CSML 2026

# On the Width Scaling of Neural Optimizers

## A Matrix Operator Norm Perspective

**Yiping Lu**

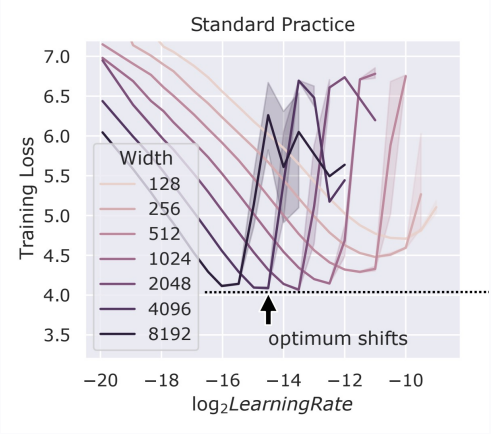
Beijing International Center for Mathematical Research
Peking University

Industrial Engineering & Management Science, Northwestern University

Joint work with Ruihan Xu (University of Chicago) and Jiajin Li (UBC)

$\|\cdot\|_{p\to q}$

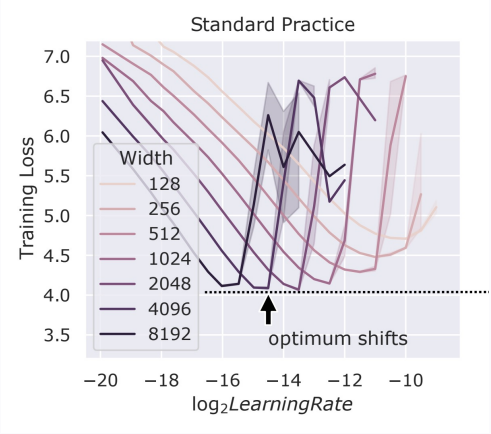
# Hyperparameter Transfer [Yang et al., NeurIPS 2021]




**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**.

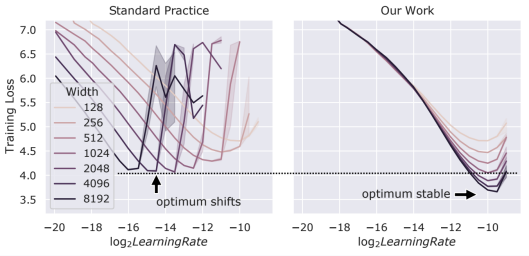
# Hyperparameter Transfer [Yang et al., NeurIPS 2021]




**Why this matters for LLMs**

Pretraining state of the art base models can cost **hundreds of millions of dollars**. Hyperparameter transfer is therefore **essential**, not optional.

# Hyperparameter Transfer [Yang et al., NeurIPS 2021]




Yang et al., NeurIPS 2021

`Adam` learning rate scaling:

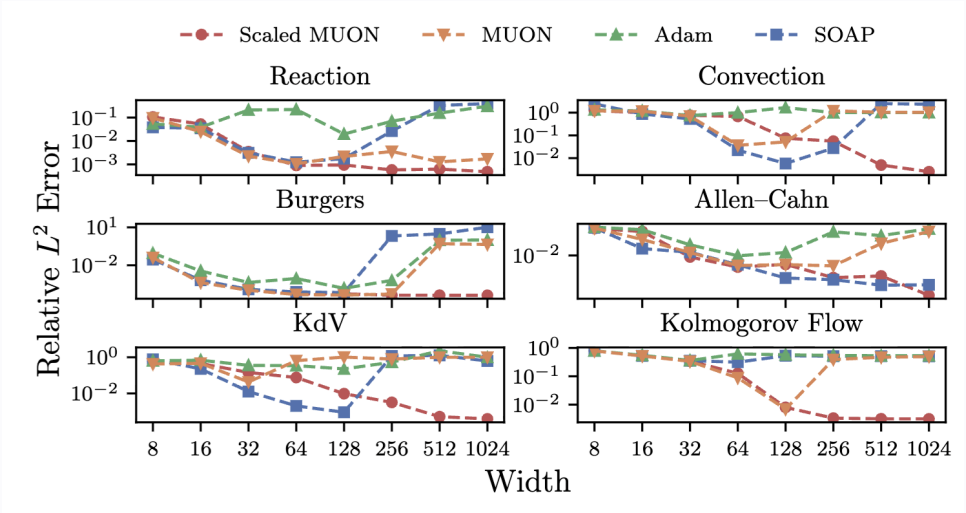
- **Weight matrices**
  - Scale learning rate by $\frac{1}{\texttt{d\_in}}$
- **Embedding matrices**
  - No learning-rate scaling

