Delving into Muon and Beyond

Deep Analysis and Extensions

Xianbiao Qi*, Marco Chen*, Jiaquan Ye, Yelin He, Rong Xiao

ICML 2026 Spotlight

Problem: Isolating the Effect of Muon

Adam: element normalization

$$\Delta \boldsymbol{W}_t = \boldsymbol{M}_t \oslash \sqrt{\boldsymbol{V}_t}$$

Adam controls update scale **coordinate by coordinate** using second-moment statistics.

Muon: matrix-level normalization

$$\boldsymbol{M}_t = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \quad \rightarrow \quad \boldsymbol{\Delta W} = \boldsymbol{U}\boldsymbol{V}^\top$$

Muon keeps the singular directions, but **sets all nonzero singular values to one**.

Problem: Existing comparisons often include QK-Norm, QK-Clip, weight decay, or optimizer-specific learning-rate treatment.

Question: Which aspects of Muon drive its behavior, and when are these effects beneficial?

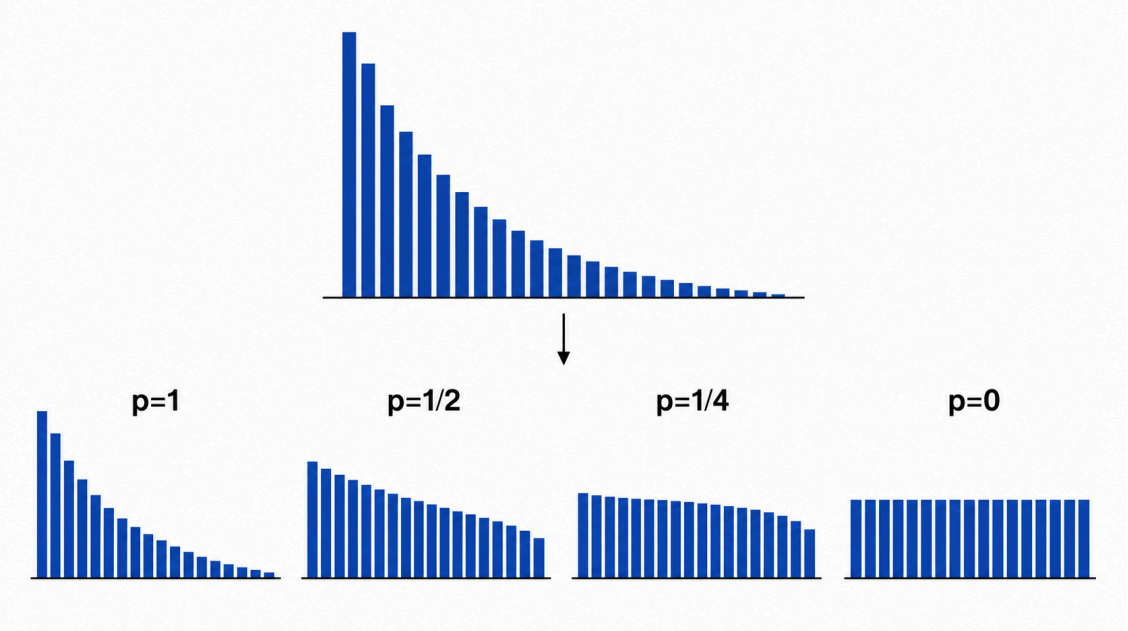
Muon as one endpoint

Given a matrix update $O_t = U\Sigma V^\top$, define

$$\Psi_p(O_t) = U\Sigma^p V^\top$$

We study

$$p \in \left\{1, \frac{1}{2}, \frac{1}{4}, 0\right\}$$



Lower p represents stronger **spectral compression**.

