Controlled LLM Training On The Spectral Sphere

Spectral Sphere Optimizer

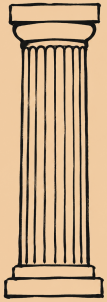
Author

Tian Xie, Haoming Luo, Haoyu Tang,
Yiwen Hu, J.K. Liu, Qingnan Ren,
Yang Wang, W.X. Zhao, Rui Yan,
Bing Su, Chong Luo, Baining Guo

GitHub

WeChat

Two Pillars



Convergence Speed

How fast does loss drop?

A: Steepest descent under a good norm

Measurable Stability

Do activations explode?

A: μP condition for feature learning

LLM training is, at its core, a pursuit of **convergence speed** grounded in the necessity of **stability**.

Fantastic Optimizers

Each one is steepest descent under a different geometry.

SSO: a fast & stable optimizer under spectral sphere

Steepest Descent

1. Linearize the loss **2.** cap the step under a norm **3.** minimize.

$$\Delta \boldsymbol{W} := \underset{\|\Delta \boldsymbol{W}\| \le \eta}{\arg\min} \Big\{ \underbrace{\mathcal{L}(\boldsymbol{W}) + \langle \boldsymbol{G}, \Delta \boldsymbol{W} \rangle}_{\text{first-order model}} \Big\}$$

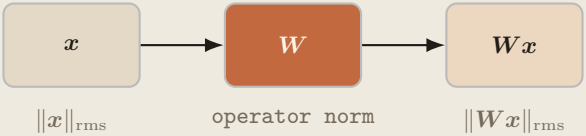
$$\Rightarrow \; \Delta \boldsymbol{W} = -\eta \cdot \underset{\|\boldsymbol{T}\| \le 1}{\arg\max} \langle \boldsymbol{G}, \boldsymbol{T} \rangle$$

Frobenius $\|\cdot\|_F$	**vector** ℓ_∞	**matrix spectral** $\|\cdot\|_2$
$\Delta \boldsymbol{W} = -\eta \boldsymbol{G}$	$\Delta \boldsymbol{W} = -\eta \operatorname{sign}(\boldsymbol{G})$	$\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \mapsto \boldsymbol{U}\boldsymbol{V}^\top$
SGD	**Adam** (no EMA)	**Muon** (msign)

Muon- steepest descent under the *spectral norm*

μP Condition: Hidden Stability

Stable training = activations stay $\Theta(1)$. RMS norm $\|\boldsymbol{x}\|_{\text{rms}} = \|\boldsymbol{x}\|_2/\sqrt{d}$.



Forward stability — a layer must preserve the activation scale :

$$\|\boldsymbol{W}\|_{\text{rms}\to\text{rms}} \quad := \quad \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{Wx}\|_{\text{rms}}}{\|\boldsymbol{x}\|_{\text{rms}}} \quad = \quad \Theta(1)$$

μP Condition: Hidden Stability

Plug in $\|\cdot\|_{\text{rms}} = \|\cdot\|_2/\sqrt{d}$: the $\sqrt{d}$'s pull out, leaving the ordinary spectral norm $\|\boldsymbol{W}\|_2$ scaled by the widths $d_{\text{in}} \to d_{\text{out}}$.

$$\|\boldsymbol{W}\|_{\text{rms}\to\text{rms}} = \sqrt{\tfrac{d_{\text{in}}}{d_{\text{out}}}} \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{W}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sqrt{\tfrac{d_{\text{in}}}{d_{\text{out}}}}\, \|\boldsymbol{W}\|_2 = \Theta(1)$$

Forward $\to$ weight

$$\|\boldsymbol{W}\|_2 \;=\; \Theta\big(\sqrt{d_{\text{out}}/d_{\text{in}}}\big)$$

Update $\to$ optimizer

$$\|\Delta\boldsymbol{W}\|_2 \;=\; \Theta\big(\sqrt{d_{\text{out}}/d_{\text{in}}}\big)$$

