Global Convergence and Rich Feature Learning in L-Layer Infinite-Width Neural Networks under μP Parametrization

Zixiang Chen*, Greg Yang*, Qingyue Zhao, Quanquan Gu

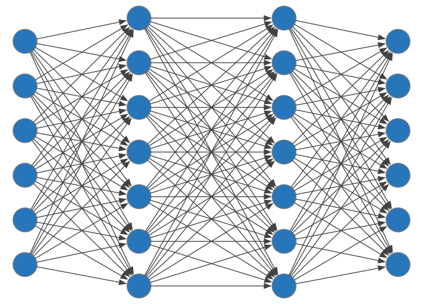
Can deep L-layer neural networks simultaneously achieve:

✅ Meaningful Feature Learning

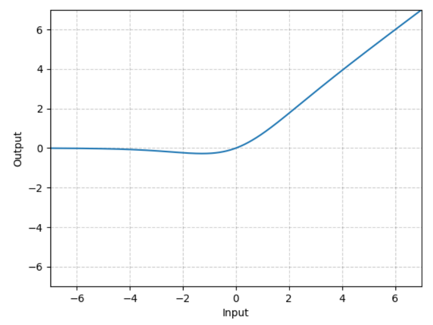
✅ Global Convergence Guarantees

Challenges

Update Rule: $W_t^l = W_{t-1}^l + \Delta W_t^l$



Multiple Layer



Non Linear Activation

[Foret et al, 2020]

Non-convex Optimization

Background & Motivation

SP - Standard Parametrization

NTP - Neural Tangent Kernel

IP - Integrable Parametrization

µP - Maximal Update Parametrization

Layer	SP		NTP		IP		μP	
	Init. Var.	LR	Init. Var.	LR	Init. Var.	LR	Init. Var.	LR
Input (W^1)	1	$\eta \cdot n^{-1}$	1	η	1	$\eta \cdot n$	1	$\eta \cdot n$
Hidden (W^l)	n^{-1}	$\eta \cdot n^{-1}$	n^{-1}	$\eta \cdot n^{-1}$	n^{-2}	η	n^{-1}	η
Output (W^{L+1})	n^{-1}	$\eta \cdot n^{-1}$	n^{-1}	$\eta \cdot n^{-1}$	n^{-2}	$\eta \cdot n^{-1}$	n^{-2}	$\eta \cdot n^{-1}$

Parametrization	Feature Learning	Feature Richness
Standard (SP)	✗	Rich
Neural Tangent (NTP)	✗	Rich
Meanfield (IP)	✓	Low
Maximal Update (μP)	✓	Rich

Neural Networks Setup

L-layer MLP under µP with input ξ:

$$h^1 = W^1\xi, \quad x^l = \phi(h^l), \quad h^{l+1} = W^{l+1}x^l, \quad f(\xi) = W^{L+1}x^L$$

GOOD Activation ϕ (details omitted)**:** Sigmoid, Tanh, SiLU, GeLU

The Limit to Infinite-width Networks

Represent features and weights via "Z" random variables:

$$h^l \rightarrow Z^{h^l(\xi)} \qquad x^l \rightarrow Z^{x^l(\xi)} \qquad W^{L+1} \rightarrow Z^{\widehat{W}^{L+1}} \qquad f(\xi) \rightarrow \mathring{f}(\xi)$$

Example: each entry of h^l behave like i.i.d. copies of the random variables $Z^{h^l(\xi)}$.

Problem Setup

Data Assumptions: Input $\xi \in S$, with $|\langle \xi_i, \xi_j \rangle| \neq |\langle \xi_i, \xi_k \rangle|$ and for distinct points $|\langle \xi_i, \xi_j \rangle| \neq 0$ for any three different points $\xi_i, \xi_j, \xi_k \in S$.

Remark: It holds with probability 1 if S are drawn from some continuous distribution like mixture Gaussian.

Error Signal: error signal $\mathring{\chi}_{t,i}$ at time step t for the i-th sample. When training with SGD to minimize the loss function L this error signal is computed as $\mathring{\chi}_{t,i} = L'(\mathring{f}_t, y_i)$, where $\mathring{f}$ is the model output and y is the label

Example: the error signal of square loss is $\mathring{\chi}_{t,i} = 2(\mathring{f}_t(\xi_i) - y_i)$.

