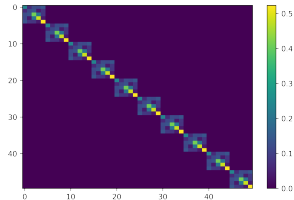
Connection to ScaledGD



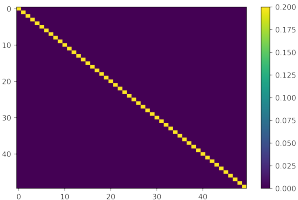
(a) Muon at $t = 0$



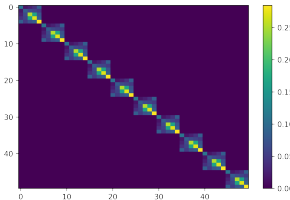
(b) Muon at $t = 500$



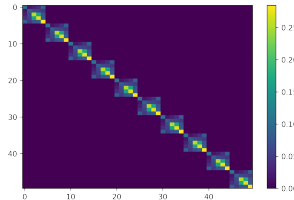
(c) Muon at $t = 1000$



(d) ScaledGD at $t = 0$

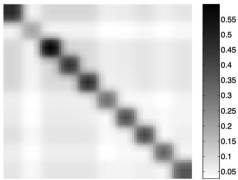


(e) ScaledGD at $t = 500$

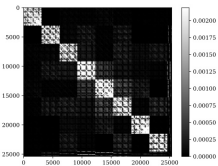


(f) ScaledGD at $t = 1000$

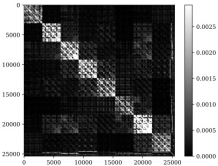
Block-diagonal structure in NN



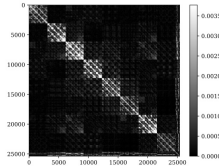
(a) Hessian of a MLP (Collobert, 2004)



(b) Hessian of a MLP at initialization



(c) Hessian of a MLP at 50% step



(d) Hessian of a MLP at 100% step

Figure: Adopted from Zhang et al. '25

Why does SignGD fail?

- SignGD only uses a diagonal preconditioner

$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \eta_t \,\mathsf{diag}\left\{|\mathsf{vec}(\nabla f(\boldsymbol{U}_t))|^{-1}\right\}\mathsf{vec}\big(\nabla f(\boldsymbol{U}_t)\big), \qquad (1)$$

  which cannot capture the block diagonal structure.

How to rigorously show SignGD is slow?

Construction of a bad instance

- **Our plan:** find a bad instance of a 2-d quadratic problem

$$\underset{\boldsymbol{z}\in\mathbb{R}^2}{\text{minimize}} \quad f(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^\top \boldsymbol{H}\boldsymbol{z}.$$

  Then, reduce our problems to this one.

- `Adam` performs well when $\boldsymbol{H}$ is (nearly) diagonal [1].

- Bad instance with condition number κ:

$$\boldsymbol{H} = \frac{1}{2}\begin{pmatrix} \kappa+1 & \kappa-1 \\ \kappa-1 & \kappa+1 \end{pmatrix}.$$

Coordinate update pattern

Upon defining $\widetilde{\boldsymbol{z}}_t = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \boldsymbol{z}_t$, we have:

Lemma

- If $|\kappa \widetilde{z}_{1,t}| > |\widetilde{z}_{2,t}|$, then

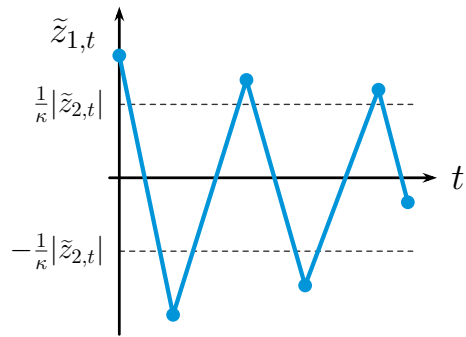
$$\widetilde{z}_{1,t+1} = \widetilde{z}_{1,t} - \sqrt{2}\eta_t \,\mathsf{sign}(\widetilde{z}_{1,t}), \qquad \widetilde{z}_{2,t+1} = \widetilde{z}_{2,t}. \tag{2}$$

- If $|\kappa \widetilde{z}_{1,t}| < |\widetilde{z}_{2,t}|$, then

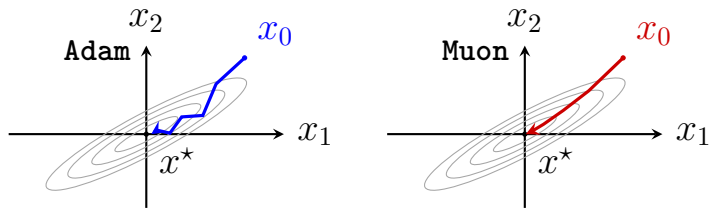
$$\widetilde{z}_{1,t+1} = \widetilde{z}_{1,t}, \qquad \widetilde{z}_{2,t+1} = \widetilde{z}_{2,t} - \sqrt{2}\eta_t \,\mathsf{sign}(\widetilde{z}_{2,t}). \tag{3}$$

Key intuition

- When $|\kappa\widetilde{z}_{1,t}| > |\widetilde{z}_{2,t}|$, we only update $\widetilde{z}_{1,t}$;
- When $|\kappa\widetilde{z}_{1,t}| < |\widetilde{z}_{2,t}|$, the learning rate η_t is already small;
- We need $O(\kappa)$ iterations to make $|\widetilde{z}_{2,t}|$ small.



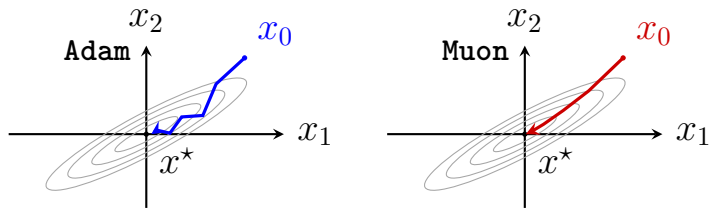
Concluding remarks



For matrix factorization & ICL of linear transformers:

- `Muon`: condition-number-free fast convergence
  - spectral dynamics: decouple into independent scalar sequences
- `Adam`: slow convergence if Hessian is not nearly diagonal

Concluding remarks



Future directions:

- understand benefits of momentum
- more case studies (recently, Gonon et al. '26, Li et al. '26, . . .)
- general theory that demonstrates benefits of `Muon`

References I


Rudrajit Das, Naman Agarwal, Sujay Sanghavi, and Inderjit S Dhillon.
Towards quantifying the preconditioning effect of Adam.
arXiv preprint arXiv:2402.07114, 2024.