Spectral family with two optimizer inputs

Momentum input

$$O_t^{\mathrm{mom}} = M_t$$

mSGD

$$\Delta W_t = \Psi_1(O_t^{\mathrm{mom}})$$

mSGDS

$$\Delta W_t = \Psi_{\frac{1}{2}}(O_t^{\mathrm{mom}})$$

mSGDQ

$$\Delta W_t = \Psi_{\frac{1}{4}}(O_t^{\mathrm{mom}})$$

mSGDZ / Muon

$$\Delta W_t = \Psi_0(O_t^{\mathrm{mom}})$$

RMS-normalized input

$$O_t^{\mathrm{rms}} = M_t \oslash \sqrt{V_t}$$

Adam

$$\Delta W_t = \Psi_1(O_t^{\mathrm{rms}})$$

AdamS

$$\Delta W_t = \Psi_{\frac{1}{2}}(O_t^{\mathrm{rms}})$$

AdamQ

$$\Delta W_t = \Psi_{\frac{1}{4}}(O_t^{\mathrm{rms}})$$

AdamZ

$$\Delta W_t = \Psi_0(O_t^{\mathrm{rms}})$$

prefix chooses input update; **suffix** chooses the spectral exponent: none $= 1$, S $= \frac{1}{2}$, Q $= \frac{1}{4}$, Z $= 0$.

Efficient computation: avoiding explicit SVD

Problem: Direct SVD is expensive for large matrix parameters and hard to parallelize.

Solution: Coupled Newton-Schulz iterations to compute spectral transforms via matmuls.

1. Rewrite spectral transforms

$$\Psi_{\frac{1}{2}}(\boldsymbol{O}_t) = \boldsymbol{U}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{V}^\top = \boldsymbol{O}_t(\boldsymbol{O}_t^\top \boldsymbol{O}_t)^{-\frac{1}{4}}.$$

so it suffices to compute the inverse fourth root of $\boldsymbol{X} = \boldsymbol{O}_t^\top \boldsymbol{O}_t$

2. Coupled Newton-Schulz

From $\boldsymbol{Y}_0 = \boldsymbol{X}$ and $\boldsymbol{Z}_0 = \boldsymbol{I}$, the iteration below converges to $\boldsymbol{X}^{\frac{1}{2}}$ and $\boldsymbol{X}^{-\frac{1}{2}}$

$$\boldsymbol{Y}_{k+1} = \frac{1}{2}\boldsymbol{Y}_k(3\boldsymbol{I} - \boldsymbol{Z}_k\boldsymbol{Y}_k)$$

$$\boldsymbol{Z}_{k+1} = \frac{1}{2}(3\boldsymbol{I} - \boldsymbol{Z}_k\boldsymbol{Y}_k)\boldsymbol{Z}_k$$

Applying the procedure twice gives the fourth-root terms needed for $p = \frac{1}{2}$.

In our setting, $K = 5$ adds about 3–4% wall-clock overhead.

Controlled experimental setup

Setup

`nanoGPT` on OpenWebText, 124M-parameter GPT-2 style model.

Controls

- Matrix and vector learning rates decoupled for all optimizers.
- All vector params: Adam with `lr = 3e-4`
- No QK-Norm or QK-Clip.
- No weight decay in primary experiments.

Comparison

Spectral compression interpolation:

$$U\Sigma^{1}V^{\top} \to U\Sigma^{\frac{1}{2}}V^{\top} \to U\Sigma^{\frac{1}{4}}V^{\top} \to U\Sigma^{0}V^{\top}$$

across two input types:

$$O_t^{\text{mom}} = M_t \quad \text{and} \quad O_t^{\text{rms}} = M_t \oslash \sqrt{V_t}$$

yields the eight optimizers we study. Each matrix optimizer is **tuned separately** over learning rate.

