Spectral Condition for μP under Width-Depth Scaling

Chenyu Zheng, Rongzhen Wang, Xinyu Zhang, Chongxuan Li

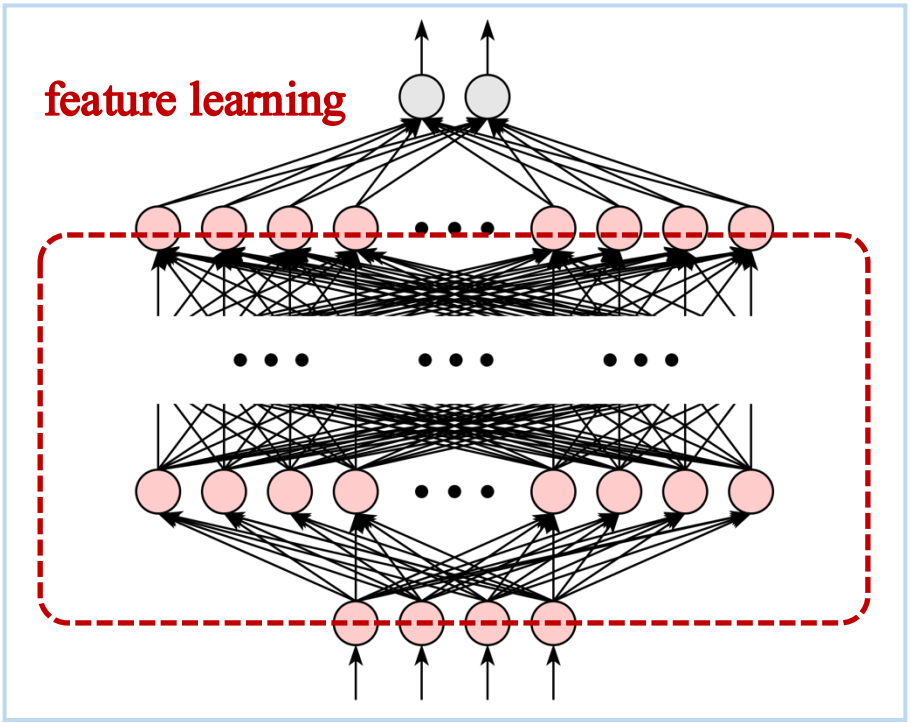
Renmin University of China & ByteDance Seed

