On the Width Scaling of Neural Optimizers: A Matrix Operator Norm Perspective

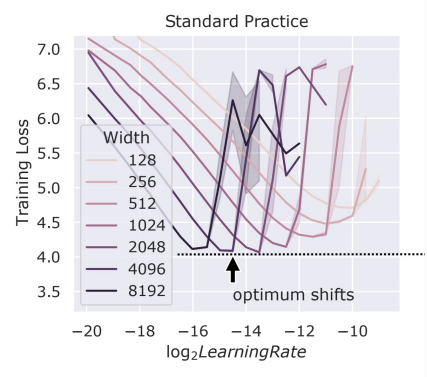
Yiping Lu

Northwestern University

FAI Seminar, March, 2026

Joint work with Ruihan Xu (U Chicago) and Jiajin Li (UBC).

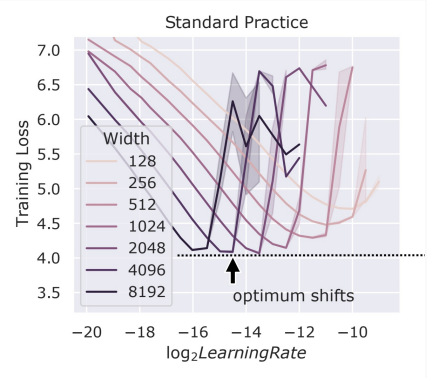
Hyperparameter Transfer [Yang et al, 2021, NeurIPS]



Key empirical observations

- The optimal learning rate **decreases as network width increases**.
- A learning rate tuned at width **512** can **diverge** or **slow dramatically** when scaled to width **2048**.

Hyperparameter Transfer [Yang et al, 2021, NeurIPS]

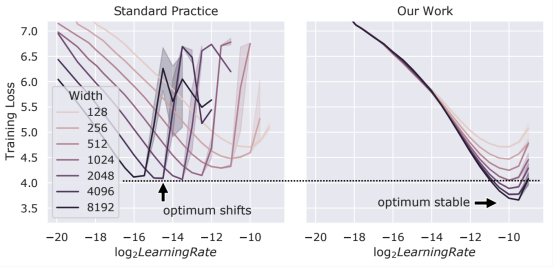


Why this matters for LLMs

Pre-training state-of-the-art base models costs $\gg$ **hundreds of millions USD**.

Hyperparameter transfer is therefore **essential**, not optional.

Hyperparameter Transfer [Yang et al, 2021, NeurIPS]



[Yang et al, 2021, NeurIPS]

`Adam` learning rate scaling rule:

- **For Weight Matrix**:
  - Scale learning rate by $\frac{1}{\texttt{d\_in}}$
- **For Embedding Matrix**:
  - No learning-rate scaling

Here, `d_in` denotes the effective input dimensionality of the parameter block.

Derivation grounded in over 100 pages of random matrix theory, with theoretical support for transfer emerging only after a single gradient step near initialization.

What this talk is about

Design **width-independent neural optimizers** *and*

develop a principled understanding of **learning-rate scaling rules**

from a **matrix operator norm perspective**.

Problem Setup: Neural Network Optimization

We consider the constrained optimization problem

$$\min_{\boldsymbol{W}_{1:\ell},\, \boldsymbol{b}_{1:\ell} \in \mathcal{C}} \; f(\boldsymbol{W}_{1:\ell}, \boldsymbol{b}_{1:\ell}) \;\coloneqq\; \mathcal{L}\big(\boldsymbol{y}_\ell(\boldsymbol{x})\big),$$

where $\boldsymbol{y}_\ell(\boldsymbol{x}) \in \mathbb{R}$ is the output of an ℓ-layer feedforward neural network evaluated at input $\boldsymbol{x} \in \mathbb{R}^d$.

Network architecture

The network is defined recursively as

$$\boldsymbol{y}_i(\boldsymbol{x}) \;\coloneqq\; \sigma\big(\boldsymbol{W}_i\, \boldsymbol{y}_{i-1}(\boldsymbol{x}) + \boldsymbol{b}_i\big), \qquad i = 1, \dots, \ell,$$

with

$$\boldsymbol{W}_1 \in \mathbb{R}^{w \times d} \quad \text{and} \quad \boldsymbol{W}_i \in \mathbb{R}^{w \times w} \;\; (i \geq 2).$$

Later, we write $\Theta := \{\boldsymbol{W}_{1:\ell},\, \boldsymbol{b}_{1:\ell}\}$ for simplicity.

Problem Setup: Neural Network Optimization

We consider the constrained optimization problem

$$\min_{\boldsymbol{W}_{1:\ell},\, \boldsymbol{b}_{1:\ell} \in \mathcal{C}} \; f(\boldsymbol{W}_{1:\ell}, \boldsymbol{b}_{1:\ell}) \;:=\; \mathcal{L}\big(\boldsymbol{y}_\ell(\boldsymbol{x})\big),$$

where $\boldsymbol{y}_\ell(\boldsymbol{x}) \in \mathbb{R}$ is the output of an ℓ-layer feedforward neural network evaluated at input $\boldsymbol{x} \in \mathbb{R}^d$.

Network architecture

The network is defined recursively as

$$\boldsymbol{y}_i(\boldsymbol{x}) \;:=\; \sigma\big(\boldsymbol{W}_i\, \boldsymbol{y}_{i-1}(\boldsymbol{x}) + \boldsymbol{b}_i\big), \qquad i = 1, \ldots, \ell,$$

with

$$\boldsymbol{W}_1 \in \mathbb{R}^{w \times d} \quad \text{and} \quad \boldsymbol{W}_i \in \mathbb{R}^{w \times w} \;\; (i \geq 2).$$

Later, we write $\boldsymbol{\Theta} := \{\boldsymbol{W}_{1:\ell},\, \boldsymbol{b}_{1:\ell}\}$ for simplicity.

Decoupled Weight Decay as a Constrained Optimization Problem [Loshchilov and Hutter, 2019]

Let $\boldsymbol{d}^t$ denote the negative update direction at iteration t. We consider updates of the form

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta \boldsymbol{d}^{t-1} - \underbrace{\eta\lambda\boldsymbol{\theta}^{t-1}}_{\text{Weight Decay}},$$

where η is the learning rate and $\lambda \in (0, 1)$ is the weight decay coefficient.

Fact

If $\|\boldsymbol{d}^t\| \leq \lambda R$ and $\|\boldsymbol{\theta}^0\| \leq R$, we have $\|\boldsymbol{\theta}^t\| \leq R$ for all $t \geq 0$.

View decoupled weight decay as enforcing an implicit norm ball constraint !