Momentum input: spectral compression stabilizes mSGD

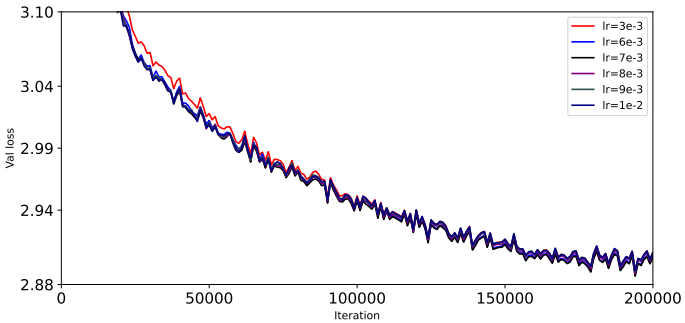
Momentum Input

$$O_t^{\text{mom}} = M_t$$

Full flattening provides **clear stabilization effects**.

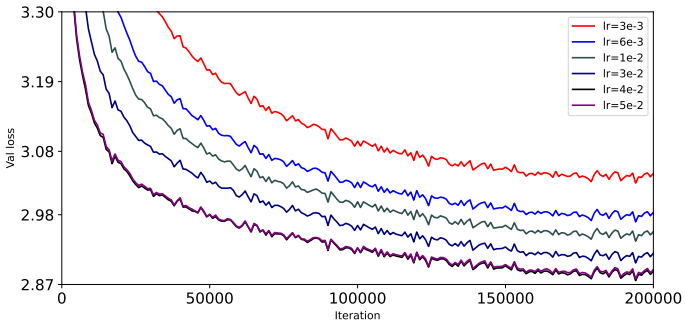
Partial compression can be competitive once training is stable.

mSGDZ / Muon, $p = 0$



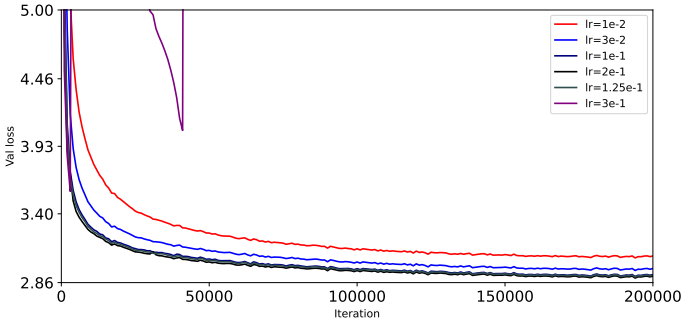
lr=3e-3: loss = 2.891, lr=6e-3: loss = 2.888,
lr=7e-3: loss = 2.887, lr=8e-3: loss = 2.889,
lr=9e-3: loss = 2.891, lr=1e-2: loss = 2.891

mSGDQ, $p = \frac{1}{4}$



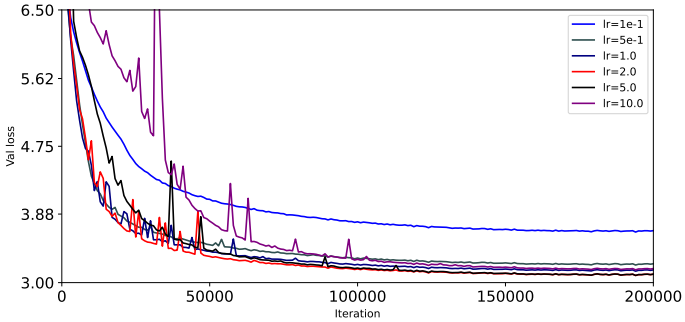
lr=3e-3: loss = 3.030, lr=6e-3: loss = 2.969,
lr=1e-2: loss = 2.938, lr=3e-2: loss = 2.904,
lr=4e-2: loss = 2.876, lr=5e-2: loss = 2.879

mSGDS, $p = \frac{1}{2}$



lr=1e-2: loss = 3.054, lr=3e-2: loss = 2.955,
lr=1e-1: loss = 2.908, **lr=2e-1: loss = 2.895,**
lr=1.25e-1: loss = 2.904, lr=3e-1: loss = inf

mSGD, $p = 1$



lr=1e-1: loss = 3.649, lr=5e-1: loss = 3.226,
lr=1.0: loss = 3.147, lr=2.0: loss = 3.092,
lr=5.0: loss = 3.092, lr=10.0: loss = 3.164

