上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

SJTU

Condensation Sheds Light on the Mathematical Foundation of Deep Neural Networks

Yaoyu Zhang

Institute of Natural Sciences & School of Mathematical Sciences

Shanghai Jiao Tong University

The Mathematics of Scientific Machine Learning and Digital Twins, Erice

饮 水 思 源 • 爱 国 荣 校

Central issue in DL theory: generalization puzzle

1995

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

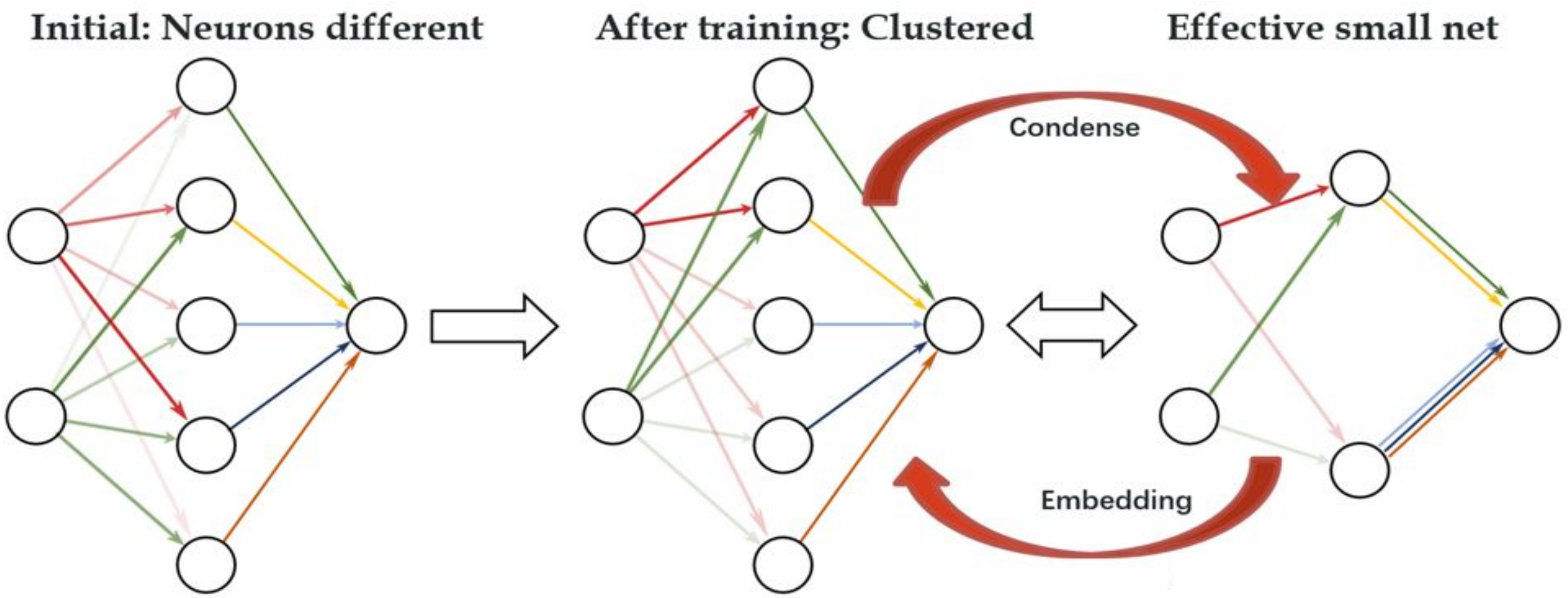
For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

generalization puzzle

Condensation phenomenon

Illustration of Condensation



$$f(x) = \sum_{i=1}^{5} a_i \sigma(\boldsymbol{w}_i^T \boldsymbol{x})$$

$$\boldsymbol{w}_1 \approx \boldsymbol{w}_2, \quad \boldsymbol{w}_3 \approx \boldsymbol{w}_4 \approx \boldsymbol{w}_5$$

$$f(x) = (a_1 + a_2)\sigma(\boldsymbol{w}_1^T \boldsymbol{x}) + (a_3 + a_4 + a_5)\sigma(\boldsymbol{w}_3^T \boldsymbol{x})$$