Spectral μP Condition: $\|\boldsymbol{W}\|_2 \;=\; \|\Delta\boldsymbol{W}\|_2 \;=\; \Theta\Big(\sqrt{d_{\text{out}}/d_{\text{in}}}\Big) \;=:\; R$

Muon: Half-Aligned μP

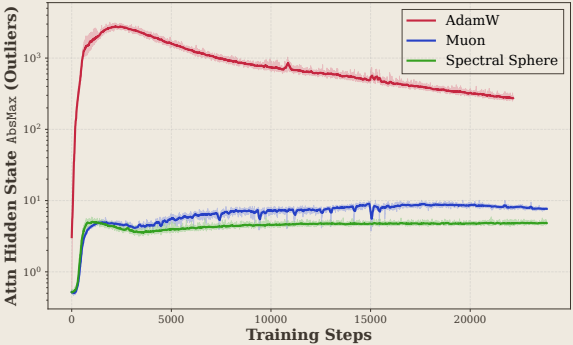
Muon controls $\|\Delta W\|_2$ ✓

Muon does *not* control $\|W\|_2$ ✗

→ effective step $\|\Delta W\|_2/\|W\|_2$ drifts

→ attention-logit blow-ups in long runs

→ needs *non-essential* SandwichNorm / QK-Norm / softcap patches



Attention activation AbsMax, Dense-1.7B pre-training

Spectral Sphere Optimizer

What if one optimizer could simultaneously satisfy
the steepest-descent property for **convergence speed**,
and the strict μP constraints for **fundamental stability**?

The Objective

Pin **both** the weight $\boldsymbol{W}$ *and* its update onto the spectral sphere of radius $R = \Theta(\sqrt{d_{\text{out}}/d_{\text{in}}})$. Parameterize $\Delta \boldsymbol{W} = \eta R\, \boldsymbol{\Phi}$ and solve for the unit step $\boldsymbol{\Phi}$:

$$\max_{\boldsymbol{\Phi}} \langle \boldsymbol{G}, \boldsymbol{\Phi} \rangle$$

The Objective

Pin **both** the weight $\boldsymbol{W}$ *and* its update onto the spectral sphere of radius $R = \Theta(\sqrt{d_{\text{out}}/d_{\text{in}}})$. Parameterize $\Delta \boldsymbol{W} = \eta R\, \boldsymbol{\Phi}$ and solve for the unit step $\boldsymbol{\Phi}$:

$$\max_{\boldsymbol{\Phi}} \langle \boldsymbol{G}, \boldsymbol{\Phi} \rangle$$

subject to

$$\|\boldsymbol{\Phi}\|_2 = 1$$

steepest descent under the spectral norm

The Objective

Pin **both** the weight $\boldsymbol{W}$ *and* its update onto the spectral sphere of radius $R = \Theta(\sqrt{d_{\text{out}}/d_{\text{in}}})$. Parameterize $\Delta \boldsymbol{W} = \eta R\, \boldsymbol{\Phi}$ and solve for the unit step $\boldsymbol{\Phi}$:

$$\max_{\boldsymbol{\Phi}} \langle \boldsymbol{G}, \boldsymbol{\Phi} \rangle$$

subject to

$$\|\boldsymbol{\Phi}\|_2 = 1$$

steepest descent under the spectral norm

$$\|\boldsymbol{W} - \eta R\, \boldsymbol{\Phi}\|_2 = \|\boldsymbol{W}\|_2 = R$$

$\boldsymbol{W}$ *stays* on the sphere of radius R

First Order: From Sphere to Tangent

Spectral norm is differentiable when the top singular value is simple (a.s. for random matrices). Its gradient is the **top singular pair**:

$$\boldsymbol{\Theta} \;:=\; \nabla_{\boldsymbol{W}} \|\boldsymbol{W}\|_2 \;=\; \boldsymbol{u}_1 \boldsymbol{v}_1^\top .$$

First Order: From Sphere to Tangent

Spectral norm is differentiable when the top singular value is simple (a.s. for random matrices). Its gradient is the **top singular pair**:

$$\boldsymbol{\Theta} \;:=\; \nabla_{\boldsymbol{W}} \|\boldsymbol{W}\|_2 \;=\; \boldsymbol{u}_1 \boldsymbol{v}_1^\top .$$