Decoupled Weight Decay as a Constrained Optimization Problem [Loshchilov and Hutter, 2019]

Let $\boldsymbol{d}^t$ denote the negative update direction at iteration t. We consider updates of the form

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta \boldsymbol{d}^{t-1} - \underbrace{\eta\lambda\boldsymbol{\theta}^{t-1}}_{\text{Weight Decay}},$$

where η is the learning rate and $\lambda \in (0, 1)$ is the weight decay coefficient.

Fact

If $\|\boldsymbol{d}^t\| \leq \lambda R$ and $\|\boldsymbol{\theta}^0\| \leq R$, we have $\|\boldsymbol{\theta}^t\| \leq R$ for all $t \geq 0$.

View decoupled weight decay as enforcing an implicit norm ball constraint !

Decoupled Weight Decay as a Constrained Optimization Problem [Loshchilov and Hutter, 2019]

Let $\boldsymbol{d}^t$ denote the negative update direction at iteration t. We consider updates of the form

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta \boldsymbol{d}^{t-1} - \underbrace{\eta\lambda\boldsymbol{\theta}^{t-1}}_{\text{Weight Decay}},$$

where η is the learning rate and $\lambda \in (0, 1)$ is the weight decay coefficient.

Fact

If $\|\boldsymbol{d}^t\| \leq \lambda R$ and $\|\boldsymbol{\theta}^0\| \leq R$, we have $\|\boldsymbol{\theta}^t\| \leq R$ for all $t \geq 0$.

View decoupled weight decay as enforcing an implicit norm ball constraint !

A Brief Review of Successful Optimizers — AdamW, signSGD

- AdamW [Kingma and Ba, 2014] was regarded as the gold standard optimizer for large-scale neural network training.

$$
\begin{aligned}
\boldsymbol{g}^t &= \nabla f(\boldsymbol{\theta}^{t-1}, \xi_t) \quad \text{(compute gradient)} \\
\boldsymbol{m}^t &= \beta_1 \boldsymbol{m}^{t-1} + (1-\beta_1)\boldsymbol{g}^t \quad \text{(estimate first order moment)} \\
\boldsymbol{v}^t &= \beta_2 \boldsymbol{v}^{t-1} + (1-\beta_2)\boldsymbol{g}^t \odot \boldsymbol{g}^t \quad \text{(estimate second order moment)} \\
\hat{\boldsymbol{m}}^t &= \frac{\boldsymbol{m}^t}{1-\beta_1}, \hat{\boldsymbol{v}}^t = \frac{\boldsymbol{v}^t}{1-\beta_2} \quad \text{(bias correction)} \\
\boldsymbol{\theta}^t &= \boldsymbol{\theta}^{t-1} - \eta \left(\frac{\hat{\boldsymbol{m}}^t}{\sqrt{\hat{\boldsymbol{v}}^t} + \epsilon} \right) - \eta\lambda\boldsymbol{\theta}^{t-1}
\end{aligned}
$$

- Adam approximate `signSGD` [Bernstein et al, 2018, ICML].

$$
\boldsymbol{\theta}^t \approx \boldsymbol{\theta}^{t-1} - \eta \frac{\boldsymbol{g}^t}{\sqrt{\boldsymbol{g}^t \odot \boldsymbol{g}^t}} \approx \boldsymbol{\theta}^{t-1} - \eta \cdot \operatorname{sign}(\boldsymbol{g}^t).
$$

A Brief Review of Successful Optimizers — Lion [Chen et al, 2023]

Key idea

Lion can be viewed as **SignSGD + momentum + decoupled weight decay**.

$$
\begin{aligned}
\boldsymbol{g}^t &= \nabla f(\boldsymbol{\theta}^{t-1}, \xi_t) \\
\boldsymbol{m}^t &= \beta_1 \boldsymbol{m}^{t-1} + (1-\beta_1)\boldsymbol{g}^t \\
\boldsymbol{\theta}^t &= \boldsymbol{\theta}^{t-1} - \eta \mathrm{sign}(\boldsymbol{m}^t) - \eta\lambda\boldsymbol{\theta}^{t-1}
\end{aligned}
$$

- Automatic discovering optimizers through RL/meta-learning
- Less memory compared with AdamW

A Brief Review of Successful Optimizers — Muon

Key idea

Change the optimization geometry from a **vector space** to a **matrix space**.

$$
\begin{aligned}
\boldsymbol{G}^t &= \nabla f(\boldsymbol{\Theta}^{t-1}, \xi_t) \\
\boldsymbol{M}^t &= \beta_1 \boldsymbol{M}^{t-1} + (1-\beta_1)\boldsymbol{G}^t \\
\boldsymbol{\Theta}^t &= \boldsymbol{\Theta}^{t-1} - \eta \,\mathrm{matrixsign}(\boldsymbol{M}^t) - \eta\lambda\boldsymbol{\Theta}^{t-1}
\end{aligned}
$$

- $\mathrm{matrixsign}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top) = \boldsymbol{U}\boldsymbol{V}^\top$ — can be approximated via **Newton-Schultz Iterations**

  [Jordan et al, 2024]

Two Controversial Learning Rate Scaling Rules for Muon

- μP version [Yang et al., 2023]:

$$\sqrt{\frac{\texttt{d\_out}}{\texttt{d\_in}}} \cdot \underbrace{\eta_{\texttt{base}}}_{\text{optimal LR for the base (small) model}}$$

  - Grafting in spectral geometry

- Moonlight version [Liu et al., 2025, Moonshot AI]:

$$\sqrt{\max\{\texttt{d\_out}, \texttt{d\_in}\}} \cdot \underbrace{\eta_{\texttt{Adam}}}_{\text{tuned LR using AdamW}}$$

  - Deployed in Kimi
  - Grafting in Euclidean geometry [Agarwal et al, 2020]
  - Leads to a $O(\sqrt{1/\text{width}})$ decay in optimal learning rate.

Here, `d_out` denotes the effective output dimensionality of the parameter block.