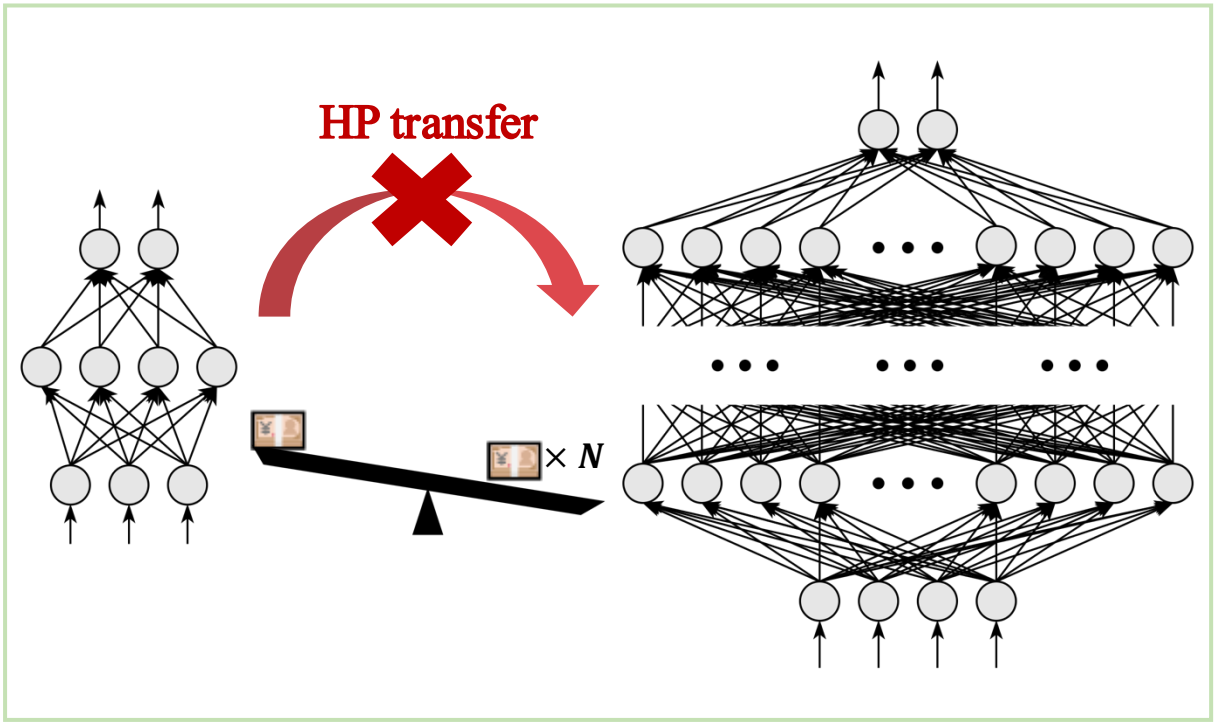
Contents

1.1 Motivation of μP

Feature learning becomes unstable or degenerate in infinite limit

Hyperparameter (HP) tuning is expensive and HP transfer from small models is difficult





1.2 Principle of μP——Stable and Fast Training

- μP desires the parameter update to be
  - Stable: preserve scale-invariant feature learning
  - Fast: maximize the feature change

Assume $h_l \in R^{n_l}$ is the hidden feature, $\Delta h_l \in R^{n_l}$ is its one-step update, μP desires:

$$\|h_l\|_{RMS} = \Theta(1), \|\Delta h_l\|_{RMS} = \Theta(1), \qquad l \in [L],$$

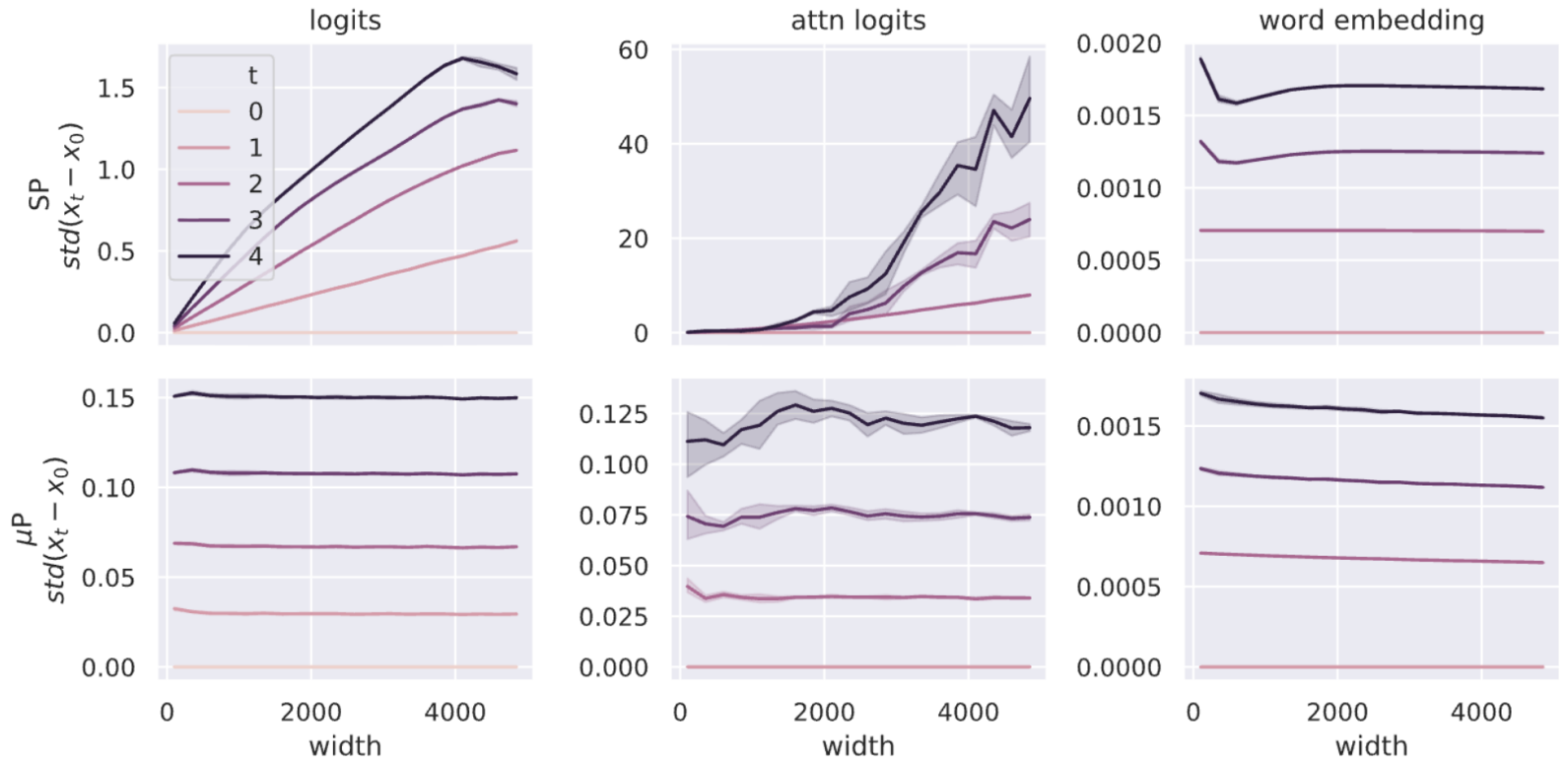
$$maximize\ \Delta W_l's\ contribution\ to\ \Delta h_L, \qquad l \in [L],$$