RMS-normalized input: smaller gains from compression

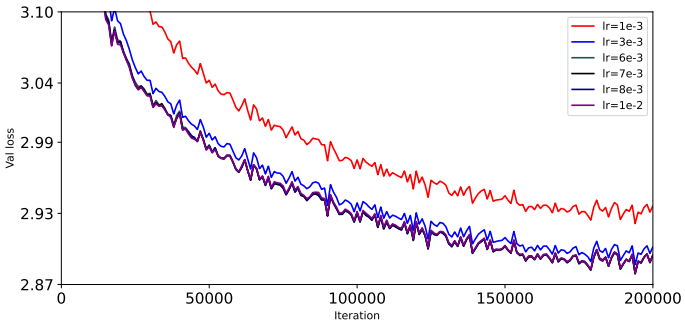
RMS-normalized Input

$$O_t^{\mathrm{rms}} = M_t \oslash \sqrt{V_t}$$

RMS normalization already controls coordinate scale.

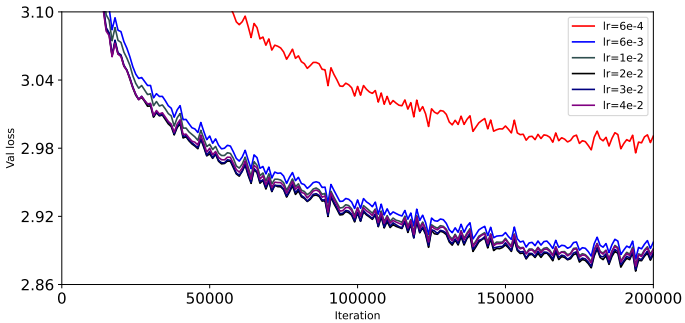
Benefit of **full spectral flattening** is more **limited**.

AdamZ, $p = 0$



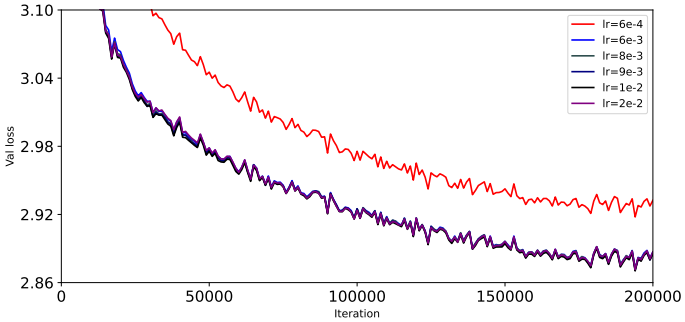
lr=1e-3: loss = 2.921, lr=3e-3: loss = 2.887,
lr=6e-3: loss = 2.880, **lr=7e-3: loss = 2.879,**
lr=8e-3: loss = 2.880, lr=1e-2: loss = 2.880

AdamQ, $p = \frac{1}{4}$



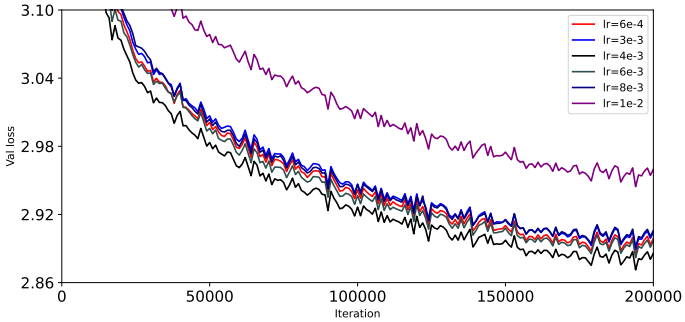
lr=6e-4: loss = 2.976, lr=6e-3: loss = 2.882,
lr=1e-2: loss = 2.877, **lr=2e-2: loss = 2.872,**
lr=3e-2: loss = 2.873, lr=4e-2: loss = 2.876