Unifying via Steepest Descent Under Different Matrix Operator Norms

Matrix Thinking

All these optimizers admit a unified interpretation as steepest descent under different matrix operator norms.

$$\boxed{\boldsymbol{D}^t = \arg\min_{\|\boldsymbol{D}\|_? \le 1} \langle \boldsymbol{M}^t, \boldsymbol{D} \rangle} \qquad (\star)$$

Definition (Operator Norm)

Let $\|\cdot\|_{\text{in}}$ and $\|\cdot\|_{\text{out}}$ be norms on $\mathbb{R}^n$ and $\mathbb{R}^m$. The operator norm of $\boldsymbol{D}$ induced by these norms is defined as

$$\|\boldsymbol{D}\|_{\text{in}\to\text{out}} := \sup_{\|\boldsymbol{x}\|_{\text{in}}=1} \|\boldsymbol{D}\boldsymbol{x}\|_{\text{out}}.$$

Unifying via Steepest Descent Under Different Matrix Operator Norms

Matrix Thinking

All these optimizers admit a unified interpretation as steepest descent under different matrix operator norms.

$$\boxed{\boldsymbol{D}^t = \arg\min_{\|\boldsymbol{D}\|_? \leq 1} \langle \boldsymbol{M}^t, \boldsymbol{D} \rangle} \qquad (\star)$$

Definition (Operator Norm)

Let $\|\cdot\|_{\text{in}}$ and $\|\cdot\|_{\text{out}}$ be norms on $\mathbb{R}^n$ and $\mathbb{R}^m$. The operator norm of $\boldsymbol{D}$ induced by these norms is defined as

$$\|\boldsymbol{D}\|_{\text{in}\to\text{out}} := \sup_{\|\boldsymbol{x}\|_{\text{in}}=1} \|\boldsymbol{D}\boldsymbol{x}\|_{\text{out}}.$$

Unifying via Steepest Descent Under Different Matrix Operator Norms

Matrix Thinking

All these optimizers admit a unified interpretation as steepest descent under different matrix operator norms.

$$\boxed{\boldsymbol{D}^t = \arg\min_{\|\boldsymbol{D}\|_? \leq 1} \langle \boldsymbol{M}^t, \boldsymbol{D} \rangle} \qquad (\star)$$

Definition (Operator Norm)

Let $\|\cdot\|_{\text{in}}$ and $\|\cdot\|_{\text{out}}$ be norms on $\mathbb{R}^n$ and $\mathbb{R}^m$. The operator norm of $\boldsymbol{D}$ induced by these norms is defined as

$$\|\boldsymbol{D}\|_{\text{in}\to\text{out}} := \sup_{\|\boldsymbol{x}\|_{\text{in}}=1} \|\boldsymbol{D}\boldsymbol{x}\|_{\text{out}}.$$

Select the Correct Matrix Operator Norm

Example (SignSGD/Adam/Lion, $\ell_1 \to \ell_\infty$)

The element-wise ℓ_∞ norm of $\boldsymbol{D}$ coincides with its $\ell_1 \to \ell_\infty$ operator norm, i.e.,

$$\|\boldsymbol{D}\|_{1\to\infty} = \max_{i,j} |\boldsymbol{D}_{i,j}| \quad \text{and} \quad \boldsymbol{D}^* = \operatorname*{arg\,min}_{\|\boldsymbol{D}\|_{1\to\infty}\leq 1} \langle \boldsymbol{M}, \boldsymbol{D}\rangle = -\mathtt{sign}(\boldsymbol{M}).$$

Example (Muon, $\ell_2 \to \ell_2$)

The spectral norm of $\boldsymbol{D}$ coincides with its $\ell_2 \to \ell_2$ operator norm, i.e.,

$$\|\boldsymbol{D}\|_{2\to 2} = \|\boldsymbol{D}\|_2 \quad \text{and} \quad \boldsymbol{D}^* = \operatorname*{arg\,min}_{\|\boldsymbol{D}\|_{2\to 2}\leq 1} \langle \boldsymbol{M}, \boldsymbol{D}\rangle = -\mathtt{matrixsign}(\boldsymbol{M}).$$

Select the Correct Matrix Operator Norm

Example (SignSGD/Adam/Lion, $\ell_1 \to \ell_\infty$)

The element-wise ℓ_∞ norm of $\boldsymbol{D}$ coincides with its $\ell_1 \to \ell_\infty$ operator norm, i.e.,

$$\|\boldsymbol{D}\|_{1\to\infty} = \max_{i,j} |\boldsymbol{D}_{i,j}| \quad \text{and} \quad \boldsymbol{D}^* = \operatorname*{arg\,min}_{\|\boldsymbol{D}\|_{1\to\infty}\leq 1} \langle \boldsymbol{M}, \boldsymbol{D}\rangle = -\mathtt{sign}(\boldsymbol{M}).$$

Example (Muon, $\ell_2 \to \ell_2$)

The spectral norm of $\boldsymbol{D}$ coincides with its $\ell_2 \to \ell_2$ operator norm, i.e.,

$$\|\boldsymbol{D}\|_{2\to 2} = \|\boldsymbol{D}\|_2 \quad \text{and} \quad \boldsymbol{D}^* = \operatorname*{arg\,min}_{\|\boldsymbol{D}\|_{2\to 2}\leq 1} \langle \boldsymbol{M}, \boldsymbol{D}\rangle = -\mathtt{matrixsign}(\boldsymbol{M}).$$

Select the Correct Matrix Operator Norm II [Bernstein and Newhouse, 2024]

Example (Column Normalization, $\ell_1 \to \ell_q$)

$$\|\boldsymbol{D}\|_{1\to q} = \max_{c\in[n]} \|\boldsymbol{D}_{:,c}\|_q, \text{ and}$$

$$\boldsymbol{D}^\star = \mathrm{colnorm}_q(\boldsymbol{M}) \quad \text{and} \quad \mathrm{colnorm}_q(\boldsymbol{M})_{:,c} := \frac{\mathrm{sign}(\boldsymbol{M}_{:,c}) \odot |\boldsymbol{M}_{:,c}|^{q^*-1}}{\|\boldsymbol{M}_{:,c}\|_{q^*}^{q^*-1}},$$

where $\frac{1}{q} + \frac{1}{q^*} = 1$.

Example (Row Normalization, $\ell_p \to \ell_\infty$)

$$\|\boldsymbol{D}\|_{p\to\infty} = \max_{r\in[m]} \|\boldsymbol{D}_{r,:}\|_{p^*},$$

$$\boldsymbol{D}^\star = \mathrm{rownorm}_p(\boldsymbol{M}) \quad \text{and} \quad \mathrm{rownorm}_p(\boldsymbol{M})_{r,:} := \frac{\mathrm{sign}(\boldsymbol{M}_{r,:}) \odot |\boldsymbol{M}_{r,:}|^{p-1}}{\|\boldsymbol{M}_{r,:}\|_p^{p-1}},$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$.