First-order Taylor expansion of the moving-weight constraint about $\boldsymbol{W}$:

$$\|\boldsymbol{W} - \eta R\, \boldsymbol{\Phi}\|_2 \;=\; \|\boldsymbol{W}\|_2 \;-\; \eta R \,\langle \boldsymbol{\Theta}, \boldsymbol{\Phi} \rangle \;+\; \mathcal{O}(\eta^2 R^2).$$

First Order: From Sphere to Tangent

Spectral norm is differentiable when the top singular value is simple (a.s. for random matrices). Its gradient is the **top singular pair**:

$$\boldsymbol{\Theta} \;:=\; \nabla_{\boldsymbol{W}} \|\boldsymbol{W}\|_2 \;=\; \boldsymbol{u}_1 \boldsymbol{v}_1^\top.$$

First-order Taylor expansion of the moving-weight constraint about $\boldsymbol{W}$:

$$\|\boldsymbol{W} - \eta R\, \boldsymbol{\Phi}\|_2 \;=\; \|\boldsymbol{W}\|_2 \;-\; \eta R\, \langle \boldsymbol{\Theta}, \boldsymbol{\Phi} \rangle \;+\; \mathcal{O}(\eta^2 R^2).$$

Matching $\|\boldsymbol{W} - \eta R\, \boldsymbol{\Phi}\|_2 \;=\; \|\boldsymbol{W}\|_2$ at first order $\Rightarrow$ **tangent constraint** $\langle \boldsymbol{\Theta}, \boldsymbol{\Phi} \rangle \;=\; 0.$

First Order: From Sphere to Tangent

Spectral norm is differentiable when the top singular value is simple (a.s. for random matrices). Its gradient is the **top singular pair**:

$$\boldsymbol{\Theta} \;:=\; \nabla_{\boldsymbol{W}} \|\boldsymbol{W}\|_2 \;=\; \boldsymbol{u}_1 \boldsymbol{v}_1^\top .$$

First-order Taylor expansion of the moving-weight constraint about $\boldsymbol{W}$:

$$\|\boldsymbol{W} - \eta R\, \boldsymbol{\Phi}\|_2 \;=\; \|\boldsymbol{W}\|_2 \;-\; \eta R\, \langle \boldsymbol{\Theta}, \boldsymbol{\Phi} \rangle \;+\; \mathcal{O}(\eta^2 R^2).$$

Matching $\|\boldsymbol{W} - \eta R\, \boldsymbol{\Phi}\|_2 \;=\; \|\boldsymbol{W}\|_2$ at first order $\Rightarrow$ **tangent constraint** $\langle \boldsymbol{\Theta}, \boldsymbol{\Phi} \rangle \;=\; 0.$

the problem collapses to

$$\max_{\boldsymbol{\Phi}} \langle \boldsymbol{G}, \boldsymbol{\Phi} \rangle \quad \text{s.t.} \quad \|\boldsymbol{\Phi}\|_2 \;=\; 1, \;\; \langle \boldsymbol{\Theta}, \boldsymbol{\Phi} \rangle \;=\; 0$$

The Solution: One Rank-1 Correction

A single **Lagrange multiplier** λ folds the tangent constraint into the objective:

$$\mathcal{L}(\mathbf{\Phi}; \lambda) \;=\; \langle \boldsymbol{G} + \lambda \mathbf{\Theta},\; \mathbf{\Phi} \rangle, \qquad \|\mathbf{\Phi}\|_2 = 1.$$

The Solution: One Rank-1 Correction

A single **Lagrange multiplier** λ folds the tangent constraint into the objective:

$$\mathcal{L}(\mathbf{\Phi}; \lambda) \;=\; \langle \boldsymbol{G} + \lambda \mathbf{\Theta},\; \mathbf{\Phi} \rangle, \qquad \|\mathbf{\Phi}\|_2 = 1.$$