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

**Current theory is highly specialized**

Existing derivations require extensive random matrix arguments and justify transfer only near initialization.

# Large PINN is Hard to Train



# 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

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

where $\boldsymbol{y}_\ell(\boldsymbol{x}) \in \mathbb{R}$ is the output of an $\ell$-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 $\Theta := \{\boldsymbol{W}_{1:\ell},\, \boldsymbol{b}_{1:\ell}\}$ for simplicity.

# Decoupled Weight Decay as an Implicit Constraint

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

**Fact**

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

**View decoupled weight decay as enforcing an implicit norm ball constraint.**

**Remark**: Stationary criteria should be Frank-Wolfe gap.

# Decoupled Weight Decay as an Implicit Constraint

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 $\eta$ 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.**

**Remark**: Stationary criteria should be Frank-Wolfe gap.

# AdamW and signSGD

- AdamW is a standard optimizer for large scale neural network training. [Kingma and Ba, 2015; Loshchilov and Hutter, 2019]

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

Adam approximates `signSGD` [Bernstein et al, 2018, ICML]. $\boldsymbol{\theta}^t \approx \boldsymbol{\theta}^{t-1} - \eta \frac{\boldsymbol{g}^t}{\sqrt{\boldsymbol{g}^t \odot \boldsymbol{g}^t}} \approx \boldsymbol{\theta}^{t-1} - \eta \cdot \operatorname{sign}(\boldsymbol{g}^t)$.

# Lion: Sign Updates with Momentum [Chen et al., 2023]

**Key idea**

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

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

- Automatic discovery of optimizers through RL and meta learning
- Lower memory usage than AdamW

# Muon: Matrix Sign Updates

**Key idea**

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

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

- If $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, then $\mathrm{matrixsign}(\boldsymbol{G}) = \boldsymbol{U}\boldsymbol{V}^\top$; it can be approximated by **Newton Schulz iterations**. [Jordan et al., 2024]

# Competing Learning Rate Scaling Rules for Muon

**$\mu$P variant [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 variant [Liu et al., 2025]**

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

- Deployed in Kimi
- Euclidean grafting [Agarwal et al., 2020]
- Optimal learning rate decays as $O(1/\sqrt{\text{width}})$.

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

# Steepest Descent in Matrix Operator Geometry

**Matrix Thinking**

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

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

**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

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

# Steepest Descent in Matrix Operator Geometry

**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

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

# Choosing the Matrix Operator Norm

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

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

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

**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.,

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

# Choosing the Matrix Operator Norm

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

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

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

**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.,

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

# Column Normalization [Bernstein and Newhouse, 2024]

**Example (Geometry: $\ell_1 \to \ell_q$)**

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

$$\boldsymbol{D}^\star = -\mathrm{colnorm}_q(\boldsymbol{M}).$$

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

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

**Interpretation**

Each column is normalized in the dual geometry induced by $\ell_q$.

# Row Normalization [Bernstein and Newhouse, 2024]

**Example (Geometry: $\ell_p \to \ell_\infty$)**

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

$$\boldsymbol{D}^\star = -\mathrm{rownorm}_p(\boldsymbol{M}).$$

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

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

**Interpretation**

Each row is normalized in the dual geometry induced by $\ell_p$.

# Matrix Thinking

| SignSGD / AdamW | Column norm | Row norm | 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**
  - Choose matrix operator geometries $\|\cdot\|_{p\to q}$ with $p \le q$.
- **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 remain stable across widths and enable hyperparameter transfer.