Select the Correct Matrix Operator Norm II [Bernstein and Newhouse, 2024]

Example (Column Normalization, $\ell_1 \to \ell_q$)

$$\|\boldsymbol{D}\|_{1\to q} = \max_{c\in[n]} \|\boldsymbol{D}_{:,c}\|_q, \text{ and}$$

$$\boldsymbol{D}^\star = \mathrm{colnorm}_q(\boldsymbol{M}) \quad \text{and} \quad \mathrm{colnorm}_q(\boldsymbol{M})_{:,c} := \frac{\mathrm{sign}(\boldsymbol{M}_{:,c}) \odot |\boldsymbol{M}_{:,c}|^{q^*-1}}{\|\boldsymbol{M}_{:,c}\|_{q^*}^{q^*-1}},$$

where $\frac{1}{q} + \frac{1}{q^*} = 1$.

Example (Row Normalization, $\ell_p \to \ell_\infty$)

$$\|\boldsymbol{D}\|_{p\to\infty} = \max_{r\in[m]} \|\boldsymbol{D}_{r,:}\|_{p^*},$$

$$\boldsymbol{D}^\star = \mathrm{rownorm}_p(\boldsymbol{M}) \quad \text{and} \quad \mathrm{rownorm}_p(\boldsymbol{M})_{r,:} := \frac{\mathrm{sign}(\boldsymbol{M}_{r,:}) \odot |\boldsymbol{M}_{r,:}|^{p-1}}{\|\boldsymbol{M}_{r,:}\|_{p}^{p-1}},$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$.

Matrix Thinking

SignSGD/AdamW	Column Normalization	Row Normalization	Muon
$\|\cdot\|_{1\to\infty}$	$\|\cdot\|_{1\to q}$	$\|\cdot\|_{p\to\infty}$	$\|\cdot\|_{2\to 2}$

Choosing the right matrix operator norm is the key to designing new optimizers!

Geometry-Aware Principles for Optimizer Design

- **Low per-iteration computational cost**
  - Choosing matrix operator geometries $\|\cdot\|_{p\to q}$ with $p \leq q$.
- **Neural network landscape**
  - **Width-independent Lipschitz bound:** the network's *forward sensitivity* is bounded independently of width.
  - **Width-independent smoothness bound:** the *second-order sensitivity* is bounded independently of width.

These width-independent bounds naturally induce ***learning rate scaling rules*** that are stable across widths, enabling effective hyperparameter transfer.

Width-independent Lipschitz Continuous NN

Observation

Standard operator norms ($p \to q$) fail to yield width-independent Lipschitz bounds.

- Each linear map $\boldsymbol{W}_\ell$ sends features from one geometric space to the next pre-activation space. The operator norm of $\boldsymbol{W}_\ell$ induces a layer-wise propagation of stability, i.e.,

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{l+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_l \nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{out}} \leq \|\boldsymbol{W}_l\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{in}}, \forall i < \ell.$$

- Cross-layer stability bound using matrix operator norms requires: $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_\ell\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \|\boldsymbol{W}_{\ell+1}\|_{\text{in}\to\text{out}} \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{in}}$$

$$\Rightarrow \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \prod_{k=l}^{l+1} \|\boldsymbol{W}_k\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$

Width-independent Lipschitz Continuous NN

Observation

Standard operator norms ($p \to q$) fail to yield width-independent Lipschitz bounds.

- Each linear map $\boldsymbol{W}_\ell$ sends features from one geometric space to the next pre-activation space. The operator norm of $\boldsymbol{W}_\ell$ induces a layer-wise propagation of stability, i.e.,

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{l+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_l \nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{out}} \leq \|\boldsymbol{W}_l\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{in}}, \forall i < \ell.$$

- Cross-layer stability bound using matrix operator norms requires: $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_\ell\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$
$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \|\boldsymbol{W}_{\ell+1}\|_{\text{in}\to\text{out}} \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{in}}$$

$$\Rightarrow \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \prod_{k=l}^{l+1} \|\boldsymbol{W}_k\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$

Width-independent Lipschitz Continuous NN

Observation

Standard operator norms ($p \to q$) fail to yield width-independent Lipschitz bounds.

- Each linear map $\boldsymbol{W}_\ell$ sends features from one geometric space to the next pre-activation space. The operator norm of $\boldsymbol{W}_\ell$ induces a layer-wise propagation of stability, i.e.,

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{l+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_l \nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{out}} \leq \|\boldsymbol{W}_l\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{in}}, \forall i < \ell.$$

- Cross-layer stability bound using matrix operator norms requires: $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_\ell\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$
$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \|\boldsymbol{W}_{\ell+1}\|_{\text{in}\to\text{out}} \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{in}}$$

$$\Rightarrow \boxed{\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \prod_{k=l}^{l+1} \|\boldsymbol{W}_k\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}}$$

Width-independent Lipschitz Continuous NN

Observation

Standard operator norms ($p \to q$) fail to yield width-independent Lipschitz bounds.

- Each linear map $\boldsymbol{W}_\ell$ sends features from one geometric space to the next pre-activation space. The operator norm of $\boldsymbol{W}_\ell$ induces a layer-wise propagation of stability, i.e.,

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{l+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_l \nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{out}} \leq \|\boldsymbol{W}_l\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{in}}, \forall i < \ell.$$

- Cross-layer stability bound using matrix operator norms requires: $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_\ell\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$
$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \|\boldsymbol{W}_{\ell+1}\|_{\text{in}\to\text{out}} \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{in}}$$

Example

Consider $\boldsymbol{x} = (1, 1, \ldots, 1) \in \mathbb{R}^n$. We have $\|\boldsymbol{x}\|_\infty = 1 < \|\boldsymbol{x}\|_p = n^{1/p} < \|\boldsymbol{x}\|_1 = n$.

Width-independent Lipschitz Continuous NN

Observation

Standard operator norms ($p \to q$) fail to yield width-independent Lipschitz bounds.

- Each linear map $\boldsymbol{W}_\ell$ sends features from one geometric space to the next pre-activation space. The operator norm of $\boldsymbol{W}_\ell$ induces a layer-wise propagation of stability, i.e.,

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{l+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_l \nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{out}} \leq \|\boldsymbol{W}_l\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{in}}, \forall i < \ell.$$

- Cross-layer stability bound using matrix operator norms requires: $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_\ell\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$
$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \|\boldsymbol{W}_{\ell+1}\|_{\text{in}\to\text{out}} \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{in}}$$

Example

Consider $\boldsymbol{x} = (1, 1, \ldots, 1) \in \mathbb{R}^n$. We have $\|\boldsymbol{x}\|_\infty = 1 < \|\boldsymbol{x}\|_p = n^{1/p} < \|\boldsymbol{x}\|_1 = n$.

Width-independent Lipschitz Continuous NN

Observation

Standard operator norms ($p \to q$) fail to yield width-independent Lipschitz bounds.

