Beyond MuP: 3. Special Cases Require Special Treatment

Jianlin Su

2026-03-02

Abstract

This article explores why certain components like Embedding layers and LM Heads require special treatment in optimization, despite being matrix-valued like other layers. Using the three stability metrics introduced in earlier articles, we analyze the initialization and steepest descent directions for these special cases and show why they differ from the Muon optimizer's approach.

1 Recap

In the first article *Beyond MuP: 1. Three Features of a Good Model*, we introduced three stability metrics:

$$\text{Forward Stability:} \quad \max_{\boldsymbol{x}} \|\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\omega})\|_{\text{RMS}} = \Theta(1) \tag{1}$$

$$\text{Dependency Stability:} \quad \max_{\boldsymbol{x}_1,\boldsymbol{x}_2} \|\boldsymbol{f}(\boldsymbol{x}_1;\boldsymbol{\omega}) - \boldsymbol{f}(\boldsymbol{x}_2;\boldsymbol{\omega})\|_{\text{RMS}} = \Theta(1) \tag{2}$$

$$\text{Update Stability:} \quad \max_{\boldsymbol{x}} \|\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\omega}+\Delta\boldsymbol{\omega}) - \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\omega})\|_{\text{RMS}} = \Theta(1) \tag{3}$$

These metrics share a common format: compute the RMS of the output, then take the maximum over inputs. Here, $\boldsymbol{x}$ is the input, $\boldsymbol{\omega}$ the parameters, and $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\omega})$ can represent a layer, block, or the entire model depending on our ability to compute the maximum.

In the previous article *Beyond MuP: 2. Linear Layers and Steepest Descent*, we showed how to compute these metrics for linear layers by adding In Norms, and derived the Muon optimizer using the steepest descent principle.

The key contribution of the "Beyond MuP" series is not the steepest descent itself, but the identification of suitable stability metrics for arbitrary layers.

2 Embedding Layers

Consider an Embedding layer: input is an index i, output is the corresponding vector $\boldsymbol{f}(i;\boldsymbol{E}) = \boldsymbol{E}_i$, where $\boldsymbol{E}$ is a $|V| \times d$ matrix and $\boldsymbol{E}_i \triangleq \boldsymbol{E}_{i,:}$ is the i-th row.

Its stability metrics are:

$$\text{Forward Stability:} \quad \max_{i} \|\boldsymbol{E}_i\|_{\text{RMS}} = \Theta(1) \tag{4}$$

$$\text{Dependency Stability:} \quad \max_{i,j} \|\boldsymbol{E}_i - \boldsymbol{E}_j\|_{\text{RMS}} = \Theta(1) \tag{5}$$

$$\text{Update Stability:} \quad \max_{i} \|\Delta\boldsymbol{E}_i\|_{\text{RMS}} = \Theta(1) \tag{6}$$

Since $\max_{i,j} \|\boldsymbol{E}_i - \boldsymbol{E}_j\|_{\text{RMS}} \leq 2\max_i \|\boldsymbol{E}_i\|_{\text{RMS}}$, these essentially measure the max row norm (scaled by $1/\sqrt{d}$) of $\boldsymbol{E}$ or $\Delta\boldsymbol{E}$.

To find the steepest descent direction for Embedding layers, we solve:

$$\min_{\Delta\boldsymbol{E}} \langle \boldsymbol{G}, \Delta\boldsymbol{E}\rangle \qquad \text{s.t.} \qquad \max_i \underbrace{\|\Delta\boldsymbol{E}_i\|_{\text{RMS}}}_{\|\Delta\boldsymbol{E}_i\|_2/\sqrt{d}} \leq \eta \tag{7}$$

Using Cauchy-Schwarz inequality, we derive:

$$\langle \boldsymbol{G}, \Delta\boldsymbol{E}\rangle = \sum_{i=1}^{|V|} \langle \boldsymbol{G}_i, \Delta\boldsymbol{E}_i\rangle \geq -\sum_{i=1}^{|V|} \|\boldsymbol{G}_i\|_2 \times \|\Delta\boldsymbol{E}_i\|_2 \geq -\eta\sqrt{d}\sum_{i=1}^{|V|} \|\boldsymbol{G}_i\|_2 \tag{8}$$

Equality holds when $\Delta\boldsymbol{E}_i = -\eta\boldsymbol{G}_i/\|\boldsymbol{G}_i\|_{\text{RMS}}$, meaning the appropriate steepest descent for Embedding layers is Row-wise RMS Normalized SGD.

3 LM Heads

Now consider LM Heads. Despite appearing as linear layers (input $\boldsymbol{x} \in \mathbb{R}^d$, weight $\boldsymbol{W} \in \mathbb{R}^{d\times|V|}$, output $\boldsymbol{x}\boldsymbol{W} \in \mathbb{R}^{|V|}$), they are not suitable for Muon.