# Why Standard Operator Norms Do Not Compose

**Observation**

Standard $\ell_p \to \ell_q$ operator norms introduce width dependent factors when they are composed across layers.

For one layer, operator control gives

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

**Cross layer requirement**

To iterate without a dimensional penalty, require $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$.

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

# Why Standard Operator Norms Do Not Compose

**Observation**

Standard $\ell_p \to \ell_q$ operator norms introduce width dependent factors when they are composed across layers.

For one layer, operator control gives

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

**Cross layer requirement**

To iterate without a dimensional penalty, require $\|\cdot\|_{\text{in}} \leq \|\cdot\|_{\text{out}}$.

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

# A Width Dependent Embedding Gap

The two consecutive bounds involve different norms of the same intermediate feature:

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

**Example (The all ones vector)**

For $\boldsymbol{x} = (1, \ldots, 1) \in \mathbb{R}^n$,

$$\|\boldsymbol{x}\|_\infty = 1,$$
$$\|\boldsymbol{x}\|_p = n^{1/p},$$
$$\|\boldsymbol{x}\|_1 = n.$$

**Consequence**

Standard norm embeddings inject powers of the width into the stability bound.

# Mean Normalized Norms Compose Across Layers

**Definition (Mean-Normalized Norm)**

We introduce a width aware geometry through 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 $\ell_p$ embeddings.

**Fact**

*For $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le p \le q \le \infty$,*

$$\|\boldsymbol{x}\|_{(p,mean)} \le \|\boldsymbol{x}\|_{(q,mean)}.$$

- Choose $\|\cdot\|_{(p,\mathtt{mean})\to(q,\mathtt{mean})}$ with $p \le q$.

# Mean Normalized Norms Compose Across Layers

**Definition (Mean-Normalized Norm)**

We introduce a width aware geometry through 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 $\ell_p$ embeddings.

**Fact**

*For $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le p \le q \le \infty$,*

$$\|\boldsymbol{x}\|_{(p,mean)} \le \|\boldsymbol{x}\|_{(q,mean)}.$$

- Choose $\|\cdot\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})}$ with $p \le q$.

# Mean Normalized Norms Compose Across Layers

**Definition (Mean-Normalized Norm)**

We introduce a width aware geometry through 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 $\ell_p$ embeddings.

**Fact**

*For $\boldsymbol{x} \in \mathbb{R}^n$ and $1 \le p \le q \le \infty$,*

$$\|\boldsymbol{x}\|_{(p,mean)} \le \|\boldsymbol{x}\|_{(q,mean)}.$$

- Choose $\|\cdot\|_{(p,\mathrm{mean})\to(q,\mathrm{mean})}$ with $p \le q$.

# Width Independent Lipschitz Bound

**Theorem (Lipschitz bound in mean normalized geometry)**

*Assume* $1 \le p \le q < \infty$*, bounded inputs, and bounded first derivatives of* $\sigma$ *and* $\mathcal{L}$*. Define*

$$\Omega_C := \{\Theta : \|\Theta\|_{\text{block}} \le C\},$$

*where*

$$\|\Theta\|_{\text{block}} := \max\left\{ \|W_1\|_{1\to(q,\text{mean})}, \max_{2\le i\le \ell-1} \|W_i\|_{(p,\text{mean})\to(q,\text{mean})}, \|W_\ell\|_{(p,\text{mean})\to\infty}, \max_i \|b_i\|_\infty \right\}.$$

*Then the loss is M Lipschitz on* $\Omega_C$*, where M is independent of the width.*

**Implication**

The forward sensitivity of the network remains controlled as width grows.