- Each linear map $\boldsymbol{W}_\ell$ sends features from one geometric space to the next pre-activation space. The operator norm of $\boldsymbol{W}_\ell$ induces a layer-wise propagation of stability, i.e.,

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{l+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_l \nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{out}} \leq \|\boldsymbol{W}_l\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_l\|_{\text{in}}, \forall i < \ell.$$

- Cross-layer stability bound using matrix operator norms requires: $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$

$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{out}} \leq \|\boldsymbol{W}_\ell\|_{\text{in}\to\text{out}} \|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_\ell\|_{\text{in}}$$
$$\left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+2}\right\|_{\text{out}} \leq \|\boldsymbol{W}_{\ell+1}\|_{\text{in}\to\text{out}} \left\|\nabla_{\boldsymbol{W}_i} \boldsymbol{y}_{\ell+1}\right\|_{\text{in}}$$

Example

Consider $\boldsymbol{x} = (1, 1, \ldots, 1) \in \mathbb{R}^n$. We have $\|\boldsymbol{x}\|_\infty = 1 < \|\boldsymbol{x}\|_p = n^{1/p} < \|\boldsymbol{x}\|_1 = n$.

How to make the matrix operator norm play nicely together?

Definition (Mean-Normalized Norm)

We introduce a new, width-aware geometry via the mean-normalized norms $\|\cdot\|_{(p,\mathtt{mean})}$:

$$\|\boldsymbol{x}\|_{(p,\mathtt{mean})} := \left(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i|^p\right)^{1/p} = n^{-1/p}\|\boldsymbol{x}\|_p.$$

- The factor $n^{-1/p}$ precisely cancels the dimensional scaling of the ℓ_p embeddings.

Fact

Let $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le p \le q \le \infty$, then $\|\boldsymbol{x}\|_{(p,mean)} \le \|\boldsymbol{x}\|_{(q,mean)}$.

- Choosing the matrix operator norm $\|\cdot\|_{(p,\mathtt{mean})\to(q,\mathtt{mean})}$ with $p \le q$!

How to make the matrix operator norm play nicely together?

Definition (Mean-Normalized Norm)

We introduce a new, width-aware geometry via the mean-normalized norms $\|\cdot\|_{(p,\mathrm{mean})}$:

$$\|\boldsymbol{x}\|_{(p,\mathrm{mean})} := \left(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i|^p\right)^{1/p} = n^{-1/p}\|\boldsymbol{x}\|_p.$$

- The factor $n^{-1/p}$ precisely cancels the dimensional scaling of the ℓ_p embeddings.

Fact

Let $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le p \le q \le \infty$, then $\|\boldsymbol{x}\|_{(p,mean)} \le \|\boldsymbol{x}\|_{(q,mean)}$.

- Choosing the matrix operator norm $\|\cdot\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})}$ with $p \le q$!

How to make the matrix operator norm play nicely together?

Definition (Mean-Normalized Norm)

We introduce a new, width-aware geometry via the mean-normalized norms $\|\cdot\|_{(p,\mathrm{mean})}$:

$$\|\boldsymbol{x}\|_{(p,\mathrm{mean})} := \left(\frac{1}{n}\sum_{i=1}^{n}|\boldsymbol{x}_i|^p\right)^{1/p} = n^{-1/p}\|\boldsymbol{x}\|_p.$$

- The factor $n^{-1/p}$ precisely cancels the dimensional scaling of the ℓ_p embeddings.

Fact

Let $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le p \le q \le \infty$, then $\|\boldsymbol{x}\|_{(p,mean)} \le \|\boldsymbol{x}\|_{(q,mean)}$.

- Choosing the matrix operator norm $\|\cdot\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})}$ with $p \le q$!

Width-independent Lipschitz Bound

Theorem (Width-Indepedent Lipschitz Bound Under $(p, \text{mean}) \to (q, \text{mean})$ Geometry)

Suppose that (1) $1 \leq p \leq q < \infty$; (2) the input is bounded; (3) both σ and $\mathcal{L}$ have bounded first derivatives. Define the parameter set as (4) $\Omega_C := \{\boldsymbol{\Theta} : \|\boldsymbol{\Theta}\|_{\text{block}} \leq C\}$, where the block norm $\|\boldsymbol{\Theta}\|_{\text{block}}$ is defined as

$$\max\left\{ \|\boldsymbol{W}_1\|_{1\to(q,\text{mean})}, \max_{2\leq i\leq K-1} \|\boldsymbol{W}_i\|_{(p,\text{mean})\to(q,\text{mean})}, \|\boldsymbol{W}_K\|_{(p,\text{mean})\to\infty}, \max_i \|\boldsymbol{b}_i\|_\infty \right\}.$$

Then the loss function is M-Lipschitz continuous on Ω_C, where M is independent of the width.

Width-independent L-Smooth NN

Definition (L-Smoothness)

Let $f : \Omega \to \mathbb{R}$ be differentiable on a convex set Ω equipped with a norm $\|\cdot\|$. We say that f is *L-smooth* with respect to $\|\cdot\|$ if, for all $\boldsymbol{Z}, \boldsymbol{Z}' \in \Omega$,

$$\|\nabla f(\boldsymbol{Z}) - \nabla f(\boldsymbol{Z}')\|_* \leq L\|\boldsymbol{Z} - \boldsymbol{Z}'\|,$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

Theorem (Smoothness Bound Under $(p, \text{mean}) \to (q, \text{mean})$ Geometry)

Suppose that (1) $1 \leq p \leq q < \infty$; (2) the input is bounded; (3) both σ and $\mathcal{L}$ have bounded second derivatives. Then, we have

$$\|\nabla f(\Theta) - \nabla f(\Theta')\|_{block,\star} \leq L\, w^{\max\left(0, \frac{2}{q} - \frac{1}{p}\right)} \|\Theta - \Theta'\|_{\text{block}} \quad \forall\, \Theta, \Theta' \in \Omega_C,$$

where $L > 0$ is independent of the width.

$q \geq 2p$!!!