3.1 Responsibility to Loss

The reason is that LM Heads are responsible for computing the loss during training:

$$\ell(\boldsymbol{x}, t; \boldsymbol{W}) = \log\sum_{i=1}^{|V|} e^{\langle \boldsymbol{x}, \boldsymbol{\omega}_i\rangle} - \langle \boldsymbol{x}, \boldsymbol{\omega}_t\rangle = \log\sum_{i=1}^{|V|} e^{\langle \boldsymbol{x}, \boldsymbol{\omega}_i - \boldsymbol{\omega}_t\rangle} \tag{9}$$

where $\boldsymbol{\omega}_i \triangleq \boldsymbol{W}_{:,i}$ is the i-th column of $\boldsymbol{W}$.

The three stability metrics cannot be computed exactly due to the complex nonlinearity, but we derive tight upper bounds.

3.2 Forward Stability

$$\begin{aligned}
\ell(\boldsymbol{x}, t; \boldsymbol{W}) = \log\sum_{i=1}^{|V|} e^{\langle \boldsymbol{x}, \boldsymbol{\omega}_i - \boldsymbol{\omega}_t\rangle} &\leq \log\left(|V| \max_i e^{\langle \boldsymbol{x}, \boldsymbol{\omega}_i - \boldsymbol{\omega}_t\rangle}\right) \\
&= \log|V| + \max_i \langle \boldsymbol{x}, \boldsymbol{\omega}_i - \boldsymbol{\omega}_t\rangle \\
&\leq \log|V| + \max_i \|\boldsymbol{x}\|_2 \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_t\|_2
\end{aligned} \tag{10}$$

$$\begin{aligned}
\text{Forward Stability:} \quad \max_{t, \|\boldsymbol{x}\|_{\text{RMS}}=1} \ell(\boldsymbol{x}, t; \boldsymbol{W}) &\leq \log|V| + d\max_{i,t} \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_t\|_{\text{RMS}} \\
&\leq \log|V| + 2d\max_i \|\boldsymbol{\omega}_i\|_{\text{RMS}}
\end{aligned} \tag{11}$$

To ensure $\Theta(1)$, the initialization variance should be $\Theta(1/d^2)$.

3.3 Key Inequality

We prove:

$$\left|\log\sum_{i=1}^{n} e^{a_i} - \log\sum_{i=1}^{n} e^{b_i}\right| \leq \max_i |a_i - b_i| \tag{12}$$

Using this and Cauchy-Schwarz, we derive bounds for the other metrics.

3.4 Dependency Stability

$$\begin{aligned}|\ell(\boldsymbol{x}_1, t_1; \boldsymbol{W}) - \ell(\boldsymbol{x}_2, t_2; \boldsymbol{W})| &\leq \max_i |\langle \boldsymbol{x}_1, \boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_1}\rangle - \langle \boldsymbol{x}_2, \boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_2}\rangle| \\ &\leq \max_i (\|\boldsymbol{x}_1\|_2 \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_1}\|_2 + \|\boldsymbol{x}_2\|_2 \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_2}\|_2) \\ &= d \max_i (\|\boldsymbol{x}_1\|_{\text{RMS}} \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_1}\|_{\text{RMS}} + \|\boldsymbol{x}_2\|_{\text{RMS}} \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_2}\|_{\text{RMS}})\end{aligned} \tag{13}$$

$$\begin{aligned}\text{Dependency Stability:} \quad \max_{\substack{t_1, t_2, \\ \|\boldsymbol{x}_1\|_{\text{RMS}} = 1 \\ \|\boldsymbol{x}_2\|_{\text{RMS}} = 1}} |\ell(\boldsymbol{x}_1, t_1; \boldsymbol{W}) - \ell(\boldsymbol{x}_2, t_2; \boldsymbol{W})| &\leq d \max_{i, t_1, t_2} (\|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_1}\|_{\text{RMS}} + \|\boldsymbol{\omega}_i - \boldsymbol{\omega}_{t_2}\|_{\text{RMS}}) \\ &\leq 4d \max_i \|\boldsymbol{\omega}_i\|_{\text{RMS}}\end{aligned} \tag{14}$$

3.5 Update Stability

$$\begin{aligned}|\ell(\boldsymbol{x}, t; \boldsymbol{W} + \Delta\boldsymbol{W}) - \ell(\boldsymbol{x}, t; \boldsymbol{W})| &\leq \max_i |\langle \boldsymbol{x}, \boldsymbol{\omega}_i + \Delta\boldsymbol{\omega}_i - \boldsymbol{\omega}_t - \Delta\boldsymbol{\omega}_t\rangle - \langle \boldsymbol{x}, \boldsymbol{\omega}_i - \boldsymbol{\omega}_t\rangle| \\ &= \max_i |\langle \boldsymbol{x}, \Delta\boldsymbol{\omega}_i - \Delta\boldsymbol{\omega}_t\rangle| \\ &\leq d \max_i \|\boldsymbol{x}\|_{\text{RMS}} \|\Delta\boldsymbol{\omega}_i - \Delta\boldsymbol{\omega}_t\|_{\text{RMS}}\end{aligned} \tag{15}$$

$$\begin{aligned}\text{Update Stability:} \quad \max_{t, \|\boldsymbol{x}\|_{\text{RMS}}=1} |\ell(\boldsymbol{x}, t; \boldsymbol{W} + \Delta\boldsymbol{W}) - \ell(\boldsymbol{x}_2, t_2; \boldsymbol{W})| &\leq d \max_{i,t} \|\Delta\boldsymbol{\omega}_i - \Delta\boldsymbol{\omega}_t\|_{\text{RMS}} \\ &\leq 2d \max_i \|\Delta\boldsymbol{\omega}_i\|_{\text{RMS}}\end{aligned} \tag{16}$$

The steepest descent for LM Heads is also Normalized SGD, but applied column-wise. Initialization standard deviation and learning rate scale as $\Theta(1/d)$, unlike the $\Theta(1)$ scaling for Embedding layers.

