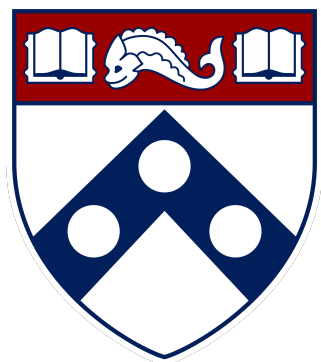
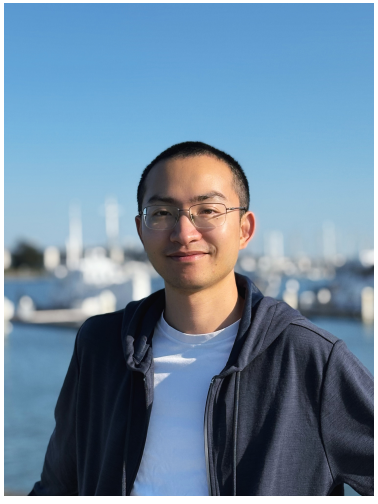
Preconditioning benefits of spectral orthogonalization in Muon



Yuxin Chen

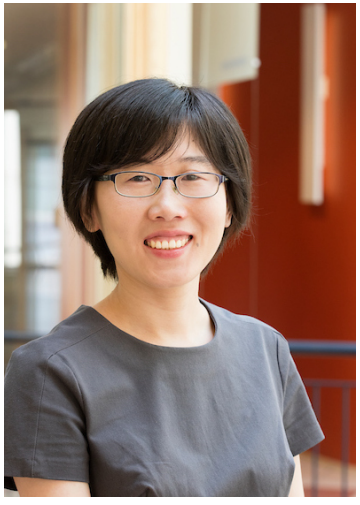
Wharton Statistics & Data Science

Coauthors



Jianhao Ma
UPenn

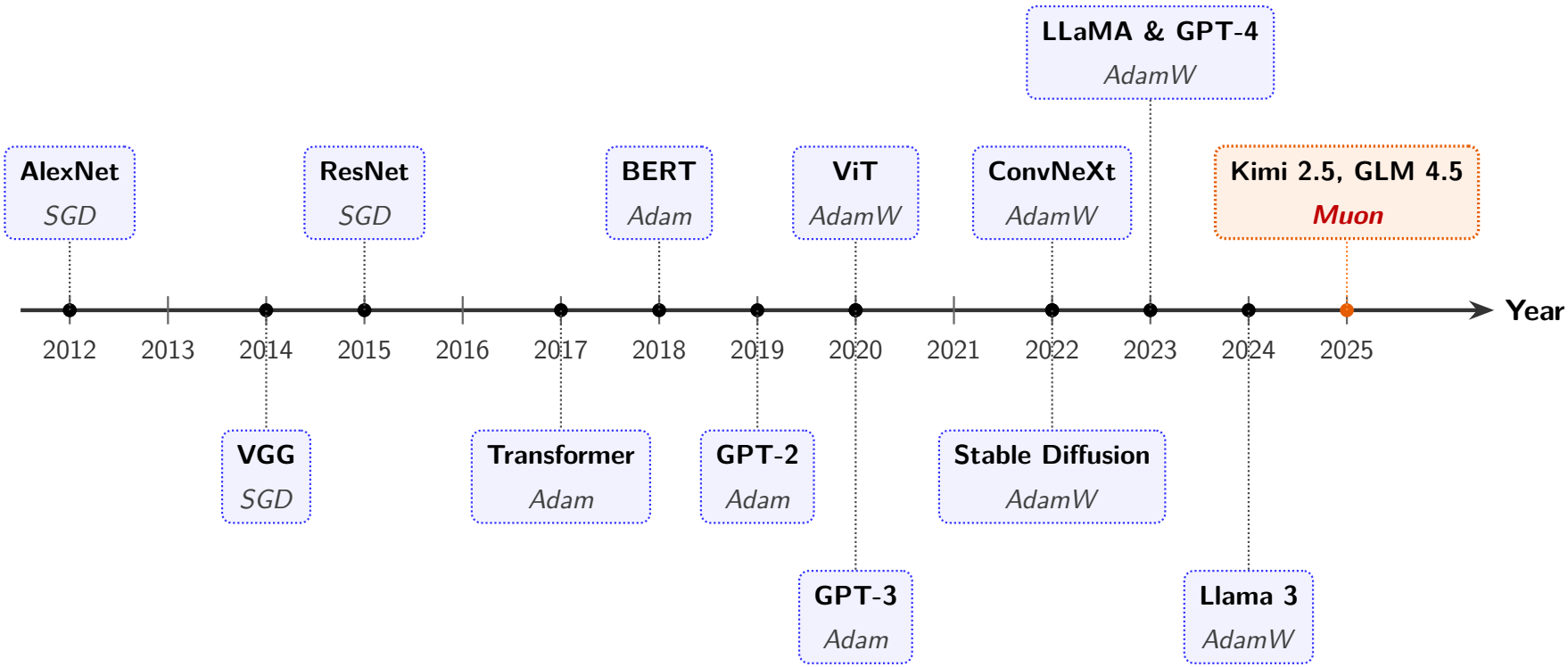
Yu Huang
UPenn



Yuejie Chi
Yale