# Width Independent Smoothness

**Definition ($L$ smoothness)**

With respect to a norm $\|\cdot\|$, the gradient satisfies

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

**Theorem (Width dependence of the smoothness constant)**

*Under the preceding boundedness assumptions, with bounded first and second derivatives of $\sigma$ and $\mathcal{L}$,*

$$\|\nabla f(\Theta) - \nabla f(\Theta')\|_{\mathrm{block},*} \leq L\, w^{\max\{0,\, 2/q - 1/p\}} \|\Theta - \Theta'\|_{\mathrm{block}},$$

*where $L$ is independent of width.*

**Width independent smoothness holds when $q \geq 2p$.**

# Width Independent Smoothness

**Definition ($L$ smoothness)**

With respect to a norm $\|\cdot\|$, the gradient satisfies

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

**Theorem (Width dependence of the smoothness constant)**

*Under the preceding boundedness assumptions, with bounded first and second derivatives of $\sigma$ and $\mathcal{L}$,*

$$\|\nabla f(\boldsymbol{\Theta}) - \nabla f(\boldsymbol{\Theta}')\|_{\mathrm{block},*} \leq L\, m^{\max\{0,\, 2/q - 1/p\}} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_{\mathrm{block}},$$

*where $L$ is independent of width.*

Width independent smoothness holds when $q \geq 2p$.

# Width Independent Smoothness

**Definition ($L$ smoothness)**

With respect to a norm $\|\cdot\|$, the gradient satisfies

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

**Theorem (Width dependence of the smoothness constant)**

*Under the preceding boundedness assumptions, with bounded first and second derivatives of $\sigma$ and $\mathcal{L}$,*

$$\|\nabla f(\boldsymbol{\Theta}) - \nabla f(\boldsymbol{\Theta}')\|_{\mathrm{block},*} \leq L\, m^{\max\{0,\, 2/q - 1/p\}} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|_{\mathrm{block}},$$

*where $L$ is independent of width.*

**Width independent smoothness holds when $q \geq 2p$.**

# Why the Condition $q \geq 2p$ Appears

Featurewise nonlinearities produce a Hadamard product in the quadratic Taylor term.

**Lemma (Mean normalized Hadamard inequality)**

*For* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ *and* $p, q \geq 1$,

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

Consequently,

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

**Operator control**

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

# Why the Condition $q \geq 2p$ Appears

Featurewise nonlinearities produce a Hadamard product in the quadratic Taylor term.

**Lemma (Mean normalized Hadamard inequality)**

*For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $p, q \geq 1$,*

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

Consequently,

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

**Operator control**

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

# Width Aware Learning Rate Scaling

## Matrix Operator Geometry Aware (MOGA) Scaling

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

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

- At the algebraic endpoint for `Adam` ($\ell_1 \to \ell_\infty$), MOGA recovers the $\mu$P factor $\boxed{1/\texttt{d\_in}}$.
- For `Muon` ($\ell_2 \to \ell_2$), MOGA recovers the $\mu$P factor $\boxed{\sqrt{\texttt{d\_out}/\texttt{d\_in}}}$.
  - The spectral condition of Yang et al. (2023) imposes the same update scale.
  - When $p = q = 2$, $L = \mathcal{O}(\sqrt{w})$; a stable step size therefore scales as $\mathcal{O}(1/\sqrt{w})$, matching Moonlight.

# MOGA Optimizer

**Part I** Initialization and column geometry

**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 $\boldsymbol{\Theta}^0$

$\boldsymbol{M}^0 \leftarrow \mathbf{0}$ ▷ $\boldsymbol{M}^t$ 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 $\ell$ **do**