Width-independent L-Smooth NN

Definition (L-Smoothness)

Let $f : \Omega \to \mathbb{R}$ be differentiable on a convex set Ω equipped with a norm $\|\cdot\|$. We say that f is *L-smooth* with respect to $\|\cdot\|$ if, for all $\boldsymbol{Z}, \boldsymbol{Z}' \in \Omega$,

$$\|\nabla f(\boldsymbol{Z}) - \nabla f(\boldsymbol{Z}')\|_* \leq L\|\boldsymbol{Z} - \boldsymbol{Z}'\|,$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

Theorem (Smoothness Bound Under $(p, \text{mean}) \to (q, \text{mean})$ Geometry)

Suppose that (1) $1 \leq p \leq q < \infty$; (2) the input is bounded; (3) both σ and $\mathcal{L}$ have bounded second derivatives. Then, we have

$$\|\nabla f(\boldsymbol{\Theta}) - \nabla f(\boldsymbol{\Theta}')\|_{block,\star} \leq L\, w^{\max\left(0,\, \frac{2}{q} - \frac{1}{p}\right)} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_{\text{block}} \quad \forall\, \boldsymbol{\Theta}, \boldsymbol{\Theta}' \in \Omega_C,$$

where $L > 0$ is independent of the width.

$q \geq 2p$!!!

Width-independent L-Smooth NN

Definition (L-Smoothness)

Let $f : \Omega \to \mathbb{R}$ be differentiable on a convex set Ω equipped with a norm $\|\cdot\|$. We say that f is *L-smooth* with respect to $\|\cdot\|$ if, for all $\boldsymbol{Z}, \boldsymbol{Z}' \in \Omega$,

$$\|\nabla f(\boldsymbol{Z}) - \nabla f(\boldsymbol{Z}')\|_* \leq L\|\boldsymbol{Z} - \boldsymbol{Z}'\|,$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$.

Theorem (Smoothness Bound Under $(p, \text{mean}) \to (q, \text{mean})$ Geometry)

Suppose that (1) $1 \leq p \leq q < \infty$; (2) the input is bounded; (3) both σ and $\mathcal{L}$ have bounded second derivatives. Then, we have

$$\|\nabla f(\boldsymbol{\Theta}) - \nabla f(\boldsymbol{\Theta}')\|_{block,\star} \leq L\, w^{\max\left(0,\, \frac{2}{q} - \frac{1}{p}\right)} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_{\text{block}} \quad \forall\, \boldsymbol{\Theta}, \boldsymbol{\Theta}' \in \Omega_C,$$

where $L > 0$ is independent of the width.

$$\boxed{q \geq 2p} \text{ !!!}$$

Theoretical Insights

The directional Hessian can be reduced to controlling this quadratic term:

Feature- or point-wise nonlinearities induce element-wise multiplication in the quadratic Taylor term.

$$\left|\nabla^2 f(\boldsymbol{W})[\Delta \boldsymbol{W}^1, \Delta \boldsymbol{W}^2]\right| \lesssim \|(\Delta \boldsymbol{W}^1 \boldsymbol{x}) \odot (\Delta \boldsymbol{W}^2 \boldsymbol{x})\|_{(p,\mathrm{mean})}$$

$$\lesssim w^{\max\left(0, \frac{2}{q} - \frac{1}{p}\right)} \|\Delta \boldsymbol{W}^1 \boldsymbol{x}\|_{(q,\mathrm{mean})} \|\Delta \boldsymbol{W}^2 \boldsymbol{x}\|_{(q,\mathrm{mean})}.$$

Due to the definition of the matrix operator norm, we have

$$\left\|\Delta \boldsymbol{W}^1 \boldsymbol{x}\right\|_{(q,\mathrm{mean})} \leq \left\|\Delta \boldsymbol{W}^1\right\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})} \|\boldsymbol{x}\|_{(p,\mathrm{mean})} \leq \|\boldsymbol{x}\|_{(p,\mathrm{mean})}.$$

Lemma (Hölder-type Inequality for Hadamard Product)

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. For any $p, q \geq 1$, we have

$$\|\boldsymbol{x} \odot \boldsymbol{y}\|_{(p,\mathrm{mean})} \leq n^{\max\left(0, \frac{2}{q} - \frac{1}{p}\right)} \|\boldsymbol{x}\|_{(q,\mathrm{mean})} \|\boldsymbol{y}\|_{(q,\mathrm{mean})}.$$

Theoretical Insights

The directional Hessian can be reduced to controlling this quadratic term:

Feature- or point-wise nonlinearities induce element-wise multiplication in the quadratic Taylor term.

$$\begin{aligned}\left|\nabla^2 f(\boldsymbol{W})[\Delta\boldsymbol{W}^1, \Delta\boldsymbol{W}^2]\right| &\lesssim \|(\Delta\boldsymbol{W}^1\boldsymbol{x}) \odot (\Delta\boldsymbol{W}^2\boldsymbol{x})\|_{(p,\mathrm{mean})} \\ &\lesssim w^{\max\left(0,\, \frac{2}{q}-\frac{1}{p}\right)} \|\Delta\boldsymbol{W}^1\boldsymbol{x}\|_{(q,\mathrm{mean})} \|\Delta\boldsymbol{W}^2\boldsymbol{x}\|_{(q,\mathrm{mean})}.\end{aligned}$$

Due to the definition of the matrix operator norm, we have

$$\left\|\Delta\boldsymbol{W}^1\boldsymbol{x}\right\|_{(q,\mathrm{mean})} \leq \left\|\Delta\boldsymbol{W}^1\right\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})} \|\boldsymbol{x}\|_{(p,\mathrm{mean})} \leq \|\boldsymbol{x}\|_{(p,\mathrm{mean})}.$$

Lemma (Hölder-type Inequality for Hadamard Product)

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. For any $p, q \geq 1$, we have

$$\|\boldsymbol{x} \odot \boldsymbol{y}\|_{(p,\mathrm{mean})} \leq n^{\max\left(0,\, \frac{2}{q}-\frac{1}{p}\right)} \|\boldsymbol{x}\|_{(q,\mathrm{mean})} \|\boldsymbol{y}\|_{(q,\mathrm{mean})}.$$

Width Aware Learning Rate Scaling Rule

Matrix Operator Geometry Aware (MOGA) Scaling

For a matrix $\boldsymbol{D} \in \mathbb{R}^{\texttt{d\_out} \times \texttt{d\_in}}$ and $1 \le p \le q \le \infty$, we have

$$\|\boldsymbol{D}\|_{(p,\texttt{mean})\to(q,\texttt{mean})} = \boxed{\frac{\texttt{d\_in}^{1/p}}{\texttt{d\_out}^{1/q}}} \|\boldsymbol{D}\|_{p\to q}.$$

- In the case of `Adam` ($\ell_1 \to \ell_\infty$), `MOGA` scaling exactly recovers the μP scaling $\boxed{\frac{1}{\texttt{d\_in}}}$.
- For `Muon` ($\ell_2 \to \ell_2$), `MOGA` scaling recovers the μP versioned Muon $\boxed{\sqrt{\frac{\texttt{d\_out}}{\texttt{d\_in}}}}$.
  - [Yang et al, 2023] proposed a **spectral condition** requiring the update direction to have spectral norm $\sqrt{\texttt{d\_out}/\texttt{d\_in}}$.
  - However, when $p = q = 2$, we have $L = \mathcal{O}(\sqrt{w})$, which matches the learning rate scaling rule in the **Moonlight** version.