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

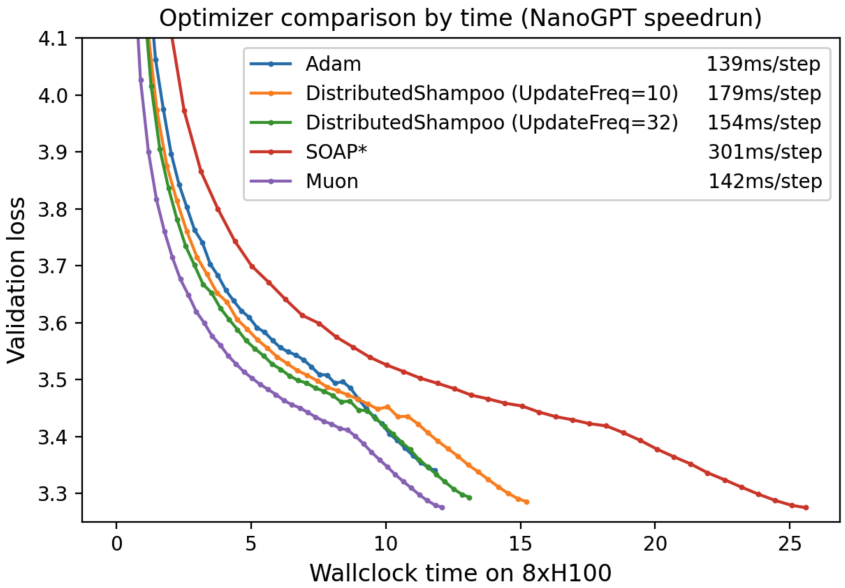
Evolution of deep learning optimizers



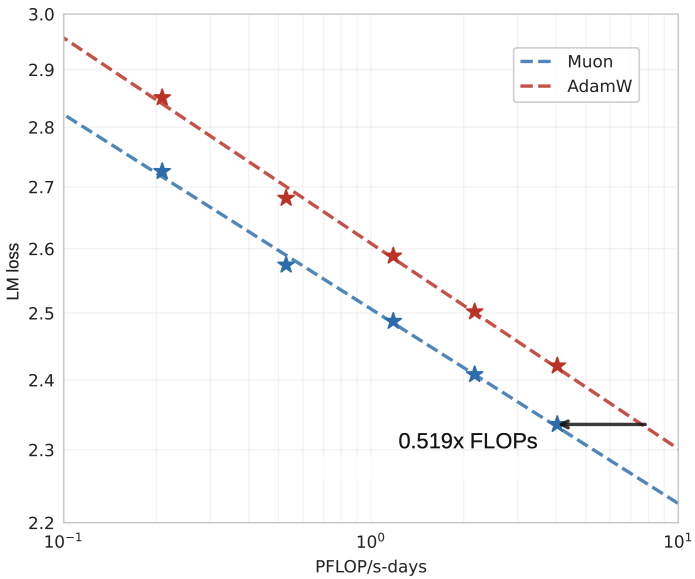
Muon (MomentUm Orthogonalized by Newton-Schulz) optimizer
— Keller Jordan blog '24

Companies are experimenting with Muon for LLM training

- Moonshot AI, Z.ai, OpenAI, DeepSeek, ...