AdamS, $p = \frac{1}{2}$



lr=6e-4: loss = 2.918, lr=6e-3: loss = 2.873,
lr=8e-3: loss = 2.872, lr=9e-3: loss = 2.871,
lr=1e-2: loss = 2.870, lr=2e-2: loss = 2.871

Adam, $p = 1$



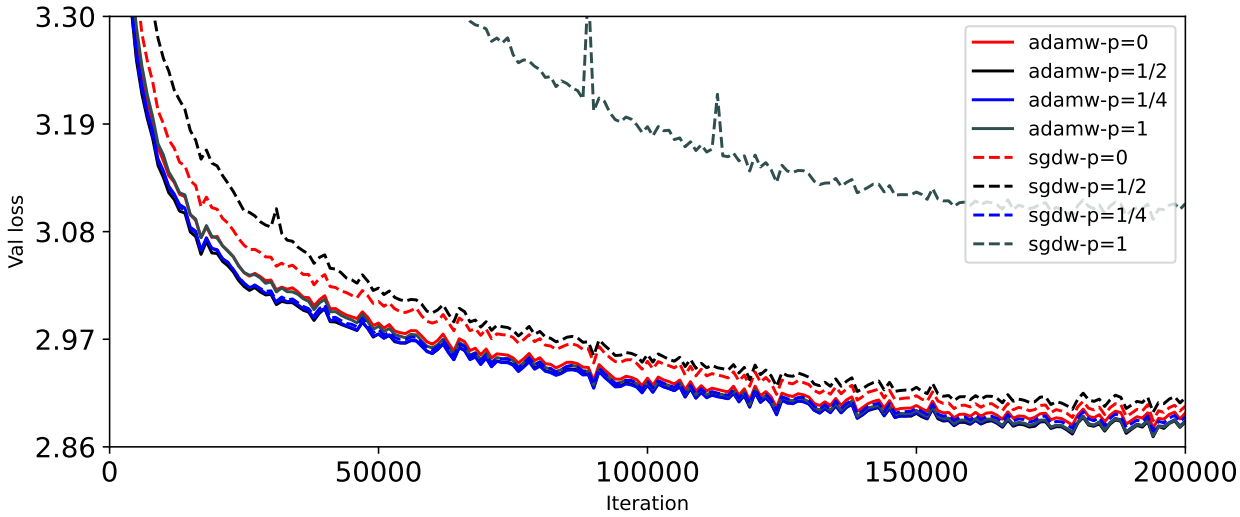
lr=6e-4: loss = 2.884, lr=3e-3: loss = 2.889,
lr=4e-3: loss = 2.871, lr=6e-3: loss = 2.882,
lr=8e-3: loss = 2.891, lr=1e-2: loss = 2.945

Results: spectral compression helps most for raw momentum

Best tuned runs

1. **Muon stabilizes mSGD.**
2. **Partial compression can be stronger.**
3. **RMS normalization reduces the margin between spectral variants.**

In general, the RMS-normalized variants outperform their momentum counterparts.