Matrix Operator Geometry Aware (MOGA) Optimizer

Algorithm 1 MOGA: Matrix-Operator-Geometry-Aware Steepest Descent

Require: Learning rates $\{\eta^t\}_{t\geq 0}$, momentum parameters $\beta_1, \beta_2 \in (0,1)$, initial parameters $\boldsymbol{\Theta}^0$

$\boldsymbol{M}^0 \leftarrow \boldsymbol{0}$ ▷ $\boldsymbol{M}^t$ has the same block structure as $\boldsymbol{\Theta}^t$

for $t = 1, 2, \ldots$ **do**

  Sample stochastic gradient: $\boldsymbol{G} \leftarrow \nabla F(\boldsymbol{\Theta}^{t-1}; \xi)$, $\xi \sim \mathbb{P}$

  Exponential moving average: $\boldsymbol{M}^t \leftarrow \beta_1 \boldsymbol{M}^{t-1} + (1-\beta_1)\boldsymbol{G}$

  Lookahead (Nesterov-type) momentum: $\tilde{\boldsymbol{M}}^t \leftarrow \beta_2 \boldsymbol{M}^{t-1} + (1-\beta_2)\boldsymbol{G}$

  if Descent under $(1, \text{mean}) \to (q, \text{mean})$ **then** ▷ $q \geq 2$

    for $i = 1$ to ℓ **do**

      $$\boldsymbol{W}_i^t \leftarrow \boldsymbol{W}_i^{t-1} - \eta^t \cdot \boxed{\frac{(d_{\text{out}}^i)^{1/q}}{d_{\text{in}}^i}} \cdot \text{colnorm}_q\left(\tilde{\boldsymbol{M}}^t_{\boldsymbol{W}_i^{t-1}}\right)$$

      $$\boldsymbol{b}_i^t \leftarrow \boldsymbol{b}_i^{t-1} - \eta^t \,\text{sign}\left(\tilde{\boldsymbol{M}}^t_{\boldsymbol{b}_i}\right)$$ ▷ or steepest descent under (q, mean) norm

    end for

  else if Descent under $(p, \text{mean}) \to \infty$ **then**

    for $i = 1$ to ℓ **do**

      $$\boldsymbol{W}_i^t \leftarrow \boldsymbol{W}_i^{t-1} - \eta^t \cdot \boxed{\frac{1}{(d_{\text{in}}^i)^{1/p}}} \cdot \text{rownorm}_p\left(\tilde{\boldsymbol{M}}^t_{\boldsymbol{W}_i^{t-1}}\right)$$

      $$\boldsymbol{b}_i^t \leftarrow \boldsymbol{b}_i^{t-1} - \eta^t \,\text{sign}\left(\tilde{\boldsymbol{M}}^t_{\boldsymbol{b}_i}\right)$$

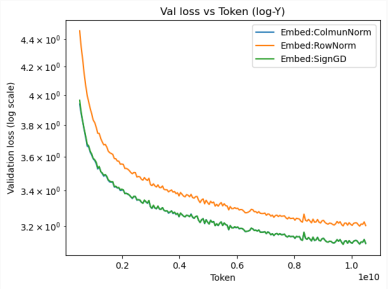
    end for

  end if

end for

Adapting to Transformers

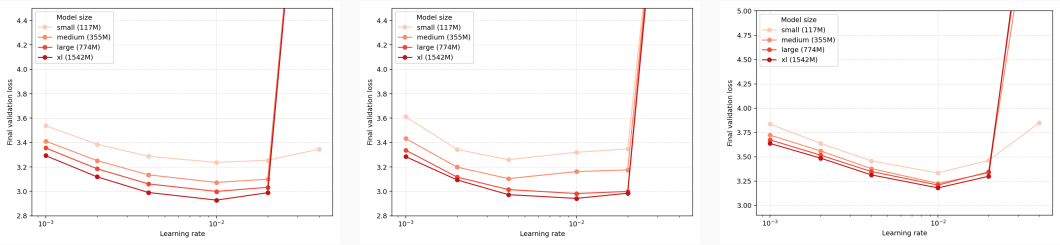
- **Linear layers in MLPs and attention:** optimized using MOGA with scaling.
- **Biases and LayerNorm parameters:** optimized using SignGD.



- **Input word and positional embeddings:** map one-hot vectors (with ℓ_1 geometry) to feature representations (with (q, mean) geometry).
- **Word unembedding:** the unembedding matrix is tied to the word embedding matrix.

Learning Rate Transfer Across Model Scales

Architecture: GPT-2 family (Small → XL)



(a) MOGA (p=1.5) **(b)** MOGA (p=2) **(c)** MOGA (p=3)

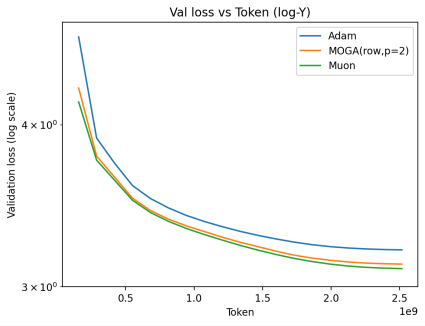
Figure 1: Learning Rate Transfer. Performance of MOGA from GPT-2 Small to GPT-XL. The optimal learning rate of MOGA is invariant to width.

LLM Pretraining — Standard Token Budget

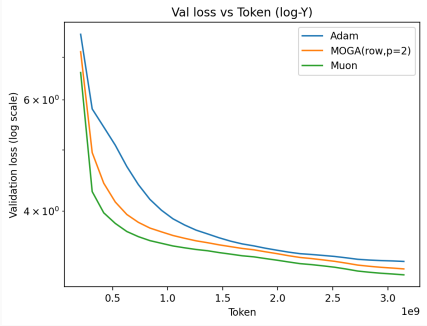
Model 1: LLaMA-130m (130M params) **Data 1**: C4 (Raffel et al., 2020).

Model 2: GPT-2 Small (117M params) **Data 2**: OpenWebText (Gokaslan et al., 2019).

Training Tokens: $\sim$ 1 Chinchilla (20$\times$ params).