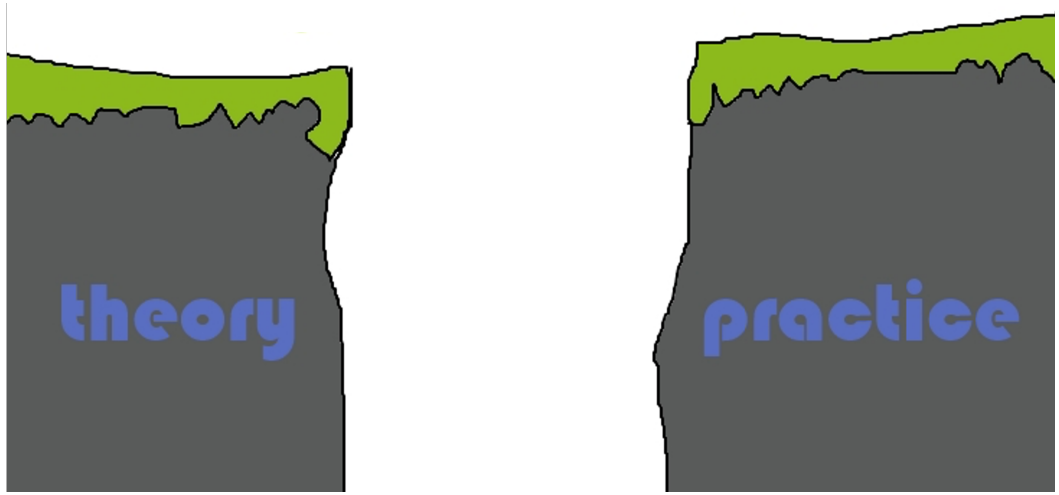


Muon's original blog

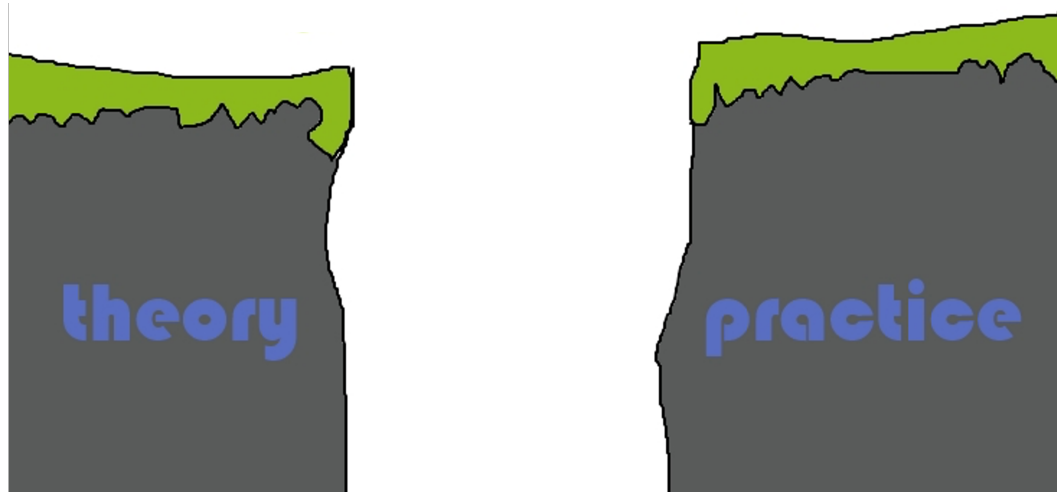


Kimi report

Theoretically, it is mysterious why Muon performs well in practice . . .



Theoretically, it is mysterious why Muon performs well in practice . . .



- *generic analysis:* Shen et al. '25, Li, Hong '25, Chen et al. '25, An et al. '25, Kovalev '25, . . .
- *one-step analysis:* Davis, Drusvyatskiy '25, Su '25, . . .
- *generalization:* 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?

Muon algorithm

$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X})$$

Muon algorithm

$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X})$$

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

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

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

Muon algorithm

$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X})$$

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

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

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

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

Muon algorithm

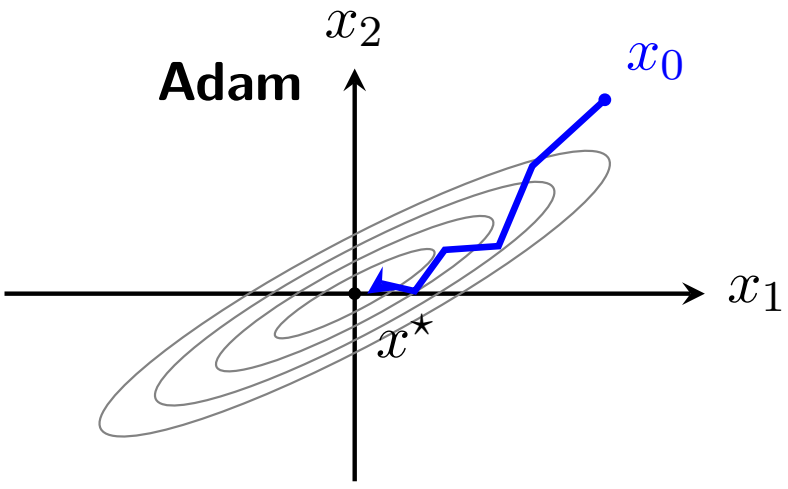
$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X})$$