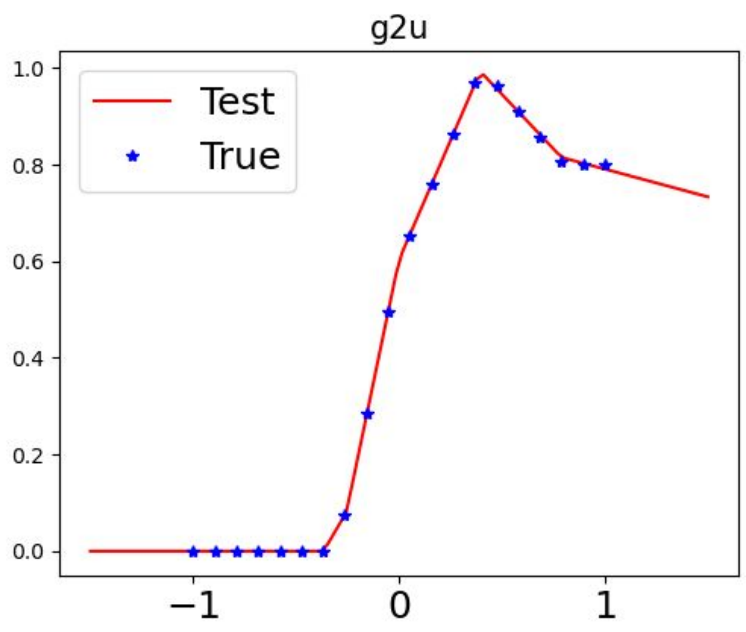
Initial: random

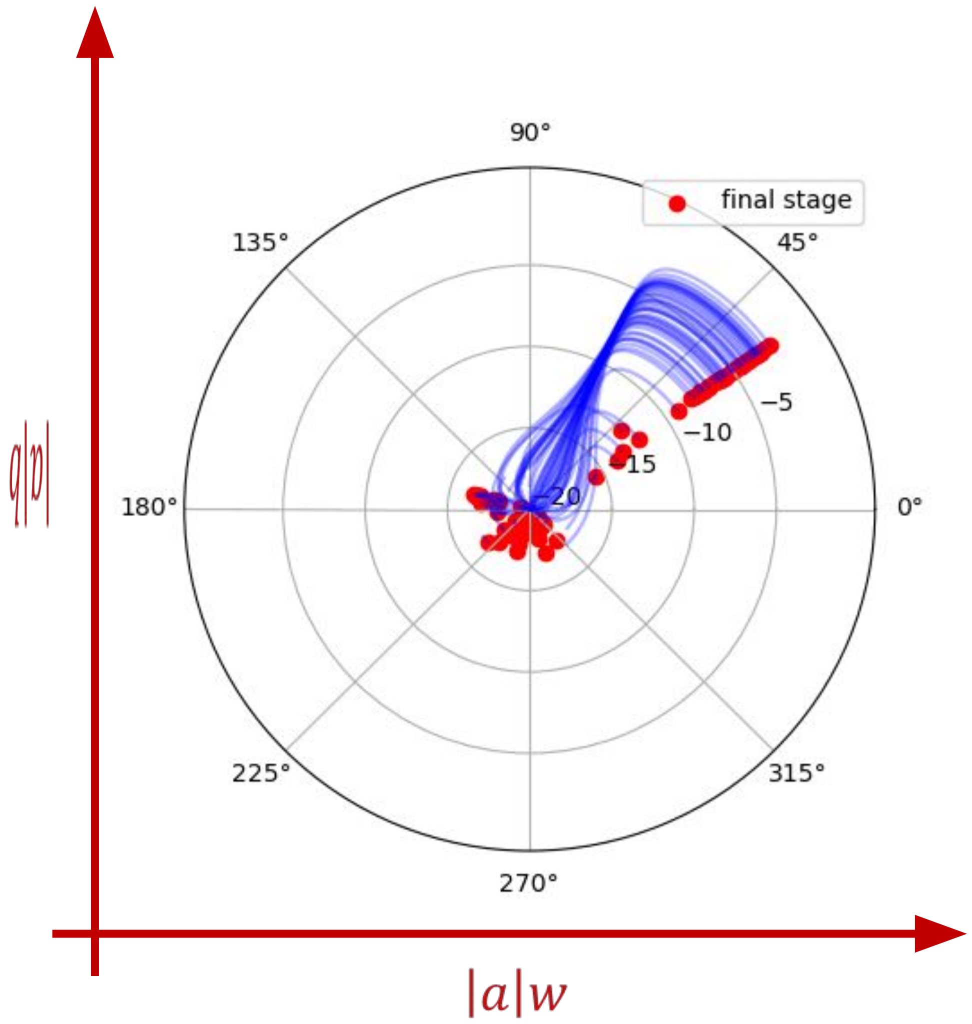
Training: condense

Effect: equiv to small net

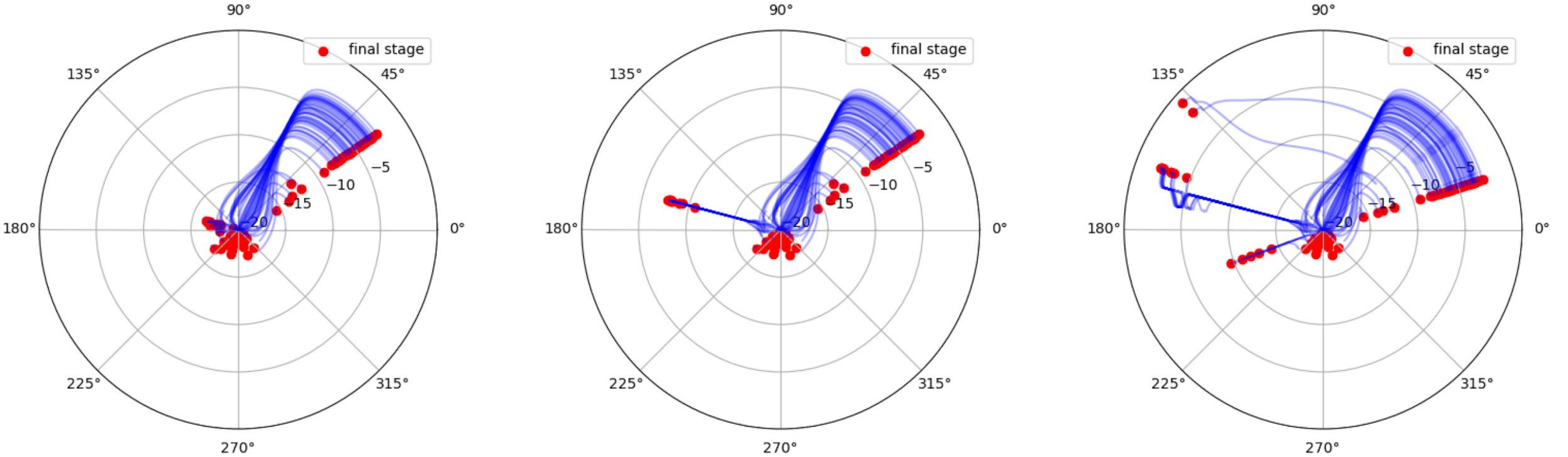
Tao Luo, Zhi-Qin John Xu, Zheng Ma, Yaoyu Zhang,
Phase diagram for two-layer ReLU neural networks at infinite-width limit, Journal of Machine Learning Research (2021)