Steepest descent under the spectral norm $\Rightarrow$ optimum is an msign:

$$\mathbf{\Phi}^\star(\lambda) \qquad = \qquad \mathrm{msign}(\boldsymbol{G} \;+\; \lambda \mathbf{\Theta})$$

The Solution: One Rank-1 Correction

A single **Lagrange multiplier** λ folds the tangent constraint into the objective:

$$\mathcal{L}(\mathbf{\Phi}; \lambda) \;=\; \langle \boldsymbol{G} + \lambda \mathbf{\Theta},\; \mathbf{\Phi} \rangle, \qquad \|\mathbf{\Phi}\|_2 = 1.$$

Steepest descent under the spectral norm $\Rightarrow$ optimum is an msign:

$$\mathbf{\Phi}^\star(\lambda) \qquad = \qquad \mathrm{msign}(\boldsymbol{G} \;+\; \lambda \mathbf{\Theta})$$

`pick` $\lambda^\star$ `so the tangent constraint holds`

$$h(\lambda) \qquad := \qquad \langle \mathbf{\Theta}, \quad \mathrm{msign}(\boldsymbol{G} \;+\; \lambda \mathbf{\Theta}) \rangle \qquad = \qquad 0$$

The Solution: One Rank-1 Correction

A single **Lagrange multiplier** λ folds the tangent constraint into the objective:

$$\mathcal{L}(\mathbf{\Phi}; \lambda) \;=\; \langle \boldsymbol{G} + \lambda \mathbf{\Theta},\; \mathbf{\Phi} \rangle, \qquad \|\mathbf{\Phi}\|_2 = 1.$$

Steepest descent under the spectral norm $\Rightarrow$ optimum is an msign:

$$\mathbf{\Phi}^\star(\lambda) \qquad = \qquad \mathrm{msign}(\boldsymbol{G} \;+\; \lambda \mathbf{\Theta})$$

`pick` $\lambda^\star$ `so the tangent constraint holds`

$$h(\lambda) \qquad := \qquad \langle \mathbf{\Theta}, \quad \mathrm{msign}(\boldsymbol{G} \;+\; \lambda \mathbf{\Theta}) \rangle \qquad = \qquad 0$$

$\lambda{=}0 \quad \Rightarrow$ exactly Muon's $\mathrm{msign}(\boldsymbol{G})$.
SSO adds *one rank-1 correction* $\lambda\mathbf{\Theta}$ that pins the step to the sphere.

Solving for $\lambda^\star$

Monotone. $h(\lambda)$ is non-decreasing, sweeping $-1 \to +1$ as λ runs over $\mathbb{R}$.

Solving for $\lambda^\star$

Monotone. $h(\lambda)$ is non-decreasing, sweeping $-1 \to +1$ as λ runs over $\mathbb{R}$.

Localized. $|\lambda^\star| \leq 2\|\boldsymbol{G}\|_*$ — a finite search window.

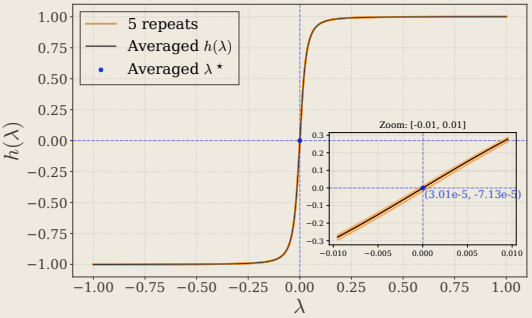
Solving for $\lambda^\star$

Monotone. $h(\lambda)$ is non-decreasing, sweeping $-1 \to +1$ as λ runs over $\mathbb{R}$.

Localized. $|\lambda^\star| \le 2\|\boldsymbol{G}\|_*$ — a finite search window.

Bracket from $\lambda{=}0$ (expand against $\operatorname{sign} h(0)$), then **bisect** to tolerance.

a 1-D root find -- a handful of msign calls per step.