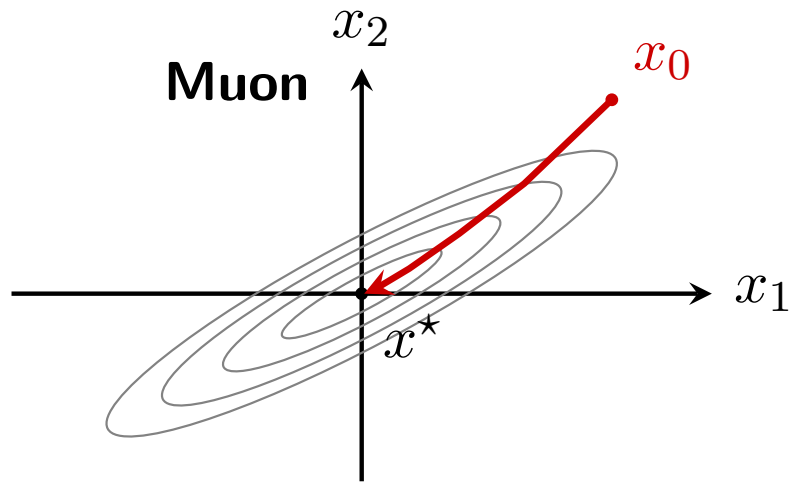
simplified Muon (or spectral gradient method): for $t = 0, 1, \cdots$

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta_t \,\text{msign}\left(\nabla f(\boldsymbol{X}_t)\right) \qquad \text{(no momentum)}$$

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

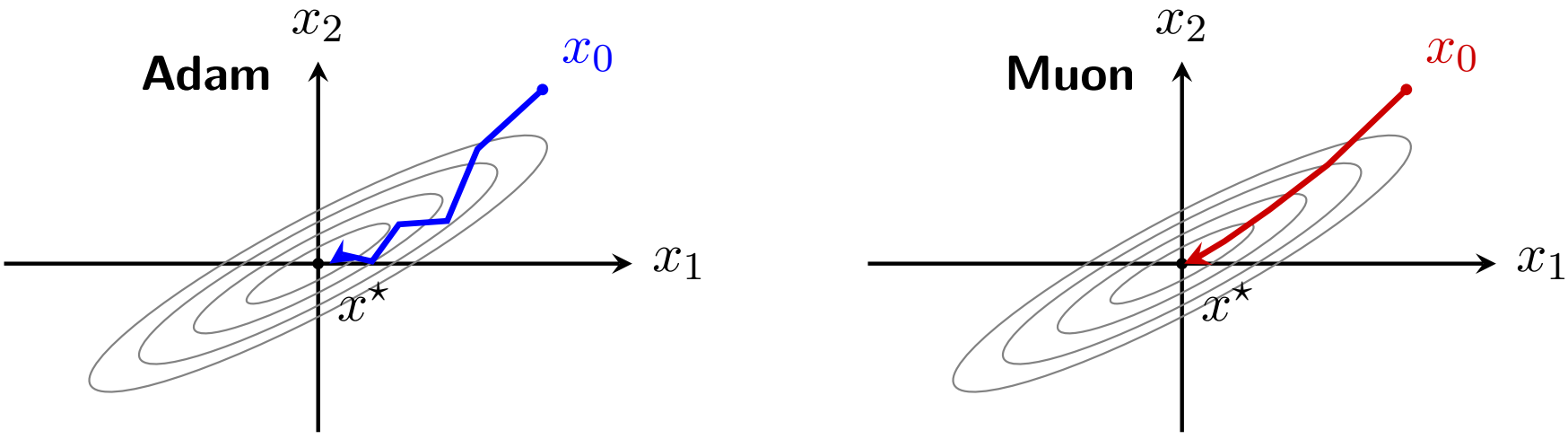
This talk: $\underbrace{\textit{end-to-end theory}}$ *of simplified Muon*





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

— Chi, Lu, Chen '19

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

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

Matrix factorization

— *Chi, Lu, Chen '19*

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

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

— *Chi, Lu, Chen '19*

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

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

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

— *Chi, Lu, Chen '19*

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

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

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.