Gradient Descent with Error Signal (last layer example)

$$Z^{\widehat{W}_t^{L+1}} = Z^{\widehat{W}_0^{L+1}} + Z^{\delta W_1^{L+1}} + \cdots + Z^{\delta W_t^{L+1}}$$

$$Z^{\delta W_t^{L+1}} = -\eta \sum \mathring{\chi}_{t-1,i} Z^{x_{t-1}^L(\xi_i)}$$

Main Results

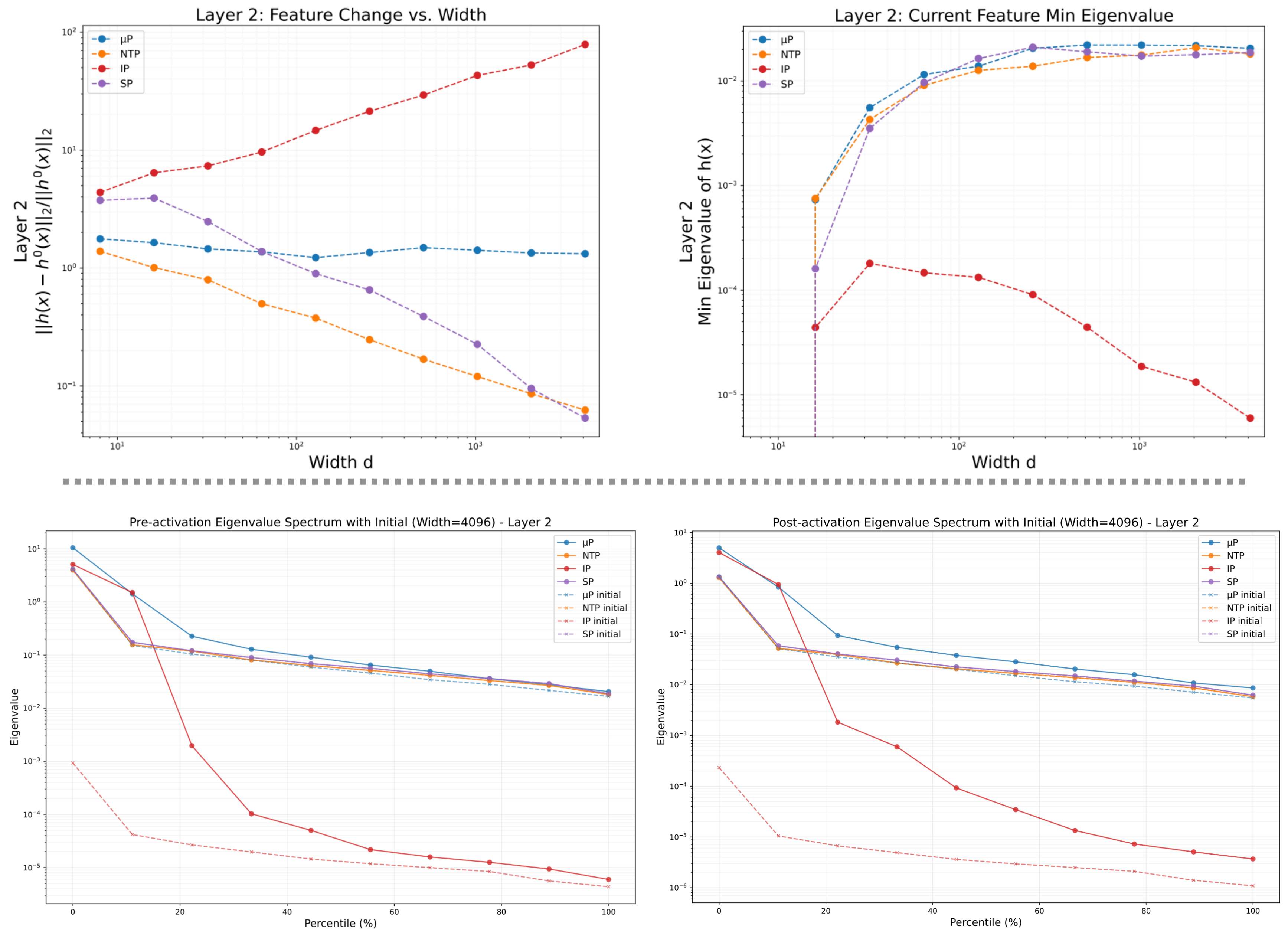
Theorem (Feature Richness). For infinite-width L-layer MLP under µP, following features remain linearly independent:

Pre-activation: $Z^{h_t^l(\xi)}$, Post-activation: $Z^{x_t^l(\xi)}$.

Corollary (Global Convergence). If the model converges at time T, then the error signal vanishes and indicates global minimum.

Why? Linear independence eliminates local minima.

Empirical Results



*: equal contribution