$h(\lambda)$ on a random 1024×3072 matrix; root $\lambda^\star$ sits close to 0

Second Order: Retraction

The $\mathcal{O}(\eta^2 R^2)$ remainder accumulates and drifts $\boldsymbol{W}$ off the sphere. Restore $\|\boldsymbol{W}\|_2 = R$ *exactly* with a one-line **retraction**:

$$\boldsymbol{W} \quad \leftarrow \quad \boldsymbol{W} \cdot \frac{R}{\|\boldsymbol{W}\|_2}$$

Second Order: Retraction

The $\mathcal{O}(\eta^2 R^2)$ remainder accumulates and drifts $\boldsymbol{W}$ off the sphere. Restore $\|\boldsymbol{W}\|_2 = R$ *exactly* with a one-line **retraction**:

$$\boldsymbol{W} \quad \leftarrow \quad \boldsymbol{W} \cdot \frac{R}{\|\boldsymbol{W}\|_2}$$

Free. A pre-update step reuses *one* power iteration for $\|\boldsymbol{W}\|_2$ (cached power iteration).

Bounded. $\|\boldsymbol{W}\|_2 = R \quad \Rightarrow \quad \|\boldsymbol{W}\|_F \leq \sqrt{\operatorname{rank}}\, R.$

Second Order: Retraction

The $\mathcal{O}(\eta^2 R^2)$ remainder accumulates and drifts $\boldsymbol{W}$ off the sphere. Restore $\|\boldsymbol{W}\|_2 = R$ *exactly* with a one-line **retraction**:

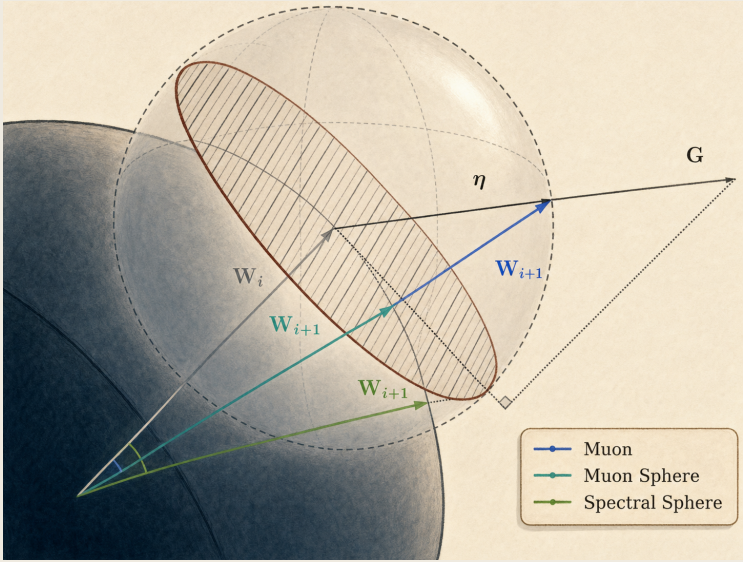
$$\boldsymbol{W} \quad \leftarrow \quad \boldsymbol{W} \cdot \frac{R}{\|\boldsymbol{W}\|_2}$$

Free. A pre-update step reuses *one* power iteration for $\|\boldsymbol{W}\|_2$ (cached power iteration).

Bounded. $\|\boldsymbol{W}\|_2 = R \quad \Rightarrow \quad \|\boldsymbol{W}\|_F \leq \sqrt{\operatorname{rank} R}$.

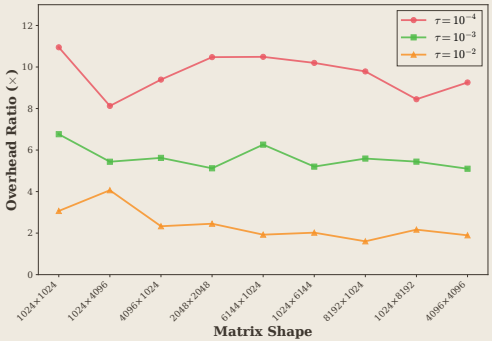
No weight decay on hidden 2D weights. One fewer hyperparameter to tune, good news for scaling law fitting.