$$f_\theta(x) = \sum_{j=1}^{m} a_j \,\mathrm{relu}(w_j x + b_j)$$



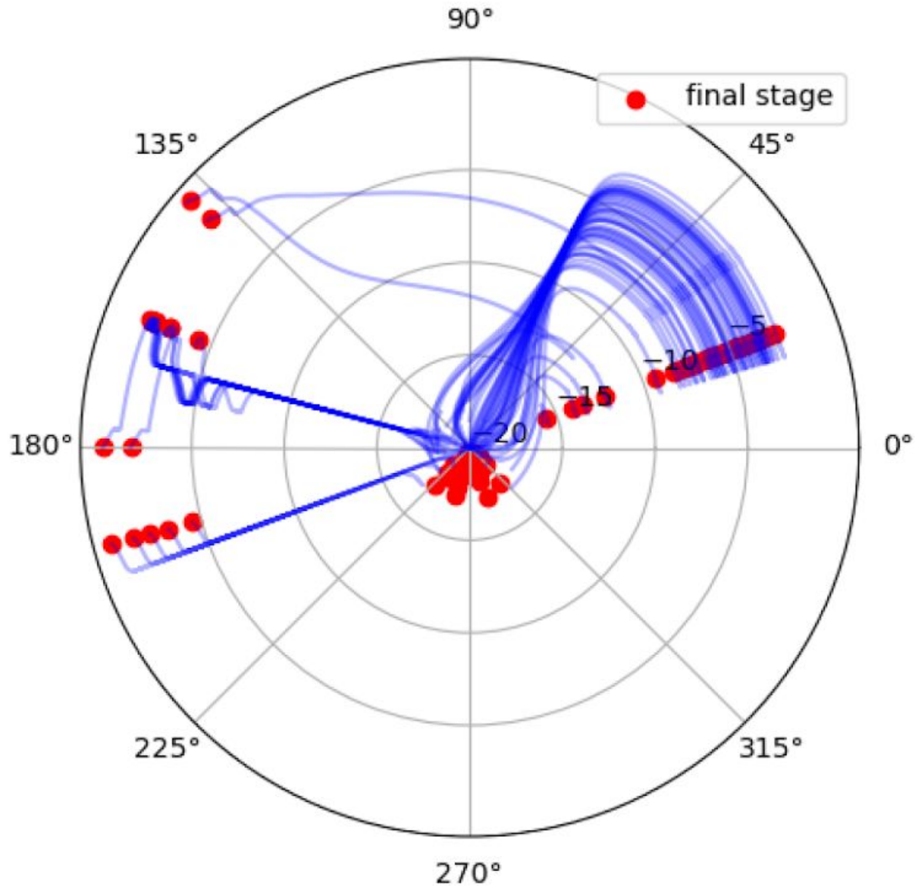


Gradual increase of neuron clusters

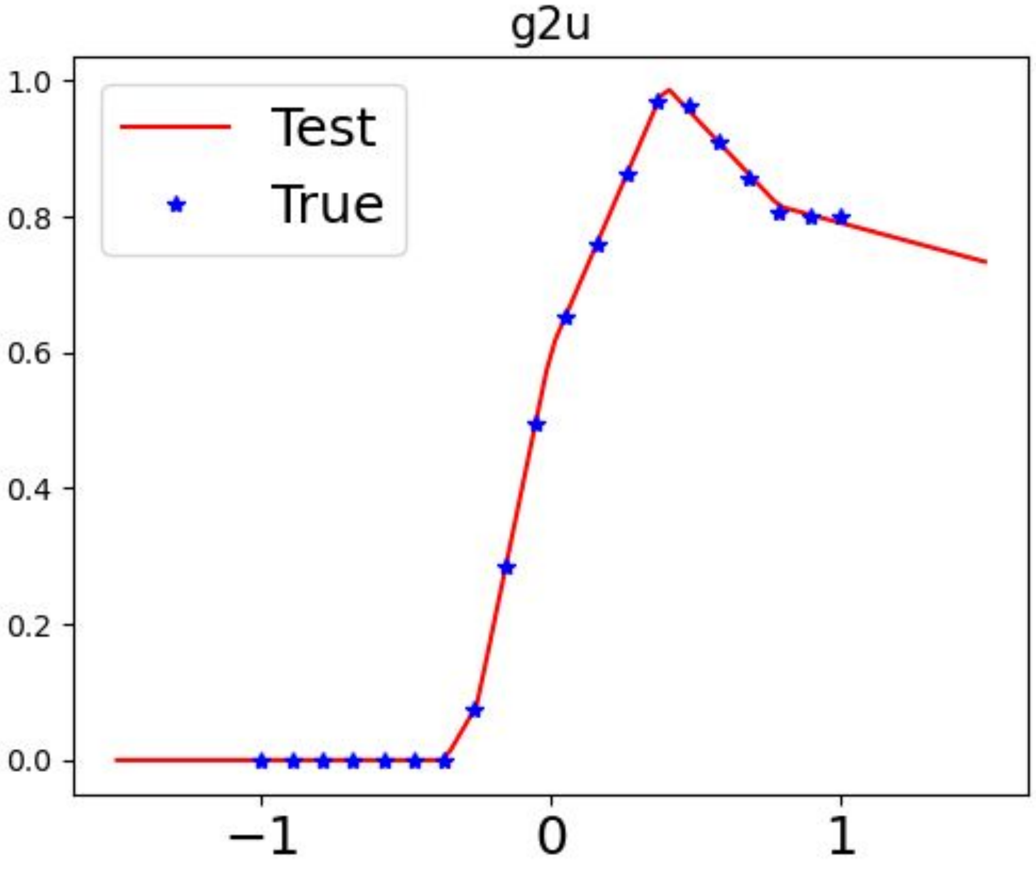


(a) epoch=100 (b) epoch=1000 (c) epoch=3000

Gradual increase of neuron clusters

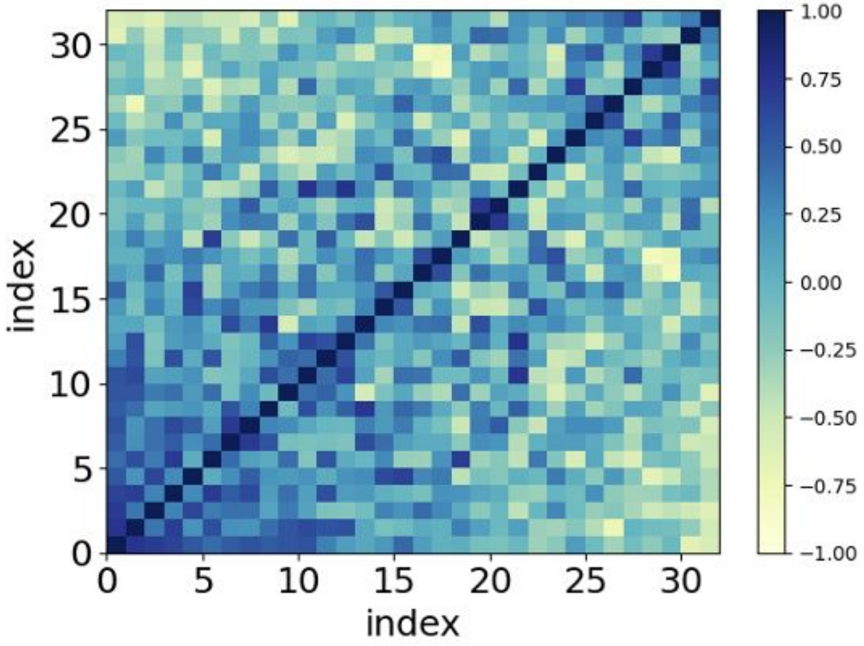