where $\|v\|_{RMS} = \frac{\|v\|_2}{\sqrt{d}} = \sqrt{\frac{\sum v_i^2}{d}}$.

Tensor Programs IV: Feature Learning in Infinite-Width Neural Network, ICML, 2021

1.3 Success of μP under Width Scaling

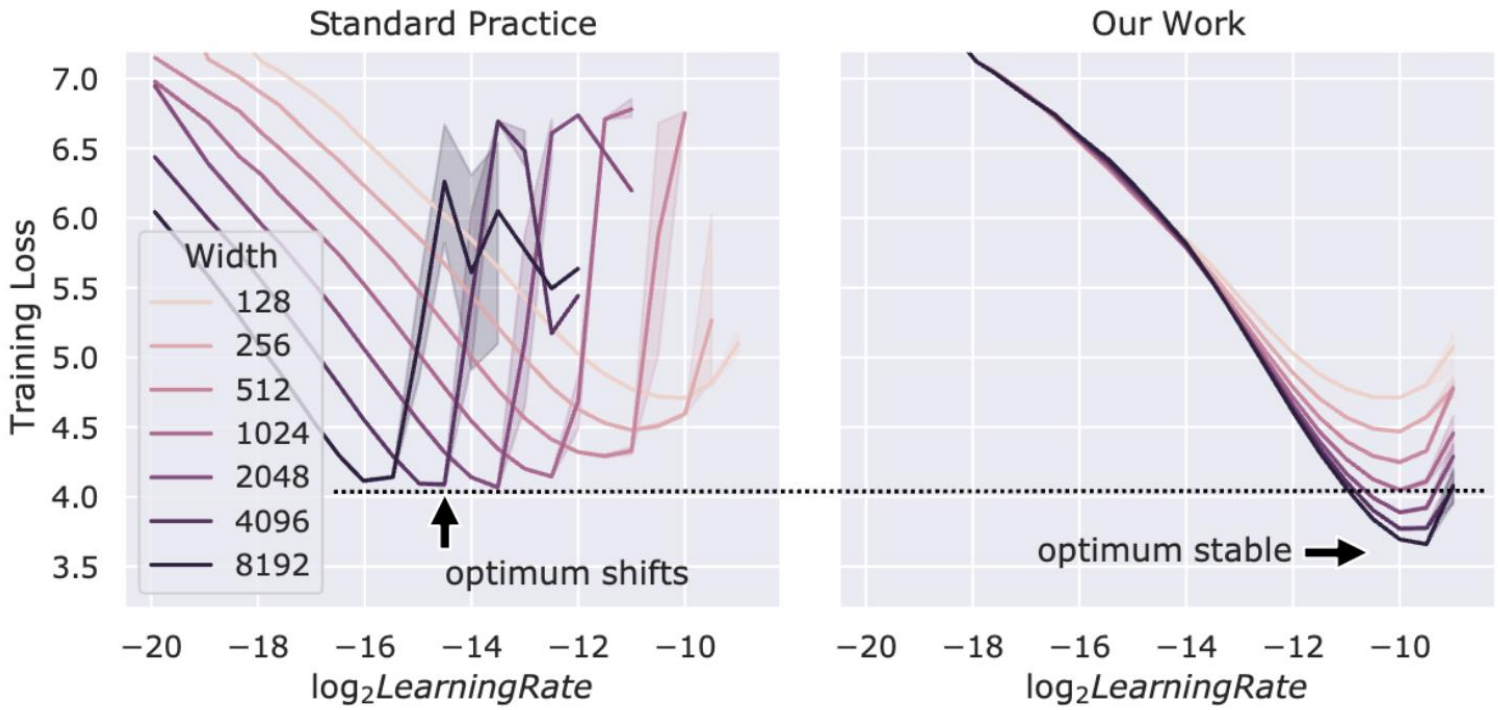
- μP maintains stable feature norms across different widths