The Geometry



Closing the Infra Cost Gap

Stage	Time (ms)	Δ vs prev.
Naive baseline (no opt.)	10 928	—
+ load balance & All-Gather	9 366	-14.3%
+ adaptive kernel & multi-stream	9 383	-8.8%
+ BF16 & torch.compile **(final)**	**7 666**	-18.3%



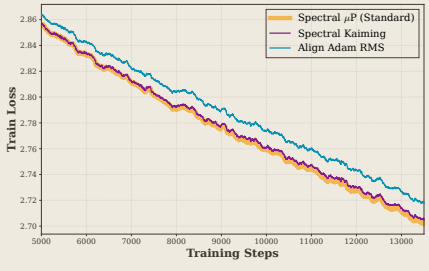
AdamW 6.73s · **Muon** 6.88s
MuonSphere 6.95s · **SSO** 7.67s

For tight infra budgets, MuonSphere (λ=0) keeps the activation control at near-zero overhead.

The Practical Recipe: LR Scaler

Unified view: each LR scaler keeps a consistent **effective step size** under a chosen **norm metric** and **initialization scheme**.

$$\frac{\|\Delta \boldsymbol{W}\|_{\mathcal{M}}}{\|\boldsymbol{W}\|_{\mathcal{M}}} = \frac{\|\eta R\, \boldsymbol{\Phi}\|_{\mathcal{M}}}{\|\boldsymbol{W}\|_{\mathcal{M}}} = \eta$$



Spectral-manifold optimization needs a scaler explicitly calibrated to the spectral norm

Spectral μP

$$R = \sqrt{d_{\text{out}}/d_{\text{in}}}$$

spectral norm; μP init

Align-Adam-RMS

$$R = 0.2\sqrt{\max(d_{\text{out}}, d_{\text{in}})}$$

l2 norm; standard init

Spectral Kaiming

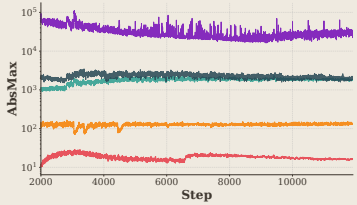
$$R = \sqrt{\max(1, d_{\text{out}}/d_{\text{in}})}$$

spectral norm; Kaiming init

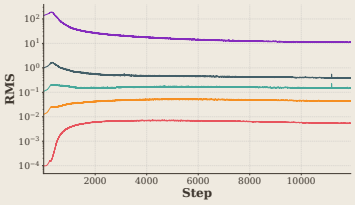
The Practical Recipe: Spectral Radius

$R = c\sqrt{d_{\text{out}}/d_{\text{in}}}$ c sets branch output magnitude relative to the residual stream.

optimization: moderate c wins, ~ 2.0 is best



FFN AbsMax: follows the radius scale

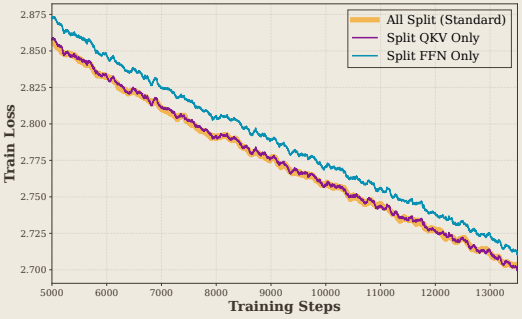


FFN RMS: clear power-law response to c

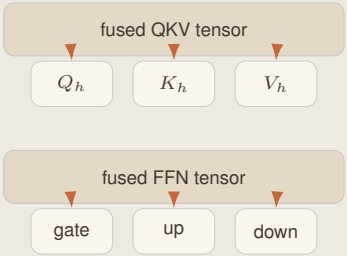
Tune c to control the signal-to-noise ratio along the deep residual path: attention/FFN branches become **directly scale-controlled** instead of drifting freely.

The Practical Recipe: Module Granularity

Megatron may fuse QKV or SwiGLU matrices into one tensor, but those slices do not play one functional role. SSO treats each slice as an **atomic function unit**.



splitting QKV per head gives the dominant gain



Best granularity: split attention QKV per head; separate FFN gate/up.