(f) epoch=100000

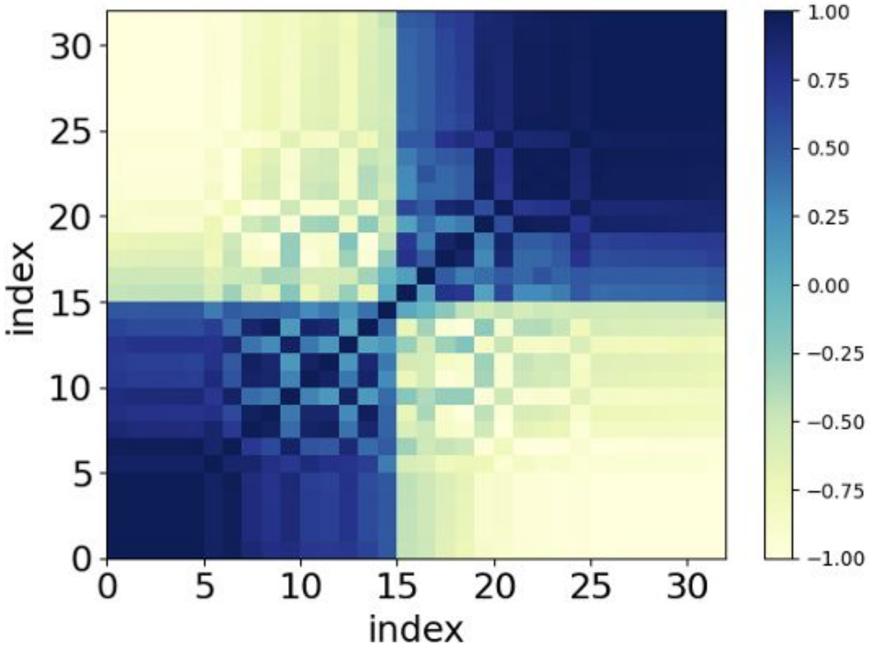


Condensation in CNN on MNIST

Cosine similarity: $$D(\boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{\boldsymbol{u}_1^\intercal \boldsymbol{u}_2}{(\boldsymbol{u}_1^\intercal \boldsymbol{u}_1)^{1/2}(\boldsymbol{u}_2^\intercal \boldsymbol{u}_2)^{1/2}}.$$