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

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

    **end for**

# MOGA Optimizer

**Part II** Row geometry and loop closure

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

  **for** $i = 1$ to $\ell$ **do**

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

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

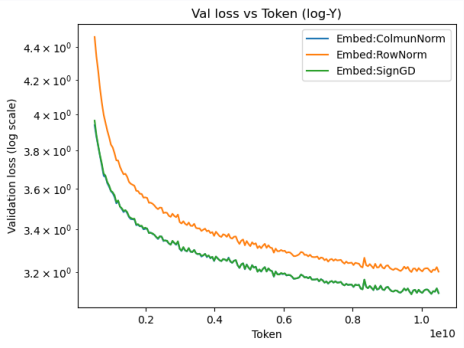
  **end for**

**end if**

**end for**

# Adapting to Transformers

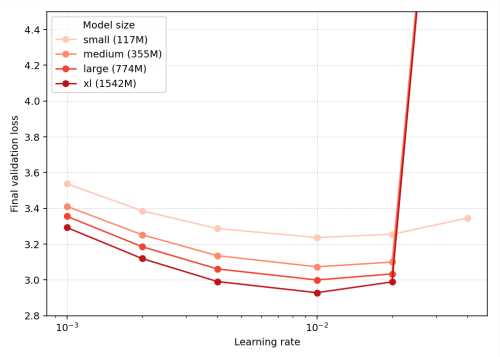
- **Linear layers in MLPs and attention:** optimized using MOGA with scaling.
- **Biases and LayerNorm parameters:** optimized using SignGD.
- **Input word and positional embeddings:** map one-hot vectors (with $\ell_1$ geometry) to feature representations (with $(q, \mathrm{mean})$ geometry).
- **Word unembedding:** the unembedding matrix is tied to the word embedding matrix.



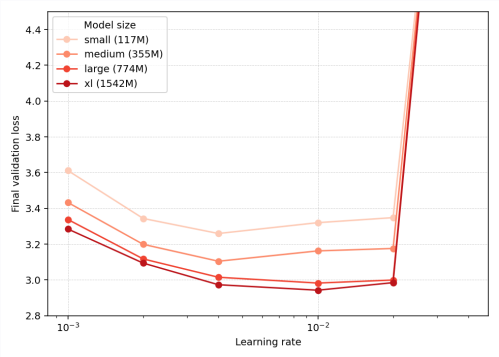

Embedding and output layer comparison

# Learning Rate Transfer Across Model Scales

**Architecture**: GPT-2 family (Small $\rightarrow$ XL)



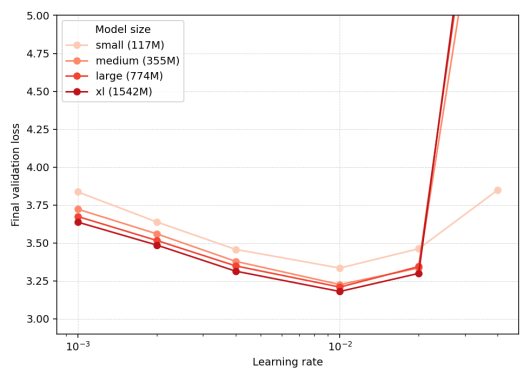

**(a)** MOGA ($p = 1.5$)




**(b)** MOGA ($p = 2$)

The optimal learning rate remains stable across model widths.

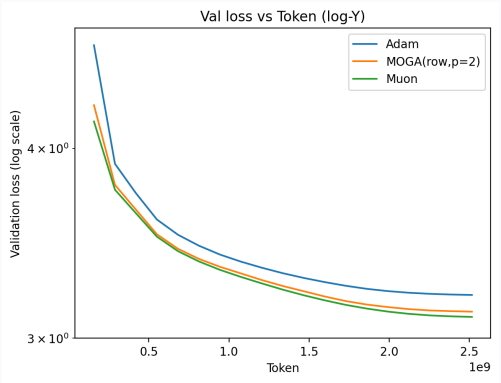
# Learning Rate Transfer Across Model Scales