Empirical Evidence

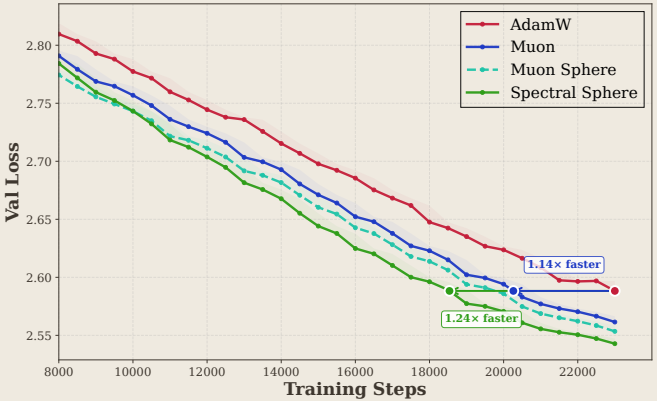
100B-token pre-training on OLMo-Mix-1124

deterministic data order, global batch ∼4M tokens, BF16 training

Scaling	Dense	MoE	DeepNet
70M → 1.8B	1.8B	8B-A1B	200 layers
LR transfer	speed	routing	depth

We test the optimizer where instability usually appears: width scaling, activation magnitude, MoE routing, and very deep residual paths.

Dense 1.8B: Faster Loss Descent



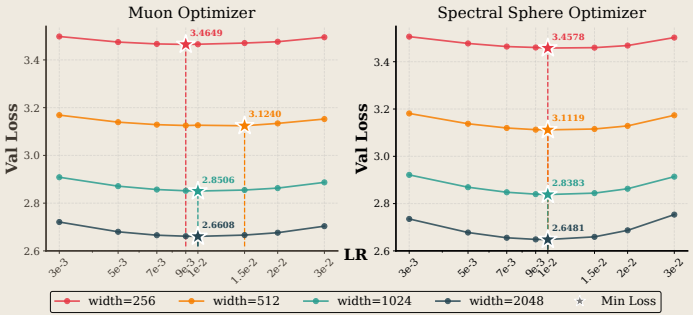
validation loss on 100B-token Dense 1.7B training

Reference point. AdamW reaches validation loss 2.588 at 23k steps.

Same loss sooner. Muon: 12% fewer steps; **SSO**: 19% fewer steps.

SSO gives a $1.24\times$ speed-up over AdamW, even in a setup tuned for AdamW.

μP Width Transfer



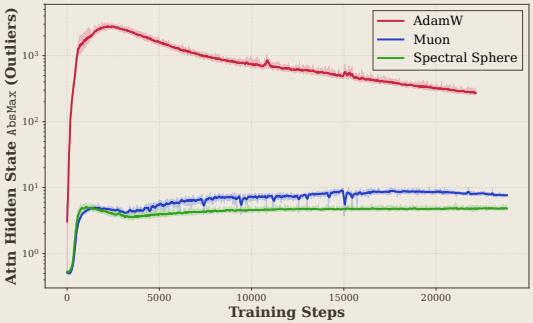
learning-rate sweep across $25\times$ model sizes: 70M $\rightarrow$ 1.8B

Target behavior. μP should make the optimal LR transfer across width.

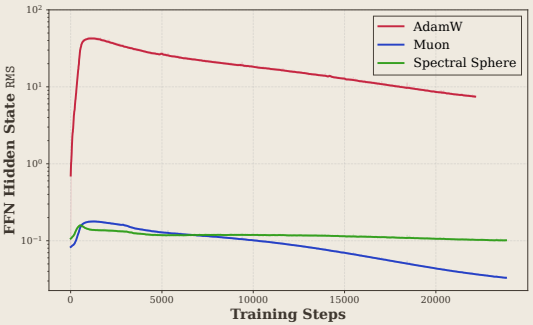
Muon is not enough. It controls updates, but its optimal LR still drifts with width.

SSO transfers. Stable optimal LR, with lower optimal loss than **Muon**.