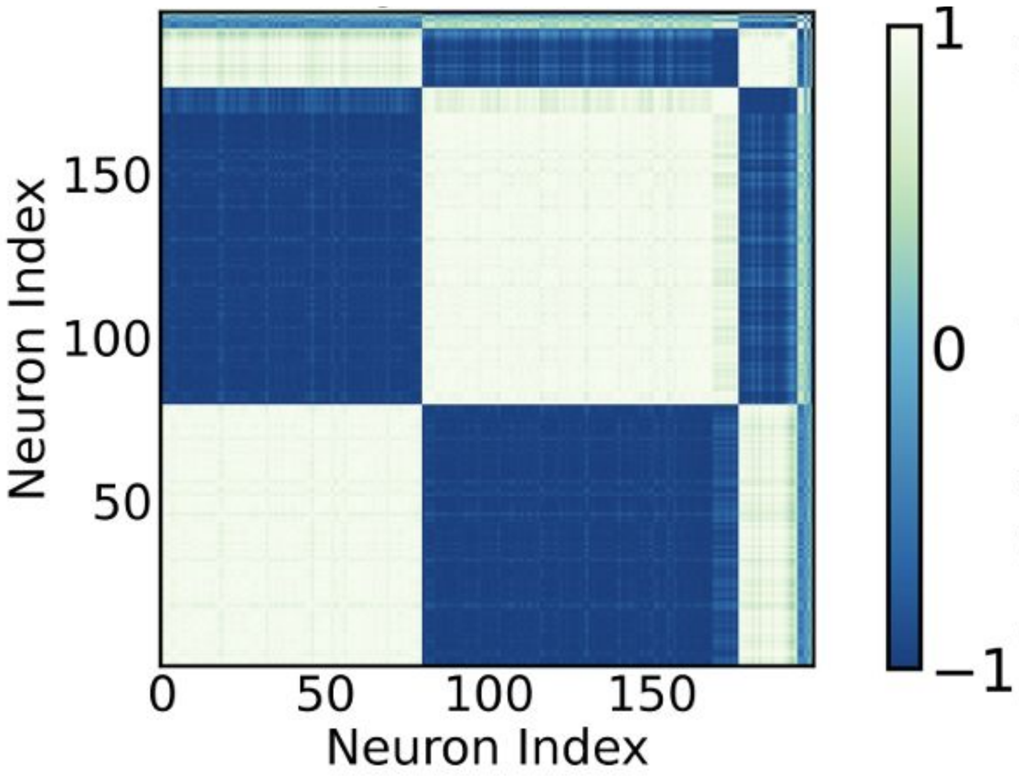


(b) initial weight



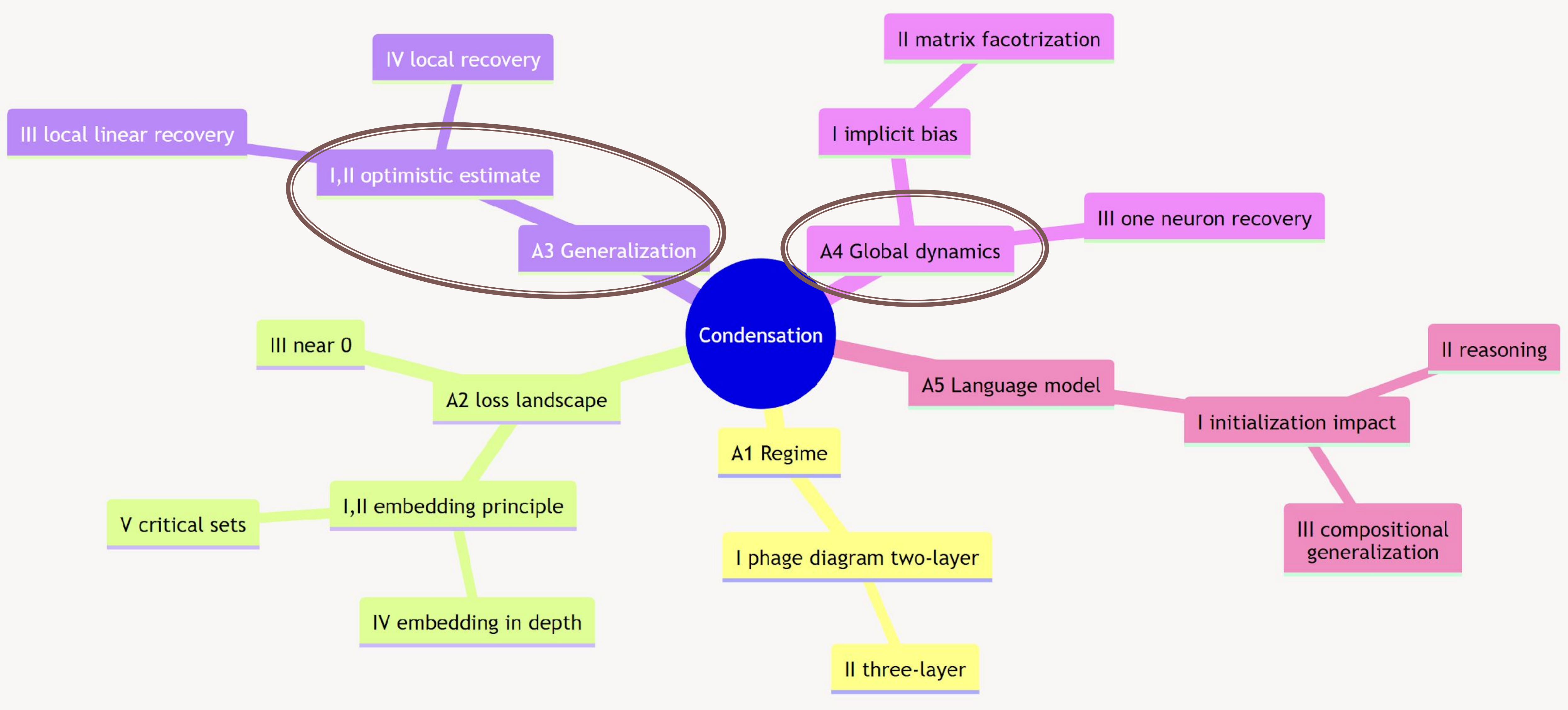
(e) final weight

Condensation in Transformer



$$A_\theta(X) = \sum_{i=1}^{h} \operatorname*{softmax}_{\text{row}} \left(\frac{X W_{Q_i} W_{K_i}^\top X^\top}{\sqrt{d}} \right) X W_{V_i} W_{O_i}^\top$$

Overview of our works on condensation



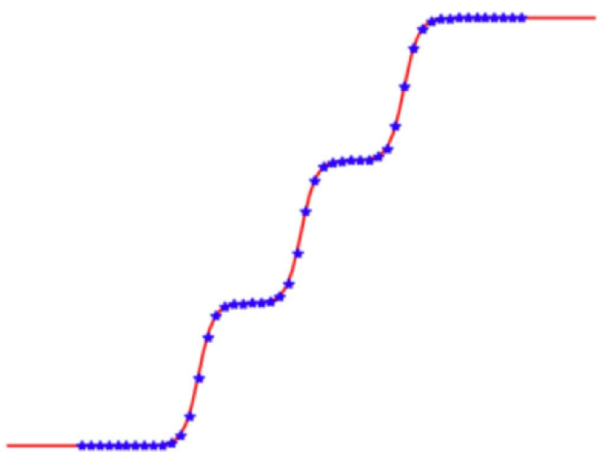
Condensation explains generalization puzzle

Yaoyu Zhang, Zhongwang Zhang, Leyang Zhang, Zhiwei Bai, Tao Luo, Zhi-Qin John Xu, "Optimistic Estimate Uncovers the Potential of Nonlinear Models," Journal of Machine Learning 2025.