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

1.3 Success of μP under Width Scaling

- μP enables the robust HP transfer from narrow to wide models



Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, NeurIPS, 2022

1.4 Spectral Condition for μP under Width Scaling

- Minimal theoretical setup: linear MLP ($h_l = W_l h_{l-1}$), one datapoint, one update

Spectral Condition: Assume $W_l \in R^{n_l \times n_{l-1}}$ is the weight matrix in l-th layer, $\Delta W_l \in R^{n_l \times n_{l-1}}$ is its one-step update. When $n \to \infty$, to realize μP principle, we need

$$\|W_l\|_{RMS} = \Theta(1), \|\Delta W_l\|_{RMS} = \Theta(1), \qquad l \in [L],$$

where $\|W\|_{RMS} = \sup_{x \neq 0} \frac{\|Wx\|_{RMS}}{\|x\|_{RMS}} = \sqrt{\frac{n_{l-1}}{n_l}} \|W\|_2$.

- It directly applies to different optimizers like SGD, Adam, Muon
- It generalizes to modern architectures like Transformer

A Spectral Condition for Feature Learning, arxiv, 2023

1.5 Extending μP to Width-Depth Scaling

- Existing studies on μP under width-depth scaling are **disparate**
  - Different architectures: $h_{l+1} = h_l + \alpha_l F_l(h_l)$, different $F_l(h_l)$
  - Specified optimizer: SGD or Adam
  - Complex derivation: Tensor Program or DMFT
  - Different results in different cases

1.5 Extending μP to Width-Depth Scaling

- Existing studies on μP under width-depth scaling are **disparate**
  - Different architectures: $h_{l+1} = h_l + \alpha_l F_l(h_l)$, different $F_l(h_l)$
  - Specified optimizer: SGD or Adam
  - Complex derivation: Tensor Program or DMFT
  - Different results in different cases

Q: Can we establish a simple and unified theory under width-depth scaling?

Contents

2.1 Theoretical Setup

- Compared to setup under width scaling, we additionally introduce:
  - Residual connection: essential for depth scaling
  - Residual block with finite k layers: modern models have deep blocks (e.g., FFN)

Multi-layer FFN with residual connection:

$$h_0 = \alpha_0 W_0 x,$$

$$h_l = h_{l-1} + \alpha_l \prod_{i=1}^{k} W_l^{(i)} \, h_{l-1}, \forall \, l \in [L],$$

$$h_{L+1} = \alpha_{L+1} W_{L+1} h_L,$$

with width n, depth $L \to \infty$.

2.2 Spectral Condition when $k = 1$

- Consider $h_l = h_{l-1} + \alpha_l W_l h_{l-1}$ (one-layer residual block)

Initialization condition:

- Input and output layers: $\alpha_0 \|W_0\|_{RMS}, \alpha_{L+1} \|W_{L+1}\|_{RMS} = \Theta(1)$
- Hidden layers: $\alpha_l \|W_l\|_{RMS} = O\left(\frac{1}{\sqrt{L}}\right), \forall\, l \in [L]$

Update condition:

- Input and output layers : $\alpha_0 \|\Delta W_0\|_{RMS}, \alpha_{L+1} \|\Delta W_{L+1}\|_{RMS} = \Theta(1)$
- Hidden layers: $\alpha_l \|\Delta W_l\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$

- Since $\|W_l\|_{RMS} = \Theta(1)$ according to width-scaling μP, we have $\alpha_l = O(1/\sqrt{L})$, which recovers Depth-μP