Hidden Activations: Bounded Dynamics



attention AbsMax: outlier magnitude

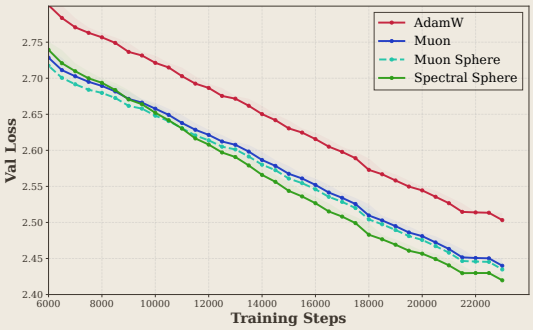


FFN RMS: activation scale

SSO keeps activation RMS at $\Theta(1)$ scale throughout training.

AdamW reaches $\sim 100\times$ larger activation magnitudes; **Muon** still drifts mildly.

MoE 8B-A1B: Better Router Balance

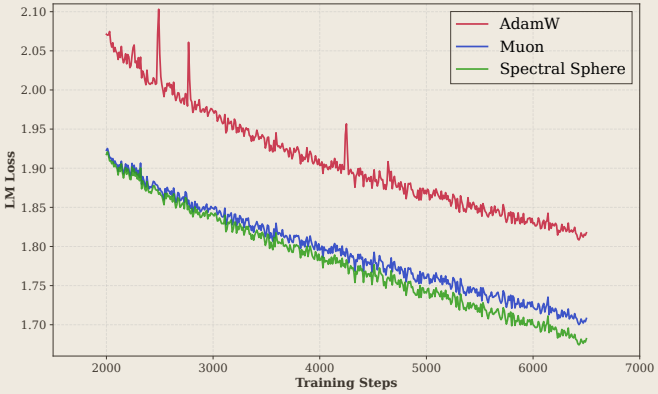


MoE validation loss

router MaxVio; lower means better load balance

Spectral normalization promotes balanced expert utilization: **SSO** has the lowest validation loss and the smallest routing violation.

DeepNet 200L: Depth Stress Test



training loss after extending the 28-layer baseline to 200 layers

Why this test. Very deep residual paths expose slow activation drift and loss spikes.

Spectral methods. Lower loss and smoother dynamics than AdamW.

SSO. Best stability and lowest loss under the 200-layer stress test.

Why Spectral Sphere?

Stiefel	Hyperball	Spectral sphere
all $\sigma_i = 1$	$\|\boldsymbol{W}\|_F \leq C$	$\sigma_{\max} = R$
too rigid	**too loose**	**a balance?**
no internal spectrum	one σ can dominate	worst-case bound, free interior

SSO bounds only $\sigma_{\max}$ — the worst-case activation bound μP needs — and lets the interior spectrum evolve freely.

Take-Away

SSO Steepest Descent on the Spectral Sphere
= Muon + Tangent Correction + Weight Retraction.

✓ **Speed:** $1.24\times$ **AdamW on Dense 1.7B**

✓ **Stability: bounded activations**

✓ **Practicality: drop-in for Megatron**

unakar666@gmail.com · github.com/Unakar/Spectral-Sphere-Optimizer

Acknowledgements

Many ideas in this talk were inspired by Jianlin Su's writing at `kexue.fm`.

Recommended reading from `kexue.fm`

1. Steepest Descent on Manifolds: 1. SGD + Hypersphere
   `kexue.fm/archives/11196`
2. Steepest Descent on Manifolds: 2. Muon + Orthogonality
   `kexue.fm/archives/11215`
3. Steepest Descent on Manifolds: 3. Muon + Stiefel
   `kexue.fm/archives/11221`
4. Steepest Descent on Manifolds: 4. Muon + Spectral Sphere
   `kexue.fm/archives/11241`
5. Beyond MuP: 4. Holding the Line on Parameter Stability
   `kexue.fm/archives/11729`

You think you're smarter than Su-shen?

kexue.fm is all you need.

Q&A

Any questions?

GitHub

WeChat