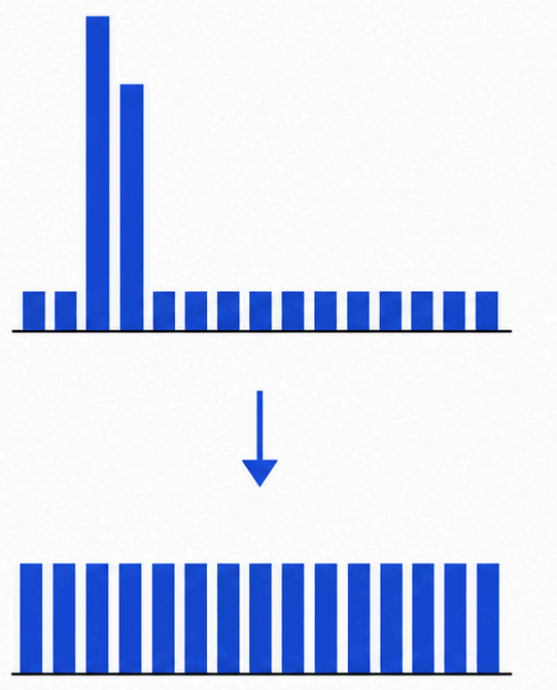


adamw-p=0: loss = 2.879,
adamw-p=1/4: loss = 2.872,
sgdw-p=0: loss = 2.887,
sgdw-p=1/4: loss = 2.876,

adamw-p=1/2: loss = 2.870,
adamw-p=1: loss = 2.871,
sgdw-p=1/2: loss = 2.895,
sgdw-p=1: loss = 3.092

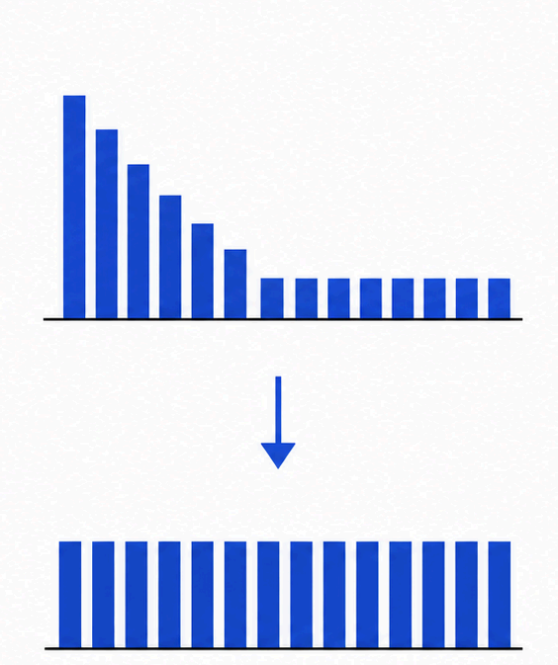
Our understanding of Muon

Dominant directions controlled



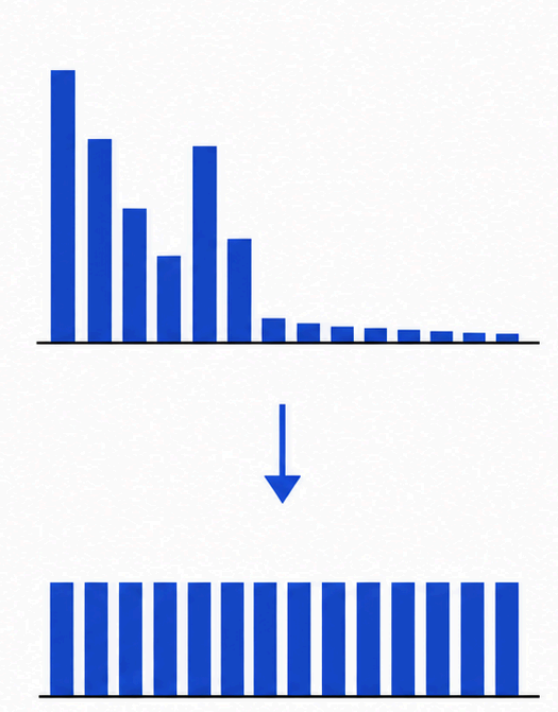
Stabilizes by reducing dominant singular directions.

Useful anisotropy suppressed



Can reduce the influence of meaningful large directions.

Noisy directions amplified



Can upweight small singular directions dominated by noise.

Muon as spectral compression

$$\Psi_p(O) = U\Sigma^p V^\top$$

Conclusions

Spectral compression is useful, especially for first-moment momentum updates.

Full orthogonalization is not always the best choice, and it does not consistently outperform Adam-style RMS normalization in our setting.

Muon should be viewed as one point in a larger collection of **spectral optimizers** that can be applied to **different inputs**.