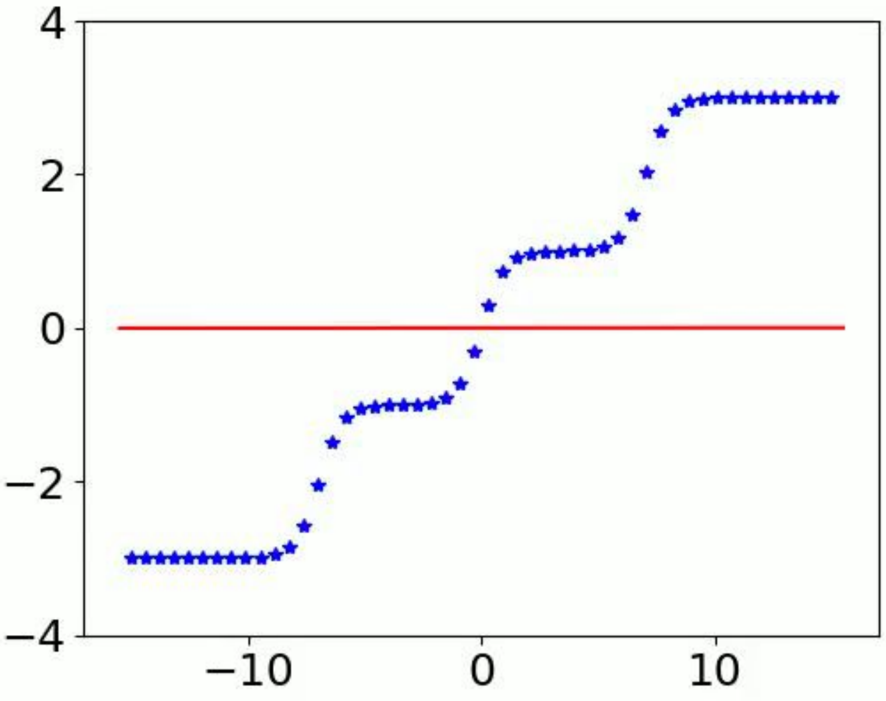
Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. JMLR 2025

Recovery under overparameterization?

Can a 500 neuron network (1500 parameters) recover a target function from 50 sample points?

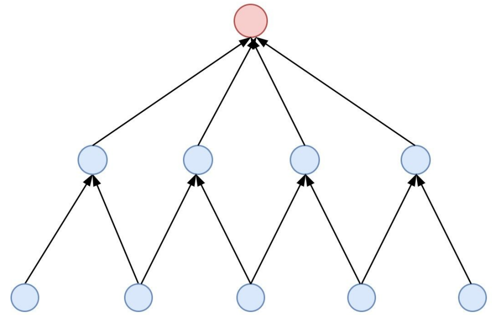


3-tanh target function

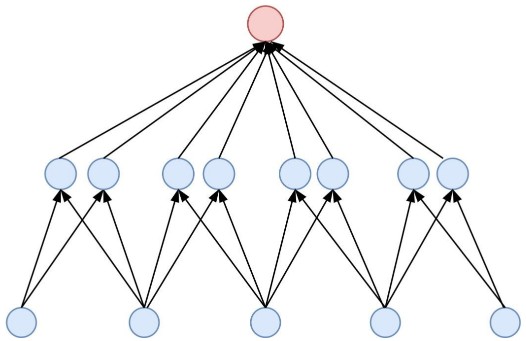


500 neuron tanh-NN

How many samples are required to recover f^*?

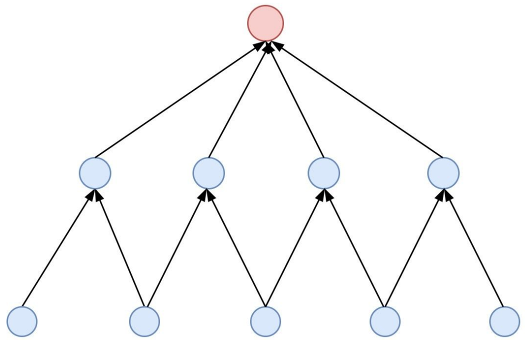
$f^* =$ 

optimistic estimate



NN_C

condense



NN_A (equiv)

parameter cout

12

Optimistic sample size estimate

Parametric model:

$$F: \mathbb{R}^M \to \mathcal{F} \subset C(\mathbb{R}^d)$$

Model rank:

$$r_{\boldsymbol{\theta}} = \dim \operatorname{span} \{\partial_{\theta_i} F(\boldsymbol{\theta})(\cdot)\}_{i=1}^{M}$$

Smaller model rank, stronger condensation!