2.3 Spectral Condition when $k = 2$ (Phase Transition)

- Consider $h_l = h_{l-1} + \alpha_l W_l^{(2)} W_l^{(1)} h_{l-1}$ (two-layer residual block)
- Essential difference from $k = 1$: high-order update term

$$\Delta \boldsymbol{h}_s(\boldsymbol{x}) = \Delta \boldsymbol{h}_0(\boldsymbol{x}) + \underbrace{\sum_{l=1}^{s} \alpha_l \boldsymbol{W}_l^{(2)} \boldsymbol{W}_l^{(1)} \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x})}_{\epsilon_0(s)} + \underbrace{\sum_{l=1}^{s} \alpha_l \boldsymbol{W}_l^{(2)} \Delta \boldsymbol{W}_l^{(1)} (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\epsilon_1^{(1)}(s)}$$

$$+ \underbrace{\sum_{l=1}^{s} \alpha_l \Delta \boldsymbol{W}_l^{(2)} \boldsymbol{W}_l^{(1)} (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\epsilon_1^{(2)}(s)} + \underbrace{\sum_{l=1}^{s} \alpha_l \Delta \boldsymbol{W}_l^{(2)} \Delta \boldsymbol{W}_l^{(1)} (\boldsymbol{h}_{l-1}(\boldsymbol{x}) + \Delta \boldsymbol{h}_{l-1}(\boldsymbol{x}))}_{\epsilon_2(s)}.$$

2.3 Spectral Condition when $k = 2$ (Phase Transition)

Initial condition:

- Hidden layers: $\alpha_l \left\| W_l^{(2)} \right\|_{RMS} \left\| W_l^{(1)} \right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$

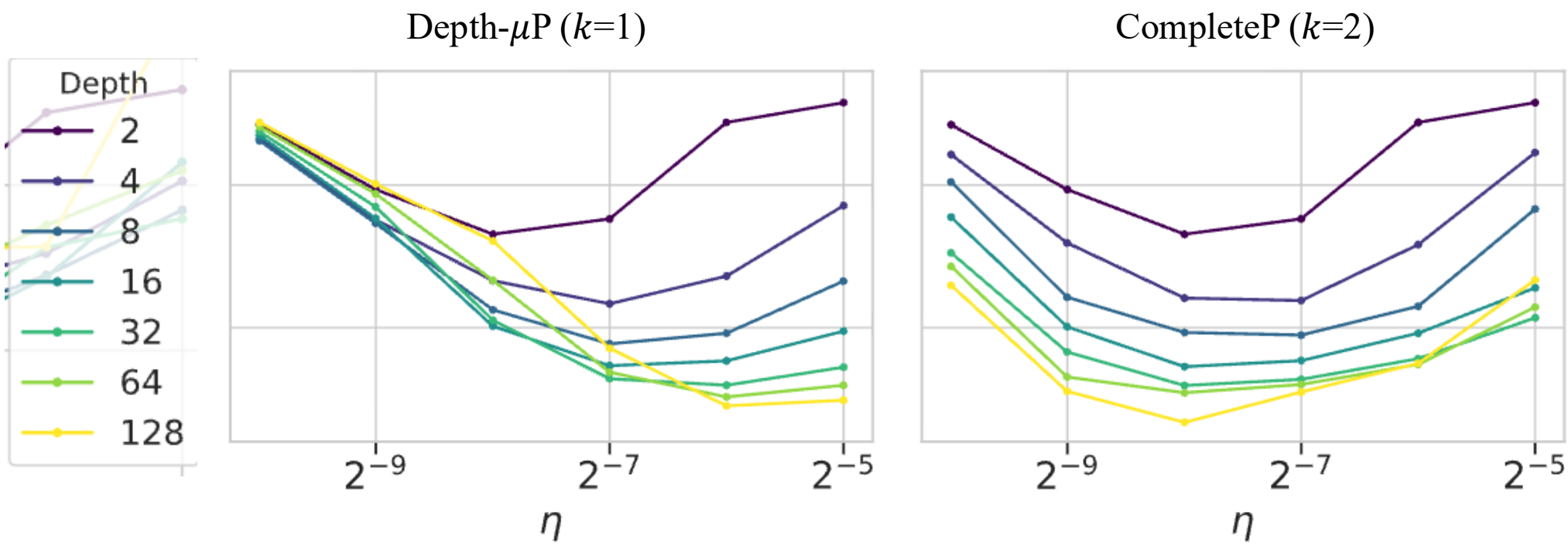
Update Condition:

- Hidden layers (1st-order): $\alpha_l \left\| \Delta W_l^{(2)} \right\|_{RMS} \left\| W_l^{(1)} \right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$
- Hidden layers (1st-order): $\alpha_l \left\| W_l^{(2)} \right\|_{RMS} \left\| \Delta W_l^{(1)} \right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$
- Hidden layers (2nd-order): $\alpha_l \left\| \Delta W_l^{(2)} \right\|_{RMS} \left\| \Delta W_l^{(1)} \right\|_{RMS} = \Theta\left(\frac{1}{L}\right), \forall\, l \in [L]$