4 Other Modules

4.1 Hadamard Product

After RMS Norm, we often multiply by a $\boldsymbol{\gamma}$ vector: $(\boldsymbol{x}/\|\boldsymbol{x}\|_{\text{RMS}}) \odot \boldsymbol{\gamma}$. This is equivalent to matrix multiplication with a diagonal matrix $\text{diag}(\boldsymbol{\gamma})$, which can be analyzed as a special linear layer.

Initialization should be identity matrix (i.e., $\boldsymbol{\gamma}$ initialized to all ones), and the optimizer becomes SignSGD.

4.2 Linear Bias Terms

For linear layers with bias $\boldsymbol{b}$, the stability metrics become:

$$\text{Forward Stability:} \quad \max_{\|\boldsymbol{x}\|_{\text{RMS}}=1} \|\boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}\|_{\text{RMS}} \tag{17}$$

$$\text{Dependency Stability:} \quad \max_{\|\boldsymbol{x}_1\|_{\text{RMS}}=\|\boldsymbol{x}_2\|_{\text{RMS}}=1} \|\boldsymbol{x}_1\boldsymbol{W} - \boldsymbol{x}_2\boldsymbol{W}\|_{\text{RMS}} \tag{18}$$

$$\text{Update Stability:} \quad \max_{\|\boldsymbol{x}\|_{\text{RMS}}=1} \|\boldsymbol{x}\Delta\boldsymbol{W} + \Delta\boldsymbol{b}\|_{\text{RMS}} \tag{19}$$

Bias is typically initialized to zero and optimized using Normalized SGD.

4.3 Attention Scaling

For attention mechanisms with $\boldsymbol{q} = \boldsymbol{x}\boldsymbol{W}_q, \boldsymbol{k} = \boldsymbol{x}\boldsymbol{W}_k$, we have:

$$|\langle \boldsymbol{q}, \boldsymbol{k} \rangle| \leq \|\boldsymbol{q}\|_2 \|\boldsymbol{k}\|_2 = d \|\boldsymbol{q}\|_{\text{RMS}} \|\boldsymbol{k}\|_{\text{RMS}} \tag{20}$$

This suggests a scaling factor of $\Theta(1/d)$, which complements the common $1/\sqrt{d}$ initialization scaling.

5 Summary

The key results are summarized below:

	Input	Parameters	Output	Initial Variance	Steepest Descent
Linear	$\boldsymbol{x}$	$\boldsymbol{W} \in \mathbb{R}^{d_{in} \times d_{out}}$ $\boldsymbol{b} \in \mathbb{R}^{d_{out}}$	$\boldsymbol{x}\boldsymbol{W} + \boldsymbol{b}$	$\boldsymbol{W}$: $\sqrt{\frac{d_{out}}{d_{in}}} \frac{1}{\sqrt{d_{in}} + \sqrt{d_{out}}}$ $\boldsymbol{b}$: 0	$\Delta \boldsymbol{W} = -\eta \sqrt{\frac{d_{out}}{d_{in}}}\, \text{msign}(\boldsymbol{G})$ $\Delta \boldsymbol{b} = -\eta \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_{\text{RMS}}}$
Embedding	i	$\boldsymbol{E} \in \mathbb{R}^{\lvert V \rvert \times d}$	$\boldsymbol{E}_{i,:}$	1	$\Delta \boldsymbol{E}_{i,:} = -\eta \frac{\boldsymbol{G}_{i,:}}{\|\boldsymbol{G}_{i,:}\|_{\text{RMS}}}$
LM Head	$\boldsymbol{x}, t$	$\boldsymbol{W} \in \mathbb{R}^{d \times \lvert V \rvert}$	$\log \sum_{i=1}^{\lvert V \rvert} e^{\langle \boldsymbol{x}, \boldsymbol{W}_{:,i} - \boldsymbol{W}_{:,t} \rangle}$	$\frac{1}{d^2}$	$\Delta \boldsymbol{W}_{:,i} = -\frac{\eta}{d} \frac{\boldsymbol{G}_{:,i}}{\|\boldsymbol{G}_{:,i}\|_{\text{RMS}}}$
RMS Norm	$\boldsymbol{x}$	$\boldsymbol{\gamma} \in \mathbb{R}^d$	$\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_{\text{RMS}}} \odot \boldsymbol{\gamma}$	1	$\Delta \boldsymbol{\gamma} = -\eta\, \text{sign}(\boldsymbol{g})$

This shows that stability metrics provide a unified framework for analyzing and optimizing different types of layers, and explain why certain components require special treatment.