Optimistic sample size（$f^* \in \mathcal{F}$）：

$$O_{f^*} = \min_{\boldsymbol{\theta} \in F^{-1}(f^*)} r_{\boldsymbol{\theta}}$$

Optimistic estimate: theory vs. experiments

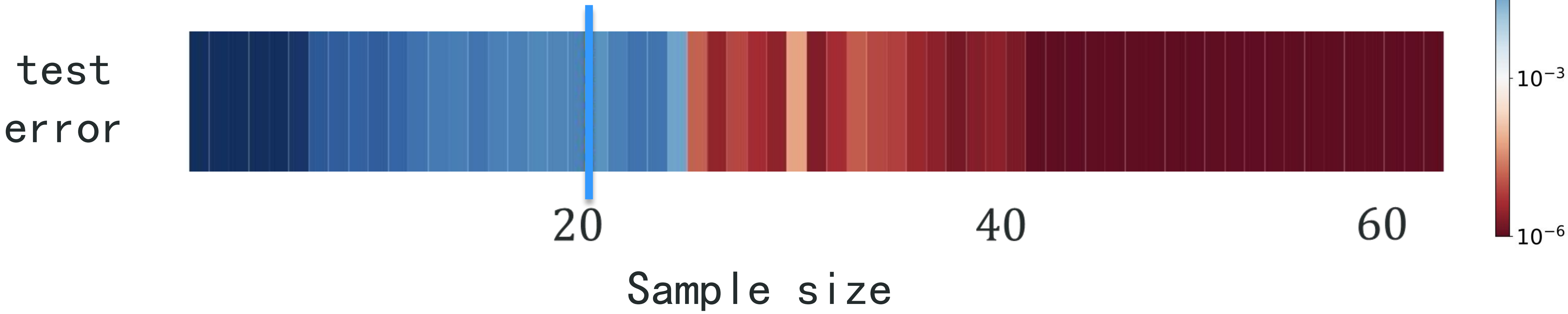
Theorem 5 (optimistic sample sizes for two-layer tanh-NN). *Given a two-layer NN* $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{m} a_i \tanh(\boldsymbol{w}_i^T \boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{\theta} = (a_i, \boldsymbol{w}_i)_{i=1}^m$, *for any target function* $f^* \in \mathcal{F}_k^{\mathrm{NN}} \backslash \mathcal{F}_{k-1}^{\mathrm{NN}}$ *with* $0 \leqslant k \leqslant m$, *the optimistic sample size*

$$O_{f_{\boldsymbol{\theta}}}(f^*) = k(d+1).$$

model size $M = 2100$!

optimistic

$O_{f^*} = 21$

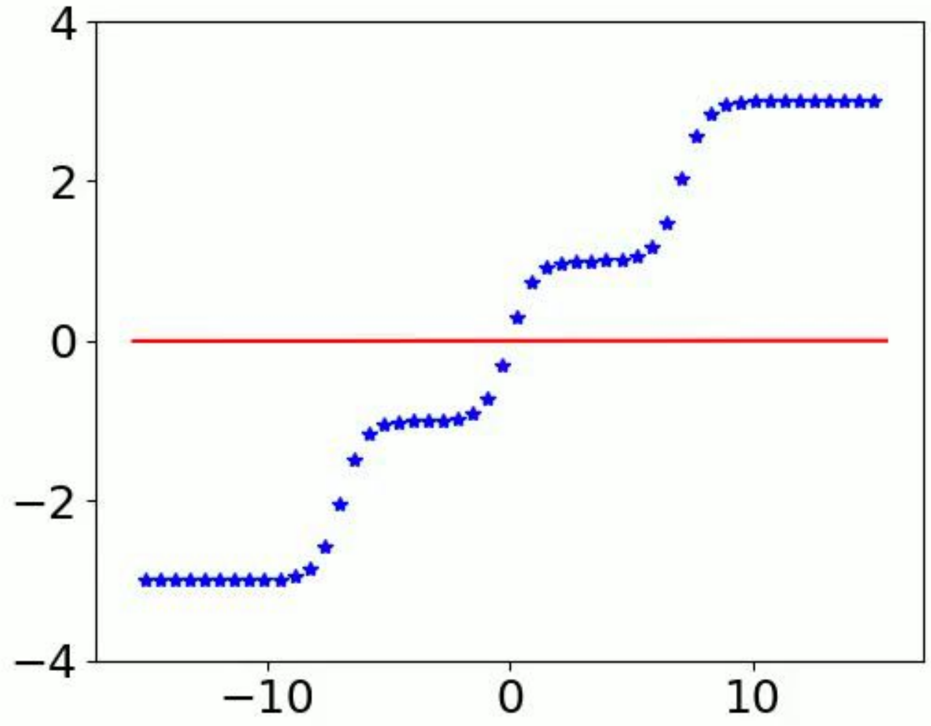


Impact of width—Deep NNs

Theorem 4 (upper bound of optimistic sample size for DNNs). *Given any NN with* M_{wide} *parameters, for any function in the function space of a narrower NN with* M_{narr} *parameters and for any* $f^* \in \mathcal{F}_{\text{narr}}$, *we have* $O_{f_{\boldsymbol{\theta}_{\text{wide}}}}(f^*) \leqslant O_{f_{\boldsymbol{\theta}_{\text{narr}}}}(f^*) \leqslant M_{\text{narr}}$.

wider network is sample efficient

Generalization puzzle explained—A simple example



Width-500 tanh-NN (~1500 parameters)

Q: Why a 1500-parameter NN can (almost) recover 3-tanh target from 50 samples?

A:

1. **More than necessary:** $50 \geq$ **9** (optimistic sample size)
2. **Strong condensation:** Initialize with small variance

Architectural symmetry induces condensation

Architectural symmetry induces condensation

Permutation symmetry of neural networks: e.g., $j, j' \in [m_{l-1}]$

$$\boldsymbol{f}^{[l]}(\boldsymbol{x};\boldsymbol{\theta}) = \sigma\left(\sum_{j=1}^{m_{l-1}} \boldsymbol{W}_{,j}^{[l-1]} \sigma\left(\boldsymbol{W}_{j}^{[l-2]} \boldsymbol{f}^{[l-2]}(\boldsymbol{x};\boldsymbol{\theta}) + b_j^{[l-2]}\right) + \boldsymbol{b}^{[l-1]}\right)$$

Definition 3.1 (structural invariant manifold (SIM)). Let $F(\boldsymbol{\theta})(\boldsymbol{x}), \boldsymbol{\theta} \in \mathbb{R}^M, \boldsymbol{x} \in \mathbb{R}^d$ be an analytic parametric model. For a subset $\mathcal{M} \subset \mathbb{R}^M$, we say $\mathcal{M}$ is a **structural invariant set** if it is invariant under $-\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$ in Eq. (1) for any real analytic loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and dataset S. Moreover, if $\mathcal{M}$ is an immersed submanifold of $\mathbb{R}^M$, we say $\mathcal{M}$ is a **structural invariant manifold.**[3]