Llama-130m model

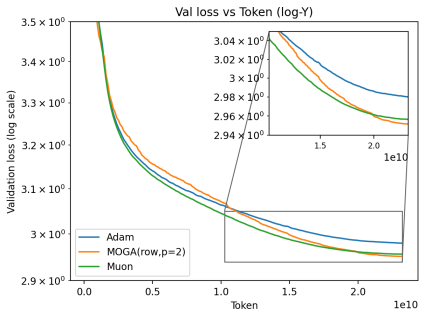


GPT-2 Small model

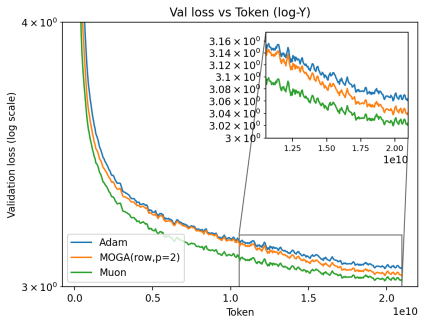
MOGA achieves almost the same speed as Muon in 1 chinchilla training.

LLM Pretraining – Large Token Budget

Training Tokens: $\sim$ 8 Chinchilla (160$\times$ params).



Llama-130m model



GPT-2 Small model

MOGA achieves faster convergence in the later stage of training when the loss is small!

[1] Muon deliver about 1.3$\times$ speedups on small models (0.1B parameters), but the gain declines to roughly 1.1$\times$ for 1.2B-parameter models trained at an 8$\times$ Chinchilla ratio (Figure 1, bottom left). [Wen et al, 2025].

Related Works on Row/Col Normalization

- Training Deep Learning Models with Norm-Constrained LMOs, arXiv:2502.07529
- A Minimalist Optimizer Design for LLM Pretraining, arXiv:2506.16659
- Mano: Restriking Manifold Optimization for LLM Training, arXiv:2601.23000
- RMNP: Row-Momentum Normalized Preconditioning for Scalable Matrix-Based Optimization, arXiv:2603.20527

Take-home Messages

- **Hyperparameter transfer** is necessary for LLM optimizer design.
- Steepest descent under different matrix operator norms (**Matrix Thinking**) offers a unified framework to understand successful existing optimizers (e.g., `SignSGD`, `AdamW`, `Muon`). **Selecting** the appropriate matrix operator norm is the key to designing new optimizers.
- Taking **width-independent Lipschitz/smoothness properties** as opt design principle.
- We introduce the mean-normalized $(p, \texttt{mean}) \to (q, \texttt{mean})$ operator norm to make the matrix operator norm play nicely together. This norm yields width-independent bounds when $q \geq 2p$, and naturally lead to width-aware learning rate rescheduling rules and new optimizers `MOGA`.
- When $p = 1, q = \infty$ we recover μ**P scaling** in the Adam Optimizer.

Take-home Messages

- **Hyperparameter transfer** is necessary for LLM optimizer design.
- Steepest descent under different matrix operator norms (**Matrix Thinking**) offers a unified framework to understand successful existing optimizers (e.g., `SignSGD`, `AdamW`, `Muon`). **Selecting** the appropriate matrix operator norm is the key to designing new optimizers.
- Taking **width-independent Lipschitz/smoothness properties** as opt design principle.
- We introduce the mean-normalized $(p, \texttt{mean}) \to (q, \texttt{mean})$ operator norm to make the matrix operator norm play nicely together. This norm yields width-independent bounds when $q \geq 2p$, and naturally lead to width-aware learning rate rescheduling rules and new optimizers `MOGA`.
- When $p = 1, q = \infty$ we recover μ**P scaling** in the Adam Optimizer.

Take-home Messages

- **Hyperparameter transfer** is necessary for LLM optimizer design.
- Steepest descent under different matrix operator norms (**Matrix Thinking**) offers a unified framework to understand successful existing optimizers (e.g., `SignSGD`, `AdamW`, `Muon`). **Selecting** the appropriate matrix operator norm is the key to designing new optimizers.
- Taking **width-independent Lipschitz/smoothness properties** as opt design principle.
- We introduce the mean-normalized $(p, \texttt{mean}) \to (q, \texttt{mean})$ operator norm to make the matrix operator norm play nicely together. This norm yields width-independent bounds when $q \geq 2p$, and naturally lead to width-aware learning rate rescheduling rules and new optimizers `MOGA`.
- When $p = 1, q = \infty$ we recover μ**P scaling** in the Adam Optimizer.

Take-home Messages

- **Hyperparameter transfer** is necessary for LLM optimizer design.
- Steepest descent under different matrix operator norms (**Matrix Thinking**) offers a unified framework to understand successful existing optimizers (e.g., `SignSGD`, `AdamW`, `Muon`). **Selecting** the appropriate matrix operator norm is the key to designing new optimizers.
- Taking **width-independent Lipschitz/smoothness properties** as opt design principle.
- We introduce the mean-normalized $(p, \texttt{mean}) \to (q, \texttt{mean})$ operator norm to make the matrix operator norm play nicely together. This norm yields width-independent bounds when $q \geq 2p$, and naturally lead to width-aware learning rate rescheduling rules and new optimizers `MOGA`.
- When $p = 1, q = \infty$ we recover μ**P scaling** in the Adam Optimizer.

Take-home Messages

- **Hyperparameter transfer** is necessary for LLM optimizer design.
- Steepest descent under different matrix operator norms (**Matrix Thinking**) offers a unified framework to understand successful existing optimizers (e.g., `SignSGD`, `AdamW`, `Muon`). **Selecting** the appropriate matrix operator norm is the key to designing new optimizers.
- Taking **width-independent Lipschitz/smoothness properties** as opt design principle.
- We introduce the mean-normalized $(p, \texttt{mean}) \to (q, \texttt{mean})$ operator norm to make the matrix operator norm play nicely together. This norm yields width-independent bounds when $q \geq 2p$, and naturally lead to width-aware learning rate rescheduling rules and new optimizers `MOGA`.
- When $p = 1, q = \infty$ we recover μ**P scaling** in the Adam Optimizer.

Thank you for your listening!

Any questions?

Spectral Condition

- The change in feature

$$\|y_l\|_2 = \Theta(\sqrt{\text{fanout}}), \text{ and } \|\Delta y_l\|_2 = \Theta(\sqrt{\text{fanout}})$$

- As a result for parameter

$$\|W_l\|_2 = \Theta\left(\sqrt{\frac{\text{fanout}}{\text{fanin}}}\right), \text{ and } \|\Delta W_l\|_2 = \Theta\left(\sqrt{\frac{\text{fanout}}{\text{fanin}}}\right)$$

Yang G, Simon J B, Bernstein J. A spectral condition for feature learning. arXiv preprint arXiv:2310.17813, 2023.