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

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

Convergence theory of `Muon`

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

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

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

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

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

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

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

- linear convergence

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

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

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

- linear convergence
- iteration complexity independent of condition number 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

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

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

- linear convergence
- iteration complexity independent of condition number κ
- geometrically decaying stepsizes

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

- *stepsize:* $\eta_t = C_{\eta,t}\sqrt{\sigma^\star_{\max}}\rho^t$ *for* $2/3 \le \rho < 1$, $C_{\eta,t} \overset{\text{i.i.d.}}{\sim} \mathsf{Unif}(1,2)$
- *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*

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

- linear convergence
- iteration complexity independent of condition number κ
- geometrically decaying stepsizes

Convergence theory of `Muon`

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

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 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`

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*

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

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 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*

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

- linear convergence
- iteration complexity independent of condition number κ
- geometrically decaying stepsizes

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

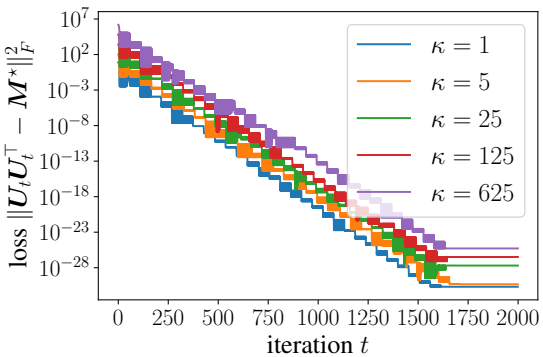
Theorem 3 (lower bound for `SignGD` (simplified `Adam`))

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

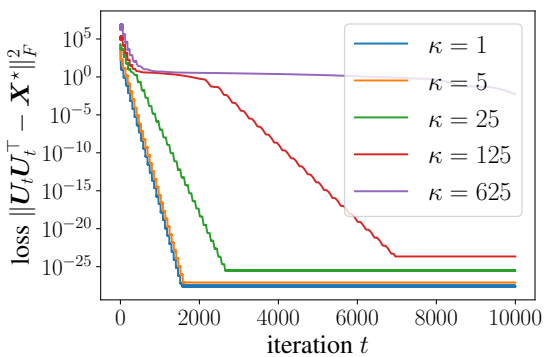
$$\frac{\kappa - 1}{4} \text{ iterations to yield } \varepsilon \text{ accuracy}$$

algorithm	parameterization	# 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	$\kappa^3 \log \frac{1}{\varepsilon}$	Stoger et al. '21
	lower bound	$\kappa \log \frac{1}{\varepsilon}$	folklore
`SignGD`	lower bound	κ	this work

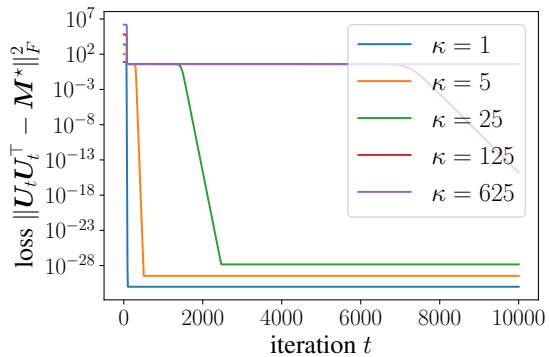
algorithm	parameterization	# iterations	paper
simplified `Muon`	exactly-parameterized	$\log \frac{1}{\varepsilon}$	this work
	over-parameterized	$\log \frac{1}{\varepsilon}$	this work
`GD`	exactly-parameterized	$\kappa \log \frac{1}{\varepsilon}$	Chi et al. '19
	over-parameterized	$\kappa^3 \log \frac{1}{\varepsilon}$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