Theorem 4.1 (invariant maps induced SIM). Let $F(\boldsymbol{\theta})(\boldsymbol{x})$ be an analytic parametric model with $\boldsymbol{\theta} \in \mathbb{R}^M$ and $\boldsymbol{x} \in \mathbb{R}^d$. Let $\{g_i\}_{i \in I}$ be family of invariant maps of F. Define $\mathcal{M} = \{\boldsymbol{\theta} \mid g_i(\boldsymbol{\theta}) = \boldsymbol{\theta}, \forall i \in I\}$. Assume $\mathcal{M}$ is an immersed submanifold of $\mathbb{R}^M$ with its tangent space satisfying $T_{\boldsymbol{\theta}}\mathcal{M} = \bigcap_{i \in I} \ker(Dg_i^\intercal(\boldsymbol{\theta}) - \mathrm{id}_M), \forall \boldsymbol{\theta} \in \mathcal{M}$. Then $\mathcal{M}$ is a SIM.[4]

SJTU

Jiajie Zhao, Yaoyu Zhang, Tao Luo, “Architecture Induces Structural Invariant Manifolds of Neural Network Training Dynamics”, arXiv:2510.09564v1 (2025).

Permutation symmetry □ sample efficiency preserving

Permutation symmetry of neural networks: e.g., $j, j' \in [m_{l-1}]$

$$f^{[l]}(x;\theta) = \sigma\left(\sum_{j=1}^{m_{l-1}} W_{,j}^{[l-1]} \sigma\left(W_j^{[l-2]} f^{[l-2]}(x;\theta) + b_j^{[l-2]}\right) + b^{[l-1]}\right)$$

Corollary:

Permutation-invariant manifolds are invariant manifolds of gradient flow.

e.g., $\left(W_{,j}^{[l-1]}, W_j^{[l-2]}, b_j^{[l-2]}\right) = \left(W_{,j'}^{[l-1]}, W_{j'}^{[l-2]}, b_{j'}^{[l-2]}\right)$

Effect of permutation symmetry

→dynamics: structural invariant manifolds exhibiting condensation.

→generalization: optimistic sample size no larger than smaller networks.

SJTU

Jiajie Zhao, Yaoyu Zhang, Tao Luo, “Architecture Induces Structural Invariant Manifolds of Neural Network Training Dynamics”, arXiv:2510.09564v1 (2025).
Yaoyu Zhang, Leyang Zhang, Zhongwang Zhang and Zhiwei Bai, Local Linear Recovery Guarantee of Deep Neural Networks at Overparameterization. JMLR 2025

permutation symmetry -> condensation -> optimistic sample efficiency preserving

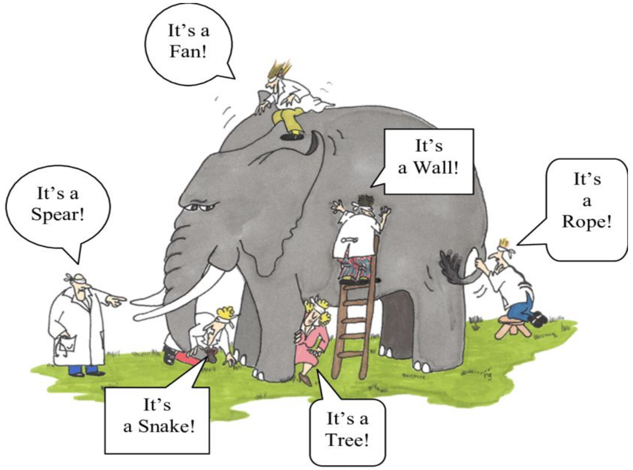
Permutation symmetric:

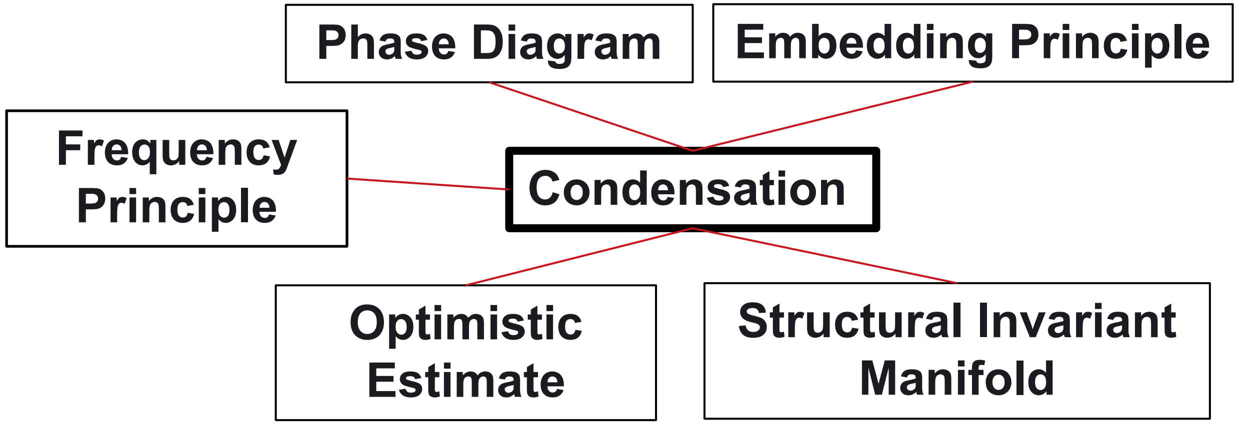
- Embedding dim: d_{model}
- Attention mat dim: d
- Heads: h

→ **Scalable!**

$$A_\theta(X) = \sum_{i=1}^{h} \operatorname{softmax}_{\text{row}} \left(\frac{X W_{Q_i} W_{K_i}^\top X^\top}{\sqrt{d}} \right) X W_{V_i} W_{O_i}^\top$$

Towards the mathematical foundation of deep learning





Homepage

FAU MoD Course

Suspension **Cumulation** **Emergence**

Wir mussen wissen. Wir werden wissen. We must know. We will know. Inscribed on his tomb in Gilttingen.

— *David Hilbert* —

Thanks!

饮 水 思 源 爱 国 荣 校