- Since $\|W_l\|_{RMS} = \Theta(1)$, we have $\alpha_l = \Theta(1/L)$, which recovers CompleteP
- $k = 2$ is the minimal meaningful setting: results when $k \geq 2$ are the same

2.4 $k = 2$ is Minimal Meaningful Setting

- Previous GPT-2 results: CompleteP ($k = 2$) outperforms Depth-μP ($k = 1$)



Don't be lazy: CompleteP enables compute-efficient deep transformers, NeurIPS, 2025

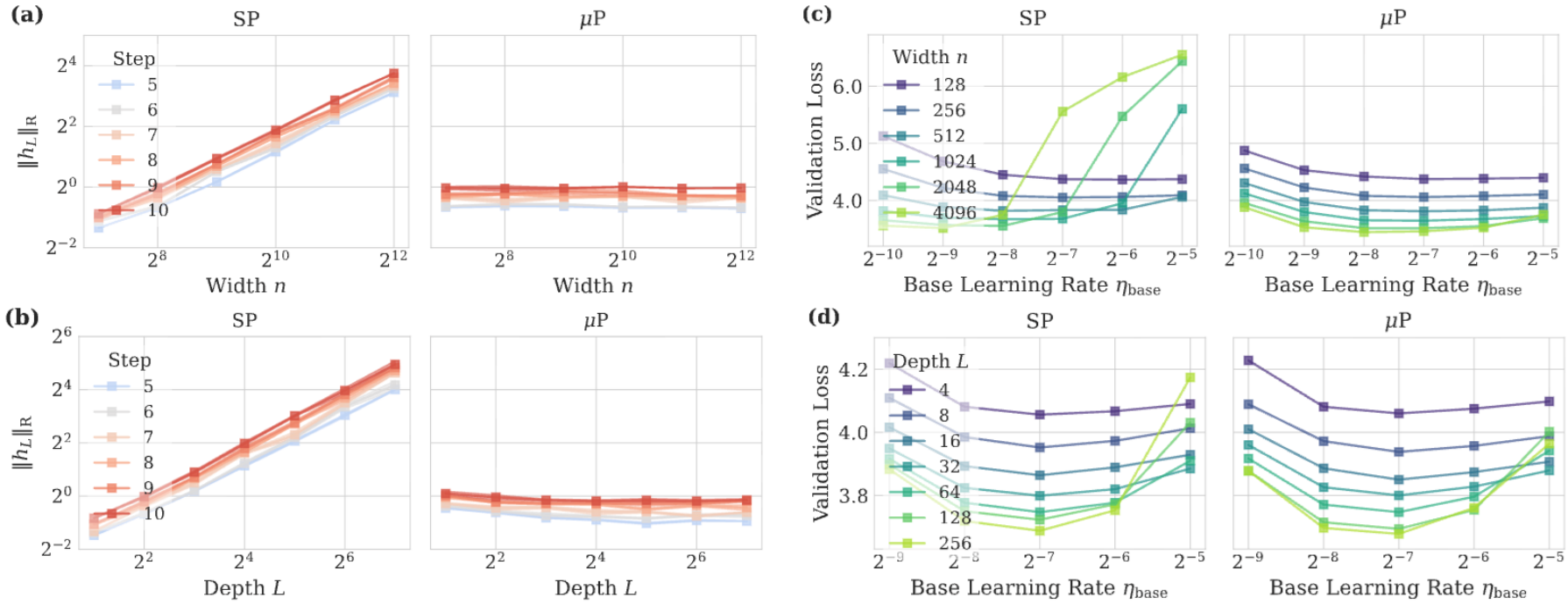
2.5 Implementation of μP under Width-Depth Scaling

- For modern optimizers
  - Muon-Kimi, Muon, Shampoo, SOAP, AdamW, Sophia, Lion, SSO, etc
- **Just set $\alpha_l = \Theta(1/L)$ in additional to the implementation of width-scaling μP!**

Contents

Feature Learning and HP Transfer

- For Muon-Kimi, μP enables stable feature learning and robust HP transfer



- More results will be released soon…

Contents

Conclusion

- **Spectral condition**: a unified spectral μP condition for width-depth scaling
- **Minimal theoretical setup**: residual block depth of $k = 2$
- **Implementation**: just set $\alpha_l = \Theta(1/\mathrm{L})$ in additional to width-scaling μP
- **Empirical validation**: stable feature learning and robust HP transfer