(a) `Muon`



(b) `SignGD`



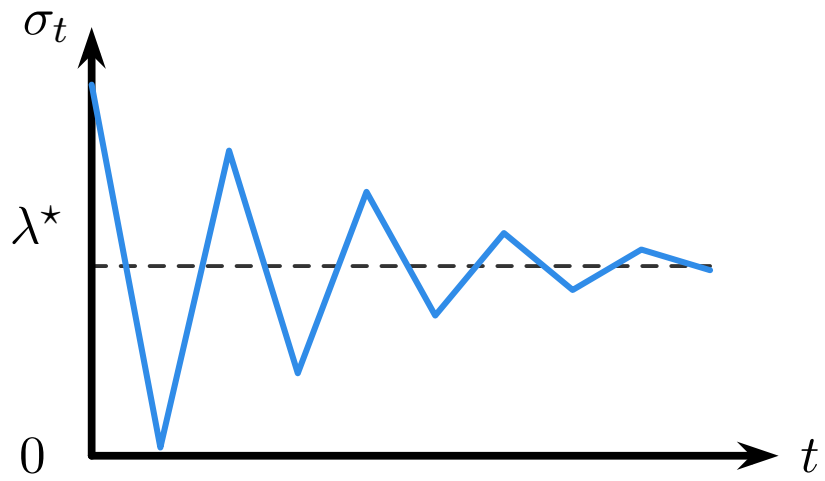
(c) `GD`

Analysis: scalar case $(d = 1)$

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

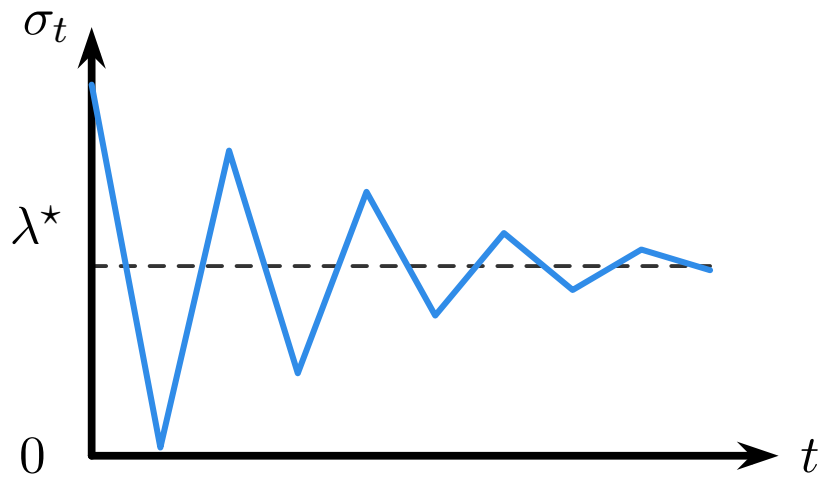
Analysis: scalar case $(d = 1)$

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



Analysis: scalar case $(d = 1)$

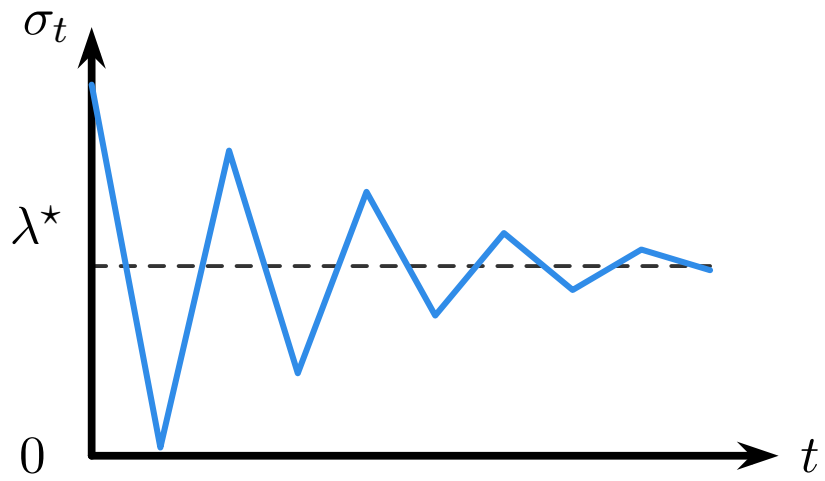
$$\sigma_{t+1} = \sigma_t - \eta_t \,\mathsf{sign}\,(\sigma_t^3 - \lambda^\star \sigma_t), \qquad t = 0, 1, \cdots$$



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

Analysis: scalar case $(d = 1)$

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



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

$$|\sigma_{t+1}^2 - \lambda^\star| \lesssim |\lambda^\star|\rho^t \qquad \textbf{(constant conv. rate)}$$

Analysis: the case w/ exact subspace alignment

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

Analysis: the case w/ exact subspace alignment

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

if $U_0 = \underbrace{V^\star}_{\text{exact subspace}} \Sigma_0 R^\top$ for ortho R, then

Analysis: 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

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

Analysis: the case w/ exact subspace alignment

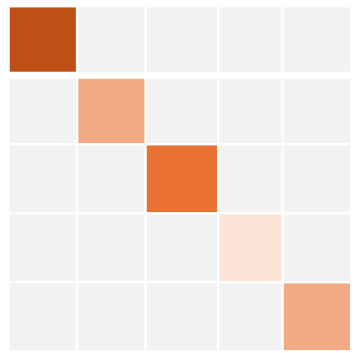
$$M^\star = V^\star \Lambda^\star V^{\star\top}$$

if $U_0 = \underbrace{V^\star}_{\text{exact subspace}} \Sigma_0 R^\top$ for ortho R, then