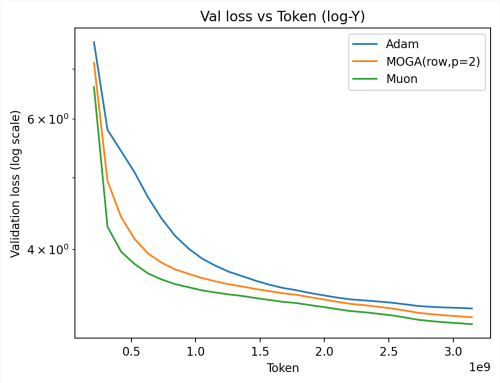

**(c)** MOGA ($p = 3$)

**Learning rate transfer.** The optimal learning rate remains stable from GPT-2 Small to XL.

# LLM Pretraining: Standard Token Budget



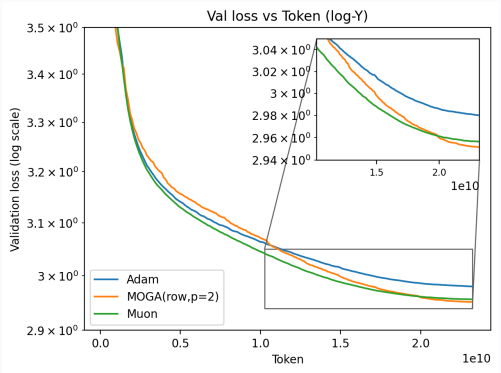

LLaMA 130M



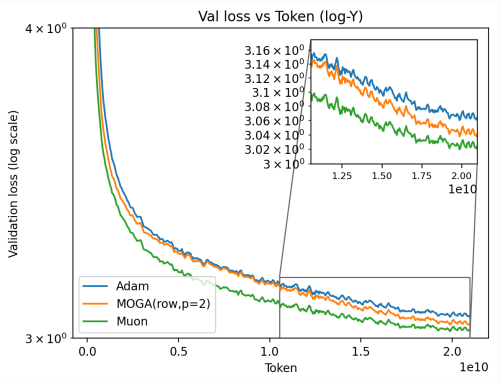

GPT-2 Small model

**MOGA converges at nearly the same rate as Muon under a $1\times$ Chinchilla budget.**

# LLM Pretraining: Large Token Budget




LLaMA 130M




GPT-2 Small model

**MOGA converges faster late in training, when the loss is small.**

# Related Work on Row and Column Normalization

- Training Deep Learning Models with Norm-Constrained LMOs, arXiv:2502.07529
- A Minimalist Optimizer Design for LLM Pretraining, arXiv:2506.16659
- Mano: Rethinking 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 essential for optimizer design at LLM scale.
- **Matrix operator geometry** unifies SignSGD, AdamW, Muon, and normalized matrix updates.
- **Mean normalized geometry** yields width independent Lipschitz and smoothness bounds when $q \geq 2p$, together with width aware learning rate scaling.
- **MOGA** recovers the $\mu$P factor for Adam at $(p, q) = (1, \infty)$ and transfers effectively across model widths.

**Future Direction:** Is there a norm unifies approxiamtion, generalization and optimization?

# Take Home Messages

- **Hyperparameter transfer** is essential for optimizer design at LLM scale.
- **Matrix operator geometry** unifies SignSGD, AdamW, Muon, and normalized matrix updates.
- **Mean normalized geometry** yields width independent Lipschitz and smoothness bounds when $q \geq 2p$, together with width aware learning rate scaling.
- **MOGA** recovers the $\mu$P factor for Adam at $(p, q) = (1, \infty)$ and transfers effectively across model widths.

**Future Direction:** Is there a norm unifies approxiamtion, generalization and optimization?

# Take Home Messages

- **Hyperparameter transfer** is essential for optimizer design at LLM scale.
- **Matrix operator geometry** unifies SignSGD, AdamW, Muon, and normalized matrix updates.
- **Mean normalized geometry** yields width independent Lipschitz and smoothness bounds when $q \geq 2p$, together with width aware learning rate scaling.
- **MOGA** recovers the $\mu$P factor for Adam at $(p, q) = (1, \infty)$ and transfers effectively across model widths.

**Future Direction:** Is there a norm unifies approxiamtion, generalization and optimization?

# Take Home Messages

- **Hyperparameter transfer** is essential for optimizer design at LLM scale.
- **Matrix operator geometry** unifies SignSGD, AdamW, Muon, and normalized matrix updates.
- **Mean normalized geometry** yields width independent Lipschitz and smoothness bounds when $q \geq 2p$, together with width aware learning rate scaling.
- **MOGA** recovers the $\mu$P factor for Adam at $(p, q) = (1, \infty)$ and transfers effectively across model widths.

**Future Direction:** Is there a norm unifies approxiamtion, generalization and optimization?

# Appendix: Spectral Condition

- **Feature scale**

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

- **Parameter scale**

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

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


# Thank you

## Questions?

**Yiping Lu**

Beijing International Center for Mathematical Research
Peking University

Industrial Engineering & Management Science ,
Northwestern University