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

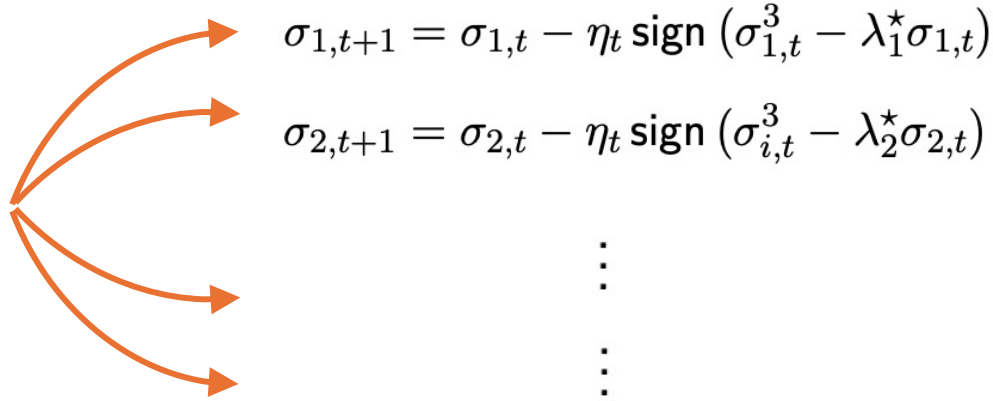
spectral dynamics:

$$\Sigma_{t+1} = \Sigma_t - \eta_t \,\mathrm{diag}\{\mathsf{sign}(\Sigma_t^3 - \Lambda^\star \Sigma_t)\}$$



Analysis: the case w/ exact subspace alignment

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



$$\sigma_{1,t+1} = \sigma_{1,t} - \eta_t \operatorname{sign}\left(\sigma_{1,t}^3 - \lambda_1^\star \sigma_{1,t}\right)$$

$$\sigma_{2,t+1} = \sigma_{2,t} - \eta_t \operatorname{sign}\left(\sigma_{i,t}^3 - \lambda_2^\star \sigma_{2,t}\right)$$

$$\vdots$$

$$\vdots$$

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

Analysis: the case w/ exact subspace alignment

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

$$\sigma_{1,t+1} = \sigma_{1,t} - \eta_t \, \mathsf{sign}\left(\sigma_{1,t}^3 - \lambda_1^\star \sigma_{1,t}\right)$$

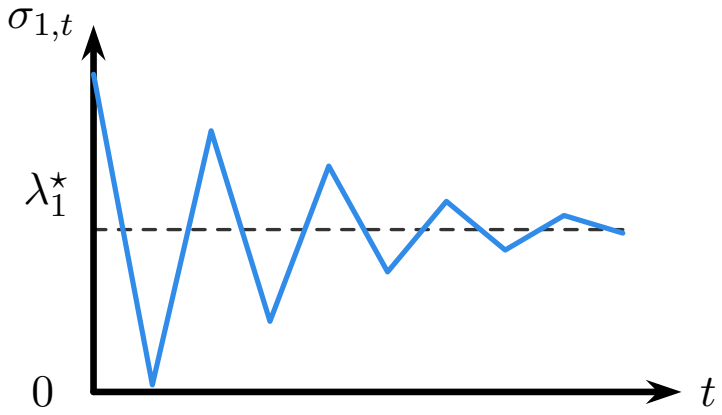
$$\sigma_{2,t+1} = \sigma_{2,t} - \eta_t \, \mathsf{sign}\left(\sigma_{i,t}^3 - \lambda_2^\star \sigma_{2,t}\right)$$

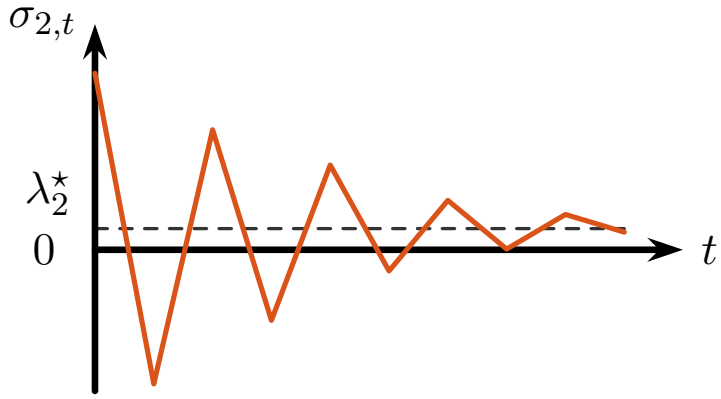
$$\vdots$$

$$\vdots$$

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 ($r \leq k < d$)

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

Analysis: more general case

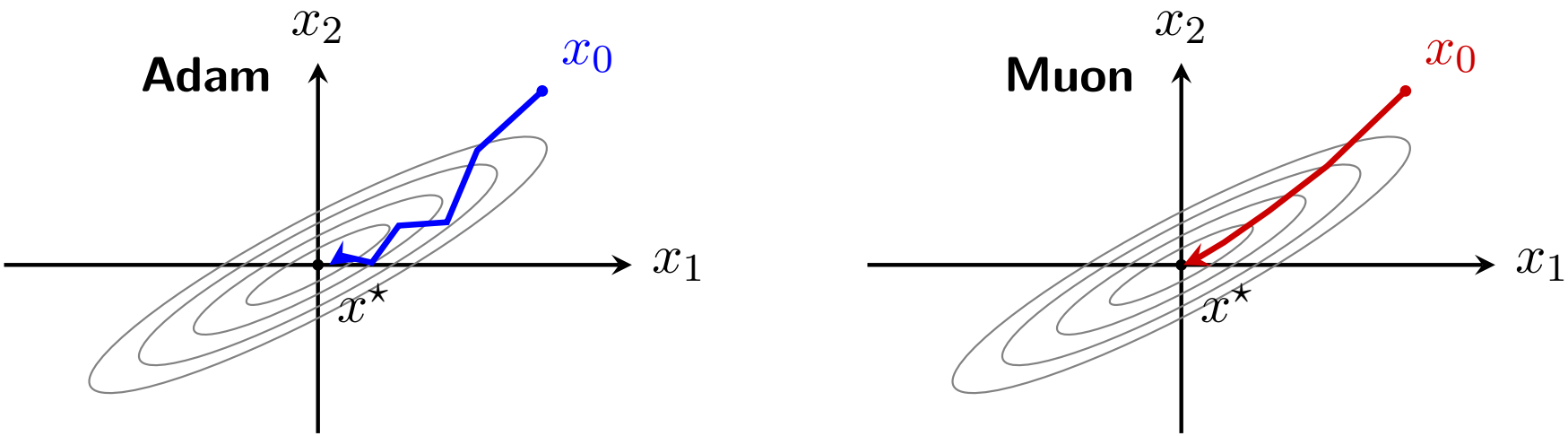
Subspace alignment?

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

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

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

Concluding remarks

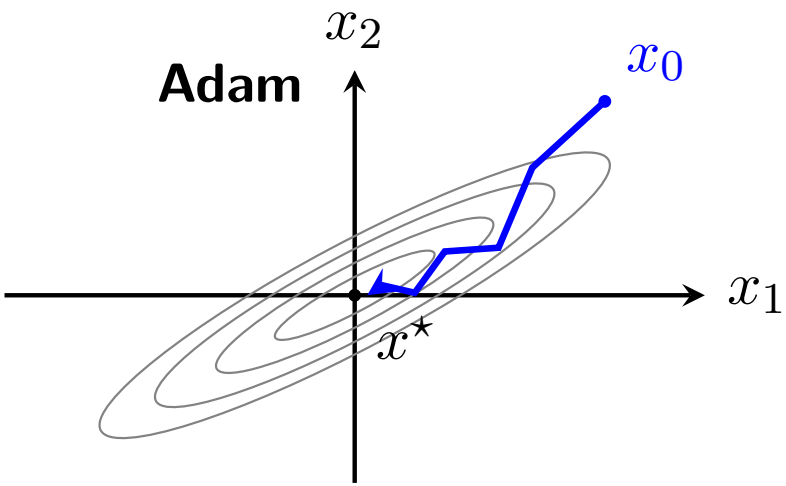


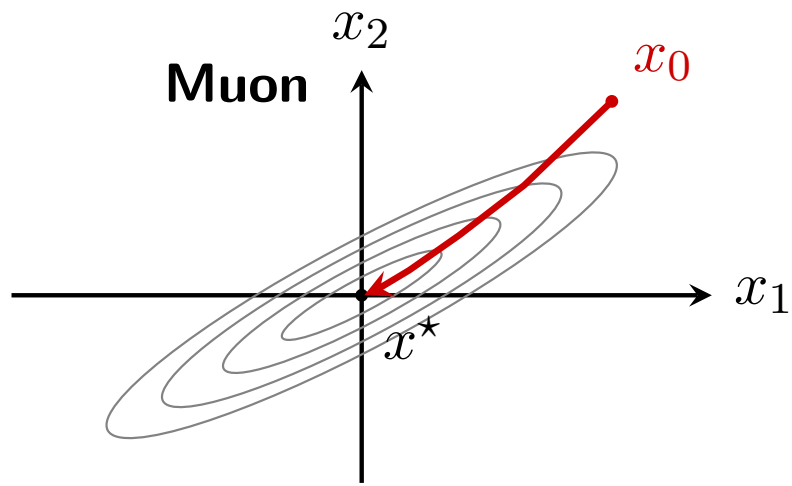
For matrix factorization & ICL of linear transformers:

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

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

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`

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